

2016 LWDA CONFERENCE

**Lernen, Wissen, Daten, Analysen
(LWDA) Conference Proceedings**

LWDA'16

September 12-14, 2016

Potsdam, Germany

Editors:

Ralf Krestel

Hasso Plattner Institute, Germany

Davide Mottin

Hasso Plattner Institute, Germany

Emmanuel Müller

Hasso Plattner Institute, Germany

Preface

LWDA 2016 conference provides a joint forum for experienced and young researchers, to bring insights to recent trends, technologies and applications and to promote interaction in the research field of big data and beyond.

The acronym LWDA expands in German to “Lernen. Wissen. Daten. Analysen.” (English: “Learning. Knowledge. Data. Analytics.”). Recent research in the field is presented and discussed from the viewpoint of machine learning, data mining, knowledge extraction, knowledge management, information retrieval, personalization, database management, information systems, big data management and big data analytics to name a few. This year, the new acronym “LWDA” reflects the successful merger of the former “LWA: Lernen, Wissen, Adaption” and “FGDB Herbsttreffen” conference series with a long tradition in the respective fields.

The LWDA conference series comprises the workshops IR, KDML, FGWM and FGDB which are organized by the respective special interest groups within the German Computer Science Society:

- FG-IR 2016 - Information Retrieval
- FG-KDML 2016 - Knowledge Discovery, Data Mining and Machine Learning
- FG-WM 2016 - Knowledge Management
- FG-DB 2016 - Database Systems

The papers published in LWDA 2016 proceedings have been selected by independent program committees from the respective fields. The program consists of five invited keynotes, one joint research session, and one joint application session with cross-disciplinary talks. In addition to these joint sessions, there are four parallel research sessions for each of the workshops focusing on more specific topics. A joint poster session gives all presenters the opportunity to discuss their work in a broader context. Furthermore, this year’s social program on-top of the joint poster session includes a barbecue event for further interaction on the first evening and on the second evening a boat tour through the picturesque lakes of Potsdam.

Our distinguished keynote speakers are:

- Prof. Dr. Manfred Stede - Universität Potsdam
- Dr. Cédric Archambeau - Amazon
- Prof. Dr. Dorothea Wagner - Karlsruher Institut für Technologie
- Dr. Sebastian Wiczorek - SAP SE
- Prof. Dr. Ulf Leser - Humboldt-Universität zu Berlin

The Hasso-Plattner-Institute (HPI) at the University of Potsdam is proud to host LWDA 2016 conference. For the technical program the organizers would like to thank the workshop chairs and their programme committees for their hard work as well as the keynote speakers for their inspiring talks. Finally we acknowledge the generous support of our thematic partners Amazon and SAP SE as well as our sponsors Idealo and M2 consulting. We hope the participants will keep the venue as an inspiring event with fruitful discussions in mind and the readers will enjoy studying the scientific contributions in this proceedings volume.

Potsdam, Germany, September 2016

Ralf Krestel
Davide Mottin
Emmanuel Müller

Conference Organization

General Chairs

Emmanuel Müller
Ralf Krestel

Hasso Plattner Institute, Germany
Hasso Plattner Institute, Germany

Program Chairs

Klaus-Dieter
Ingo Frommholz
Sebastian Furth
Stephan Günemann
Claus-Peter Klas
Felix Naumann
Ansgar Scherp

Althoff DFKI / University of Hildesheim, Germany
University of Bedfordshire, United Kingdom
denkbares GmbH, Germany
Technical University of Munich, Germany
GESIS - Leibniz Institute for Social Sciences, Germany
Hasso Plattner Institute, Germany
Kiel University and ZBW – Leibniz Information Center for Economics, Germany
Hasso Plattner Institute, Germany

Program Committee

Klaus-Dieter Althoff
Martin Atzmueller
Kerstin Bach
Wolf-Tilo Balke
Christian Bauckhage
Joachim Baumeister
Martin Becker
Axel Benjamins
Ralph Bergmann
Alexander Boehm
Stefan Conrad
Alexander Dallmann
Stefan Deßloch
Jens Dittrich
Stephan Doerfel
Susanne Durst
Michael Fellmann

Universität Hildesheim
University of Kassel
Norwegian University of Science and Technology (NTNU)
Institut für Informationssysteme, TU Braunschweig
Fraunhofer IAIS
denkbares GmbH
University of Würzburg
Universität Osnabrück
University of Trier
SAP SE
Heinrich-Heine University Duesseldorf
University of Würzburg
TU Kaiserslautern
U Saarland
University of Kassel
University of Skövde School of Business
University of Osnabrueck, Institute of Information Management and Corporate Governance
Luton University, UK
Denkbares GmbH
Technical University of Munich
Bauhaus-Universität Weimar
RWTH Aachen University
University of Bamberg
GESIS - Leibniz Institute for the Social Sciences
Martin-Luther-University Halle-Wittenberg
University of Glasgow
University of Wuerzburg
L3S Research Center

Ingo Frommholz
Sebastian Furth
Stephan Günemann
Matthias Hagen
Marwan Hassani
Andreas Henrich
Daniel Hienert
Alexander Hinneburg
Frank Hopfgartner
Andreas Hotho
Robert Jäschke

Dimitris Karagiannis	University of Vienna
Alfons Kemper	TU Muenchen
Kristian Kersting	TU Dortmund University
Claus-Peter Klas	GESIS - Leibniz Institute for Social Sciences, Germany
Andrea Kohlhase	University of Applied Sciences Neu-Ulm
Michael Kohlhase	KWARC
Ralf Krestel	Hasso Plattner Institute
Thomas Krämer	gesis
Florian Lemmerich	GESIS - Leibniz Institute for the Social Sciences, Germany
Ulf Leser	Institut für Informatik, Humboldt-Universität zu Berlin
Johannes Leveling	Elsevier
Michael Leyer	University of Rostock
Thomas Mandl	University of Hildesheim
Alke Martens	University of Rostock, Institute of Computer Science, IEF
Philipp Mayr	GESIS
Mirjam Minor	Goethe University Frankfurt
Bernhard Mitschang	University of Stuttgart
Emmanuel Mueller	Hasso Plattner Institute
Peter Mutschke	GESIS – Leibniz Institute for the Social Sciences
Hannes Mühleisen	Centrum Wiskunde & Informatica (CWI)
Henning Müller	HES-SO
Felix Naumann	Hasso Plattner Institute
Thomas Niebler	University of Wuerzburg
Miltos Petridis	CEM, Brighton University
Nico Piatkowski	TU Dortmund University, LS8, SFB876
Ulrich Reimer	University of Applied Sciences St. Gallen
Achim Rettinger	Karlsruhe Institute of Technology
Jochen Reutelshöfer	denkbares GmbH
Bodo Rieger	Universität Osnabrück, BWL / Management Support und Wirtschaftsinformatik
Thomas Roelleke	Queen Mary University of London
Thomas Roth-Berghofer	School of Computing and Engineering, University of West London
Kai Sachs	SAP AG
Kurt Sandkuhl	The University of Rostock
Christian Severin Sauer	University of West London
Ralf Schenkel	Universitaet Passau
Ansgar Scherp	ZBW - Leibniz Information Centre for Economics and Kiel University
Stefanie Scherzinger	OTH Regensburg
Ute Schmid	Faculty Information Systems and Applied Computer Science, University of Bamberg
Erich Schubert	Ludwig-Maximilians-Universität München
Robin Senge	Computational Intelligence Group @ University of Marburg
Benno Stein	Bauhaus-Universität Weimar
Gerd Stumme	University of Kassel
Jens Teubner	TU Dortmund University
Matthias Uflacker	Hasso-Plattner-Institut Potsdam
Sahar Vahdati	University of Bonn
Christian Wolff	Media Computing, Regensburg University
Arthur Zimek	Ludwig-Maximilians-Universität München

Table of Contents

Keynotes

Argument Mining: Manual and automatic annotation of short user-generated texts	1
<i>Manfred Stede</i>	
Amazon: A Playground for Machine Learning	2
<i>Cédric Archambeau</i>	
Algorithm Engineering for Graph Clustering	3
<i>Dorothea Wagner</i>	
Intelligent Enterprise	4
<i>Sebastian Wiczorek</i>	
Web-Scale Domain-Specific Information Extraction	5
<i>Ulf Leser</i>	

Invited Talks

Visualization-Driven Data Aggregation	6
<i>Uwe Jugel and Zbigniew Jerzak</i>	
Exploring the Application Potential of Relational Web Tables	7
<i>Christian Bizer</i>	
Massively distributed BIG DATA management for Enterprises	9
<i>Franz Faerber, Jonathan Dees, Westenberger Eric and Marc Hartz</i>	
Data Sciences - Ein neuer Studiengang oder nur eine (Informatik-)Vorlesung? (Panel)	10
<i>Kai-Uwe Sattler</i>	
The Evolution of HyPer	11
<i>Thomas Neumann</i>	
The digital transformation is just the beginning	12
<i>Dirk Helbing</i>	

Research Papers

Probabilistic Inference with Stochastic Discrete Clenshaw-Curtis Quadrature	13
<i>Nico Piatkowski</i>	
Sampling Methods for Random Subspace Domain Adaptation	14
<i>Christian Pölitz</i>	
Challenges with Continuous Deployment of NoSQL-backed Database Applications	26
<i>Meike Klettke, Uta Störl, Steffi Scherzinger, Stephanie Sombach and Katharina Wiech</i>	
Case Completion of Workflows for Process-Oriented Case-Based Reasoning	27
<i>Gilbert Müller and Ralph Bergmann</i>	

Human Categorization Learning as Inspiration for Machine Learning Algorithms	28
<i>Christina Zeller and Ute Schmid</i>	
Understanding online financial communities: What constitutes a valuable information exchange for users?	29
<i>Ann-Kathrin Hirzel, Michael Leyer, Nick Russell, Alistair Barros and Jürgen Moormann</i>	
Model-Driven Integration of Compression Algorithms in Column-Store Database Systems	30
<i>Juliana Hildebrandt, Dirk Habich and Wolfgang Lehner</i>	
Min-Hashing for Probabilistic Frequent Subtree Feature Spaces	42
<i>Pascal Welke, Tamas Horvath and Stefan Wrobel</i>	
SEMAFLEX - Semantic Integration of Flexible Workflow and Document Management	43
<i>Lisa Grumbach, Eric Rietzke, Markus Schwinn, Ralph Bergmann and Norbert Kuhn</i>	
Network Analysis with NetworKit - Interactive and Fast	51
<i>Henning Meyerhenke, Elisabetta Bergamini, Moritz von Looz and Christian Staudt</i>	
Applying Topic Model in Context-Aware TV Programs Recommendation	52
<i>Jing Yuan, Andreas Lommatzsch and Mu Mu</i>	
Correlated Variable Selection in High-dimensional Linear Models using Dual Polytope Projection	53
<i>Niharika Gauraha and Swapan Parui</i>	
Ähnlichkeitsbasiertes Retrieval von BPMN-2.0-Modellen	54
<i>Maximilian Pfister, Florian Fuchs and Ralph Bergmann</i>	
Assessing the Quality of Unstructured Data: An Initial Overview	62
<i>Cornelia Kiefer</i>	
An Interactive e-Government Question Answering System	74
<i>Malte Schwarzer, Jonas Düver, Danuta Ploch and Andreas Lommatzsch</i>	
From Cloud to Fog and Sunny Sensors	83
<i>Hannes Grunert, Martin Kasparick, Björn Butzin, Andreas Heuer and Dirk Timmermann</i>	
Matrix Factorization for Near Real-time Geolocation Prediction in Twitter Stream	89
<i>Nghia Duong-Trung, Nicolas Schilling, Lucas Drumond and Lars Schmidt-Thieme</i>	
Scalable Inference in Dynamic Admixture Models	101
<i>Patrick Jähnichen, Florian Wenzel and Marius Kloft</i>	
Knowledge Analytics For Workplace Learning	103
<i>Maria Anna Schett, Stefan Thalmann and Ronald K Maier</i>	
A Reference Model for Anti-Money Laundering in the Financial Sector	111
<i>Felix Timm, Andrea Zasada and Felix Thiede</i>	
Is Web Content a Good Proxy for Real-Life Interaction? A Case Study Considering Online and Offline Interactions of Computer Scientists	121
<i>Mark Kibanov, Martin Atzmueller, Jens Illig, Christoph Scholz, Alain Barrat, Ciro Cattuto and Gerd Stumme</i>	
Predicting video game properties with deep convolutional neural networks using screenshots	122
<i>Przemysław Buczkowski and Antoni Sobkowicz</i>	
Towards a case-based reasoning approach for cloud provisioning	123
<i>Eric Kübler and Mirjam Minor</i>	

Modern Tools for Old Content - in Search of Named Entities in a Finnish OCR'd Historical Newspaper Collection 1771-1910	124
<i>Kimmo Kettunen, Eetu Mäkelä, Juha Kuokkala, Teemu Ruokolainen and Jyrki Niemi</i>	
Experience - the neglected success factor in enterprises?	136
<i>Edith Maier, Werner Bruns, Sebastian Eschenbach and Ulrich Reimer</i>	
Linked Data City - Visualization of Linked Enterprise Data	145
<i>Joachim Baumeister, Sebastian Furth, Lea Roth and Volker Belli</i>	
Case Representation and Similarity Assessment in the selfBACK Decision Support System	153
<i>Kerstin Bach, Tomasz Szczepanski, Agnar Aamodt, Odd Erik Gundersen and Paul Jarle Mork</i>	
Understanding Mathematical Expressions: An Eye-Tracking Study (Resubmission)	155
<i>Andrea Kohlhase and Michael Fürsich</i>	
Synthesizing Invariants via Iterative Learning of Decision Trees	156
<i>Pranav Garg, Daniel Neider, P. Madhusudan and Dan Roth</i>	
BISHOP - Big Data Driven Self-Learning Support for High-performance Ontology Population . . .	157
<i>Daniel Knöll, Martin Atzmueller, Constantin Rieder and Klaus-Peter Scherer</i>	
The Revieal of Subject Analysis: A Knowledge-based Approach facilitating Semantic Search . . .	165
<i>Sebastian Furth, Volker Belli and Joachim Baumeister</i>	
Using key phrases as new queries in building relevance judgments automatically	175
<i>Mireille Makary, Michael Oakes and Fadi Yamout</i>	
CAPLAN: An Accessible, Flexible and Scalable Semantification Architecture	177
<i>Sebastian Furth, Volker Belli, Alexander Legler, Albrecht Striffler and Joachim Baumeister</i>	
Topical Video-On-Demand Recommendations based on Event Detection	186
<i>Tobias Dörsch and Andreas Lommatzsch</i>	
MapReduce Frameworks: Comparing Hadoop and HPCC	194
<i>Fabian Fier, Eva Höfer and Johann-Christoph Freytag</i>	
A Clustering Approach for Holistic Link Discovery (Project overview)	200
<i>Markus Nentwig, Anika Groß and Erhard Rahm</i>	
Query-driven Data Integration (Short Paper)	206
<i>Peter Schwab, Andreas Wahl, Richard Lenz and Klaus Meyer-Wegener</i>	
SABER: Window-Based Hybrid Stream Processing for Heterogeneous Architectures	212
<i>Alexandros Koliouisis, Matthias Weidlich, Raul Castro Fernandez, Alexander Wolf, Paolo Costa and Peter Pietzuch</i>	
Graph n-grams for Scientific Workflow Similarity Search	213
<i>David Luis Wiegandt, Johannes Starlinger and Ulf Leser</i>	
Discovering Data Transformations in Web Resources (Abstract)	225
<i>Ziawasch Abedjan, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti and Michael Stonebraker</i>	
Scalable Detection of Emerging Topics and Geo-spatial Events in Large Textual Streams	226
<i>Erich Schubert, Michael Weiler and Hans-Peter Kriegel</i>	
Approaches for Annotating Medical Documents	227
<i>Victor Christen, Anika Groß and Erhard Rahm</i>	

Robust Query Processing in Co-Processor-accelerated Databases	233
<i>Sebastian Breß, Henning Funke, Jens Teubner and Volker Markl</i>	
On the Evaluation of Outlier Detection: Measures, Datasets, and an Empirical Study Continued . .	234
<i>Guilherme Campos, Arthur Zimek, Jörg Sander, Ricardo Campello, Barbora Micenková, Erich Schubert, Ira Assent and Michael E. Houle</i>	
Finding Trees in Mountains – Outlier Detection on Polygonal Chains	235
<i>Michael Singhof, Daniel Braun and Stefan Conrad</i>	
SemRes: A System for Creating and Searching Semantic Documentation for Conservators.	247
<i>Ernesto William De Luca</i>	
Case-Based Decision Support on Diagnosis and Maintenance in the Aircraft Domain	249
<i>Pascal Reuss, Klaus-Dieter Althoff and Wolfram Henkel</i>	
Building an integrated CBR-Big Data Architecture based on SEASALT	257
<i>Kareem Amin</i>	
Towards rapidly developing database-supported machine learning applications	265
<i>Frank Rosner and Alexander Hinneburg</i>	
Vereinheitlichung internationaler Bibliothekskataloge	271
<i>Christian Scheel, Claudia Schmitz and Ernesto William De Luca</i>	
Sequential Modeling and Structural Anomaly Analytics in Industrial Production Environments . . .	283
<i>Martin Atzmueller, Andreas Schmidt and David Arnu</i>	
Ontology-based Communication Architecture Within a Distributed Case-Based Retrieval System for Architectural Designs	291
<i>Viktor Ayzenshtadt, Klaus-Dieter Althoff, Syed Saqib Bukhari, Andreas Dengel and Ada Mikyas</i>	
Classification of German Newspaper Comments	299
<i>Christian Godde, Konstantina Lazaridou and Ralf Krestel</i>	
k-Means Clustering via the Frank-Wolfe Algorithm	311
<i>Christian Bauckhage</i>	
Plackett-Luce Networks for Dyad Ranking	323
<i>Dirk Schäfer and Eyke Huellermeier</i>	
The Partial Weighted Set Cover Problem with Applications to Outlier Detection and Clustering . .	335
<i>Sebastian Bothe and Tamas Horvath</i>	
Variable Attention and Variable Noise: Forecasting User Activity	347
<i>Cesar Ojeda, Kostadin Cvejovski, Rafet Sifa and Christian Bauckhage</i>	
Mining Subgroups with Exceptional Transition Behavior	359
<i>Florian Lemmerich, Martin Becker, Philipp Singer, Denis Helic, Andreas Hotho and Markus Strohmaier</i>	
Harmony Assumptions: Extending Probability Theory for Information Retrieval	360
<i>Thomas Roelleke</i>	
Comparing contextual and non-contextual features in ANNs for movie rating prediction	361
<i>Ghulam Mustafa and Ingo Frommholz</i>	
Author Index	373

Argument Mining: Manual and automatic annotation of short user-generated texts

Manfred Stede

University of Potsdam, Potsdam, Germany
stede@uni-potsdam.de

In the last few years, argument mining has emerged as a new field that aims to identify argumentative portions in natural language text, and to uncover the structure of the underlying arguments. Domains that have been addressed include legal text, student essays, and customer reviews (as a follow-up step to sentiment analysis). In this talk, I suggest an annotation scheme for argumentation, and present results on automatic analysis of our argumentative microtext corpus - a collection of 115 short texts that have been produced by students in response to a trigger question, which usually bears the form “Should one (not) do X ?” I give results from a joint-inference approach to this task, present various extensions, and then discuss how the approach scales up to longer text.

Biography. Manfred Stede studied Computer Science and Linguistics at TU Berlin and Edinburgh University, and received an M.Sc. in Computer Science from Purdue University (USA). In 1996, he earned his Ph.D. at the University of Toronto with a thesis on multilingual text generation. From 1995 to 2000, he worked at TU Berlin in the large national “Verbmobil” project, which built a system for translating spoken language between German, English, and Japanese. After a short interlude at a company in Berlin, he became a professor in Applied Computational Linguistics at Potsdam University in 2001. His research mainly revolves around issues of text structure, ranging from theoretical models to its automatic analysis, with applications in, e.g., text mining and summarization. Recently, a focus of his research is on different dimensions of subjectivity in language, where speakers convey their attitudes, opinions, and arguments. The well-known computational application is Sentiment Analysis, where Stede contributed to a successful system implementing a lexicon-based approach for English. As a follow-up step, he is now interested in Argument Mining, i.e., the automatic discovery of authors claims, reasons supporting them, and possible objections.

Stede published three monographs, fifteen journal papers, and numerous conference papers and book chapters. He directed research projects funded by various German national agencies and the European Union, sometimes in collaboration with local companies.

Amazon: A Playground for Machine Learning

Cédric Archambeau

Amazon, Berlin, Germany
cedrica@amazon.com

Within Amazon, a company with over 200 millions of active consumers, over 2 million active seller accounts and over 180.000 employees, there are hundreds of problems which can be tackled with Machine Learning. In the first part of this talk, I will give an overview of a number of Machine Learning applications. I will explain how they fit within the Amazon ecosystem, the challenges we are facing and how they help us scale. While Machine Learning is routinely used in recommendation, fraud detection and ad allocation, it plays a key role in devices such as the Kindle or the Echo, as well as the automation of Kiva enabled fulfilment centres, statistical machine translation and automated Fresh produce inspection. In the second part, I will discuss how we democratize machine learning within the company. Applying complex predictive systems, such as machine learning-based systems, in the wild requires to manually tune and adjust knobs, broadly referred to as system parameters or hyperparameters. Black-box optimisation and in particular Bayesian optimisation provides a natural framework for addressing this problem by taking the human expert out of the fine tuning loop. I will introduce Bayesian optimization and discuss open problems in this area.

Biography. Cedric Archambeau is a Senior Machine Learning Scientist with Amazon, Berlin. He manages the algorithms team and served as a technical advisor to Sebastian Gunningham, Amazon Senior Vice President Seller Services. Recently, his team delivered the learning algorithms offered in Amazon Machine Learning. He is interested in large scale probabilistic inference and Bayesian optimization. He holds a visiting position in the Centre for Computational Statistics and Machine Learning at University College London. Prior to joining Amazon, he was leading the Machine Learning and Mechanism Design area at Xerox Research Centre Europe, Grenoble.

Algorithm Engineering for Graph Clustering

Dorothea Wagner

Karlsruher Institut für Technologie (KIT), Karlsruhe, Germany
dorothea.wagner@kit.edu

Graph clustering has become a central tool for the analysis of networks in general, with manifold applications e.g. in data mining, social networks, biology or complex systems. The general aim of graph clustering is to identify dense groups in networks. Countless formalizations thereof exist, among those the widespread measure modularity. However, the overwhelming majority of algorithms for graph clustering relies on heuristics, e.g., for some NP-hard optimization problem, and do not allow for any structural guarantee on their output. Moreover, networks in the real world are often large, evolve over time or come as a data stream.

The talk will discuss algorithmic aspects of graph clustering, especially quality measures and algorithms that are based on the intuition of identifying as clusters dense subgraphs that are loosely connected among one another. We will focus on the algorithm engineering methodology which consists in a cycle of design, analysis, implementation, and experimental evaluation of algorithms, bridging the gap between algorithm theory and practical applications. Special emphasis will be on clustering large networks.

Biography. Dorothea Wagner is a full professor for Informatics at the Karlsruhe Institute of Technology (KIT). Her research interests include design and analysis of algorithms and algorithm engineering, graph algorithms, computational geometry and discrete optimization, particularly applied to transportation systems, energy systems, network analysis, data mining and visualization. Among other activities she is member of the German Council of Science and Humanities (Wissenschaftsrat). From 2007 to 2014 she was vice president of the DFG (Deutsche Forschungsgemeinschaft - German Research Foundation) and 2004 to 2013 speaker of the scientific advisory board of Dagstuhl - Leibniz Center for Informatics. In 2012 she received a Google Focused Research Award, she is member of Academia Europaea and Fellow of the GI (Gesellschaft für Informatik). Dorothea Wagner obtained her diploma and Ph.D. degrees from the RWTH Aachen in 1983 and 1986 respectively; and 1992 the Habilitation degree from the TU Berlin. 1994 - 2003 she was a full professor at the University of Konstanz.

Intelligent Enterprise

Sebastian Wieczorek

SAP SE, Berlin, Germany

Machine Learning describes algorithms that can learn from experience without having to be explicitly programmed. Improved processing power, better algorithms and the availability of big data are the foundation and the reason why Machine Learning is going to take enterprise software to a new level now. We see tremendous potential for our customers. The worlds relevant enterprises rely on SAP. By using Machine Learning we are going to leverage their data for them and let our customers focus on the real job to be done.

Biography. Dr. Sebastian Wieczorek is Director at the SAP Innovation Center Network. In this role he is responsible for the development of SAPs Machine Learning Platform. He is also serving as an academic expert and reviewer for the European Commission and the German Ministry of Education and Research. In previous positions at SAP, Sebastian was coordinating all startup engagement activities of the SAP Innovation Center Network, managed EU-funded research projects, was product owner for SAP Web Analytics, team lead for the “Application Engineering Group” at SAP Research and futurist for the Development Experience team, working from Germany, Israel and the US. Besides working for SAP, Sebastian held lectures on Web Engineering at the Technical University of Darmstadt and obtained a PhD from the Technical University of Berlin (Germany) under supervision of Prof. Ina Schieferdecker. Before joining SAP, Sebastian worked as a Software Developer in Moscow (Russia) and studied computer sciences at the Technical University of Dresden (Germany) and the Northumbria University at Newcastle (United Kingdom).

Web-Scale Domain-Specific Information Extraction

Ulf Leser

Humboldt University, Berlin, Germany
leser@informatik.hu-berlin.de

Information Extraction (IE) from unstructured texts is a technology with growing importance in many applications. Three important challenges to IE are the achievement of high quality results, scalability of methods to very large corpora, and integration of IE results with other data for downstream analysis. In this talk, we will highlight recent advances and open questions in these areas by drawing from extensive experiences in developing and applying IE for biomedical research.

Biography. Ulf Leser studied computer science at the Technische Universität München and obtained his PhD in Data Integration and Query Planning from Technische Universität Berlin. After positions in research institutes and in the private sector, in 2002 he became a professor for Knowledge Management in Bioinformatics at Humboldt-Universität zu Berlin. His research focuses on scientific data management, statistical Bioinformatics, biomedical text mining and infrastructures for large-scale biomedical analysis and is typically carried out in interdisciplinary projects with domain scientists, especially from Medicine and Biology. He is speaker of the DFG-funded graduate school “SOAMED - Service-oriented architectures for medical applications”, chairman of the coordinated BMBF project “PREDICT - Comprehensive Data Integration for Personalized Ontology”, PI of the DFG research unit Stratosphere, and a board member of the DFG-excellence graduate school “BSIO - Berlin School for Integrative Oncology”.

Visualization-Driven Data Aggregation

Uwe Jugel¹ and Zbigniew Jerzak²

¹ LOVOO GmbH, Dresden, Germany

uwe.jugel@lovo.com

² SAP SE, Berlin, Germany

zbigniew.jerzak@sap.com

Visual analysis of high-volume, numerical data is traditionally required for understanding sensor data in manufacturing and engineering scenarios. However, today the visual analysis of any kind of big data has become ubiquitous and is a most-wanted feature for visual analytics tools, required for commerce, finance, sales, and an ever-growing number of industries, whose data is prevalently stored in a relational database management system (RDBMS).

Unfortunately, contemporary RDBMS-based data visualization and analysis systems have difficulties to cope with the hard latency requirements and high ingestion rates required for interactive visualizations of big data. These systems are particularly not able to effectively sample or aggregate the data, inevitably failing to visualize the millions of acquired records. A general-purpose solution for visualization-related data reduction in RDBMS-based systems has been missing. Thereby, custom solutions are tailored to specific domains, supporting only a few custom types of visualizations, while general approaches to data reduction disregard the spatial properties of data visualizations, resulting in measurable and perceivable visualization errors.

To facilitate truly interactive visualizations of the growing volume and variety of big data, computer systems need to change the way they acquire data for the purpose of data visualizations. Visualization-Driven Data Aggregation (VDDA) facilitates up to error-free visualizations of high-volume data sets, at high data reduction rates. Defined as data reduction at the query level and leveraged in a transparent query rewriting system, VDDA is applicable to any visualization system that consumes data stored in relational databases.

This is a resubmission of previously published papers by Jugel et al. [1,2].

Keywords: data visualization, data aggregation, relational databases, query processing, sampling methods

References

1. U. Jugel, Z. Jerzak, G. Hackenbroich, and V. Markl. M4: A visualization-oriented time series data aggregation. *PVLDB*, 7(10):797–808, 2014. (VLDB Best Paper).
2. U. Jugel, Z. Jerzak, G. Hackenbroich, and V. Markl. VDDA: Automatic visualization-driven data aggregation in relational databases. *The VLDB Journal*, 25(1):53–77, February 2016.

Exploring the Application Potential of Relational Web Tables

Christian Bizer

University of Mannheim, Germany
Research Group Data and Web Science
chris@informatik.uni-mannheim.de

The Web contains large amounts of HTML tables. Most of these tables are used for layout purposes, but a small subset of the tables is relational, meaning that they contain structured data describing a set of entities [1]. Relational web tables cover a wide range of topics and there is a growing body of research investigating the utility of web table data for applications such as complementing cross-domain knowledge bases [2], extending arbitrary tables with additional attributes [13, 4], and translating data values [9].

Until recently, most of the research around web tables originated from the large search engine companies as they were the only ones having access to large web crawls and thus were able to extract web table corpora from the crawls. This situation has changed in 2012 with the University of Mannheim [7] and in 2014 with the Dresden University of Technology [3] starting to extract web table corpora from the CommonCrawl, a large public web corpus.

In the talk, I will introduce the 2015 version of the Web Data Commons - Web Table Corpus [7]¹. Afterward, I will give an overview of the different efforts that are currently conducted by my group on exploring the application potential of relational web tables. These efforts include profiling the content [12, 6] of web tables by matching [11] them to cross-domain knowledge bases such as DBpedia [5], fusing web table data in order to complement cross-domain knowledge bases [10], and performing SearchJoins between a local table and a web table corpus in order to extend the local table with additional attributes [8].

References

1. Michael Cafarella, Alon Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu. Uncovering the Relational Web. In *Proceedings of the 11th International Workshop on Web and Databases*, 2008.
2. Xin Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmman, Shaohua Sun, and Wei Zhang. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proc. of the 20th SIGKDD*, pages 601–610, 2014.
3. Julian Eberius, Katrin Braunschweig, Markus Hentsch, Maik Thiele, Ahmad Ahmadov, and Wolfgang Lehner. Building the dresden web table corpus: A classification approach. In *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*, pages 41–50, 2015.

¹ <http://webdatacommons.org/webtables/>

4. Julian Eberius, Maik Thiele, Katrin Braunschweig, and Wolfgang Lehner. Top-k Entity Augmentation Using Consistent Set Covering. In *Proc. of the 27th Int. Conf. on Scientific and Statistical Database Mgmt*, 2015.
5. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*, 6:167–195, 2015.
6. Oliver Lehmborg and Christian Bizer. Web table column categorisation and profiling. In *Proceedings of the 19th International Workshop on Web and Databases*, page 4, 2016.
7. Oliver Lehmborg, Dominique Ritze, Robert Meusel, and Christian Bizer. A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 75–76, 2016.
8. Oliver Lehmborg, Dominique Ritze, Petar Ristoski, Robert Meusel, Heiko Paulheim, and Christian Bizer. The Mannheim Search Join Engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:159–166, 2015.
9. John Morcos, Ziawasch Abedjan, Ihab Francis Ilyas, Mourad Ouzzani, Paolo Papotti, and Michael Stonebraker. Dataxformer: An interactive data transformation tool. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 883–888, New York, NY, USA, 2015. ACM.
10. Yaser Oulabi, Robert Meusel, and Christian Bizer. Fusing time-dependent web table data. In *Proceedings of the 19th International Workshop on Web and Databases*, page 3, 2016.
11. Dominique Ritze, Oliver Lehmborg, and Christian Bizer. Matching HTML Tables to DBpedia. In *Proc. of the 5th Int. Conf. on Web Intelligence, Mining and Semantics*, 2015.
12. Dominique Ritze, Oliver Lehmborg, Yaser Oulabi, and Christian Bizer. Profiling the potential of web tables for augmenting cross-domain knowledge bases. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 251–261, 2016.
13. Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012.

Massively distributed BIG DATA management for Enterprises

Franz Faerber¹, Jonathan Dees¹, Westenberger Eric¹, and Marc Hartz¹

SAP SE, Germany

{franz.fauber, jonathan.dees, eric.westenberger, marc.hartz}@sap.com

More and more companies recognize the value of digitalized information and data for their business. We see a clear trend that business models are adapted or completely changed by putting data into the center of their operations. Managing these data from infrastructure, data ingestion, data consistency, data manipulation to data consumption becomes therefore even more critical especially when considering the huge amount of data, which are created in enterprises (BIG DATA). There are many systems and infrastructures available which deal with BIG DATA focusing on different aspects of the problem space like massive distribution, machine learning, and other topics. Making BIG DATA technologies available for enterprises in a way, that they can build their business on top is still an open challenge. In this presentation we demonstrate on the example of SAP HANA BIG DATA what enterprise readiness means for BIG DATA solutions and how the architecture of such a system looks like. Beside the enterprise features, special focus will be set to the extensibility of the system and the ingestion process.

Data Sciences - Ein neuer Studiengang oder nur eine (Informatik-)Vorlesung? (Panel)

Kai-Uwe Sattler

TU Ilmenau
kus@tu-ilmenau.de

Zusammenfassung. Mit der Verfügbarkeit großer Datenmengen und den Technologien zu ihrer Verarbeitung ist in den vergangenen Jahren der Bedarf an Datenanalysten enorm gestiegen. Gleichzeitig haben sich damit auch die fachlichen Anforderungen verschoben: neben fundierten Statistikkenntnissen und Domänenwissen müssen Datenanalysten heutzutage auch Kenntnisse im Bereich Machine Learning, Datenbanken, Datenintegration sowie im praktischen Umgang mit Big-Data-Technologien und damit Programmierfähigkeiten vorweisen können. Für die universitäre Ausbildung stellt sich daher die Frage, ob diese Anforderungen durch das Curriculum eines Informatikstudienganges abgedeckt sind, ggf. spezialisierte Vorlesungsangebote entwickelt werden sollten oder neue – ggf. interdisziplinäre – Studiengänge eingeführt werden müssen. Im Rahmen des Panels sollen diese Fragen von Experten aus Hochschule und Praxis diskutiert werden.

The Evolution of HyPer

Thomas Neumann

Technische Universität München
neumann@in.tum.de

Abstract. The original vision of HyPer was to build an efficient combined OLTP and OLAP system by exploiting modern hardware, in particular large main-memory capacities. This original design led to a number of further developments, inspired by both technology changes and a broadening of the original goal. Key technologies were a powerful compilation framework and a highly parallelized query processing engine. This talk highlights the evolution and major development steps of HyPer from a specialized research system to an industrial-strength database system.

The digital transformation is just the beginning ...

Dirk Helbing

ETH Zurich

`dirk.helbing@gess.ethz.ch`

The world is running into great trouble. The anthropocene challenges (including climate change, impending resource shortages, demographic change, conflict, financial and economic crises) call for entirely new solutions. As a result, we are now seeing the emergence of data-driven societies around the globe. What organizational framework should we choose? What would be the implications?

Probabilistic Inference with Stochastic Discrete Clenshaw-Curtis Quadrature

Nico Piatkowski

Artificial Intelligence Group, TU Dortmund
44227 Dortmund, Germany
nico.piatkowski@tu-dortmund.de
<http://sfb876.tu-dortmund.de/px>

Abstract. The partition function is fundamental for probabilistic graphical models—it is required for inference, parameter estimation, and model selection. Evaluating this function corresponds to discrete integration, namely a weighted sum over an exponentially large set. This task quickly becomes intractable as the dimensionality of the problem increases. We propose an approximation scheme that, for any discrete graphical model whose parameter vector has bounded norm, estimates the partition function with arbitrarily small error. Our algorithm relies on a near minimax optimal polynomial approximation to the potential function and a Clenshaw-Curtis style quadrature. Furthermore, we show that this algorithm can be randomized to split the computation into a high-complexity part and a low-complexity part, where the latter may be carried out on small computational devices. Experiments confirm that the new randomized algorithm is highly accurate if the parameter norm is small, and is otherwise comparable to methods with unbounded error.

Keywords: graphical model · quadrature · approximation

Summary. An entirely new method for approximate probabilistic inference in exponential family models is presented and discussed. Important preconditions on the models' sufficient statistics, namely χ -integrability, and the dependence of the approximation error on the parameter norm are explained. This work was originally presented at the International Conference on Machine Learning 2016 [1]. In addition to the results present at ICML, we will discuss how this technique could be applied to perform marginal inference and parameter estimation.

Acknowledgments. This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) within the collaborative research center SFB 876, project A1.

References

1. Piatkowski, N., Morik, K.: Stochastic discrete Clenshaw-Curtis quadrature. In: Proceedings of the 33rd International Conference on Machine Learning. JMLR: W&CP, vol. 48. JMLR.org (2016)

Sampling Methods for Random Subspace Domain Adaptation

Christian Pölitz

TU Dortmund University, Otto Hahn Str. 12, 44227 Dortmund

Abstract. Supervised classification tasks like Sentiment Analysis or text classification need labelled training data. These labels can be difficult to obtain, especially for complicated and ambiguous data like texts. Instead of labelling new data, domain adaptation tries to reuse already labelled data from related tasks as training data. We propose a greedy selection strategy to identify a small subset of data samples that are most suited for domain adaptation. Using these samples the adaptation is done on a subspace in a kernel defined feature space. To make this kernel approach applicable for large scale data sets, we use random Fourier features to approximate kernels by expectations.

Introduction

The usual assumption for most of the Data Mining and Machine Learning tasks is that the training data used to learn a model has the same distribution as the test data on that the model is applied. On the other hand, there are many situation where this is not true. Imagine as Data Mining task Sentiment Analysis on product reviews from Amazon. In case of a new product or product type, producers might be interested in how their products catches on. Sentiment Analysis now tries to label the reviews of the corresponding products as being positive or negative. To assign such labels, classification models are trained on some labelled reviews and than applied on the unlabelled reviews. For new product types it is reasonable to assume that we have no labelled training data at hand. Labelling the new reviews can be quite expensive. Especially identifying the sentiment in texts can be hard - even for experts. Ambiguous words or sarcasm for instance make this task difficult. Instead of starting to label the new reviews, another possibility is to reuse already labelled reviews from different products. There might be for instance already labelled reviews about books and now we get new reviews about DVDs. The idea is to leverage the reviews about books to train a classifier that is applied on reviews about DVDs. To accomplish this, we need to find a way to safely transfer the information from one domain to another.

We solve this problem by domain adaptation with the following assumptions: We have two data sets with (possible large) difference in distribution. We have data from a source domain S that is distributed via p_s together with label information y distributed via $p_s(y|x)$. On the other hand, we also have data from a target domain T that is distributed via p_t with no label information. The domain adaptation task now is to use the source domain together with its label information to find a classifier that labels the target domain best.

We expect that many data sets share similarities on latent subspaces. On product reviews for instance, a book might be described as *tedious* while a toaster might be described as *malfunctioning*. Both words have negative meaning and very likely appear together with other negative words like *bad*, *poor* or *poorly*. In a latent subspace in the space spanned by the words, we expect that these words span together a whole dimension. When we map texts of reviews from books and electronic articles onto such a subspace the words *tedious* and *malfunctioning* can be replaced by their common meaning. This will make the texts from the different domains more similar. Further, only terms alone might not be able to find such subspaces. For instance, bi-grams like *little helpful* or *hardly improving* can also span a latent subspace that is helpful for domain adaptation. Generally, n-grams should also be considered.

In order to integrate information of multiple combinations of words, kernels like polynomial kernels can be used. Kernel methods can also integrate structural information and even information from probabilistic models. Consequently, we find low dimensional representations of the data from a source and a target domain in a Reproducing Kernel Hilbert Space. These representations shall keep enough structure from the data that a classifier trained on the source domain still performs well. On the other hand, the low dimensional representation shall make the two data sets more similar. This justifies a safe application of a classifier trained on the source domain, to the target domain.

To find the subspace for the domain adaptation we propose a greedy selection strategy that finds the most useful data samples in the source domain for the domain adaptation. By this, we reduce the data size and concentrate on those samples that are potentially best suited to transfer knowledge. This idea is based on the assumption that not all source samples might be equally important for adaptability. This has been investigated for instance by in [10]. Further, we approximate kernels by random Fourier features as proposed by [19]. This tackles the quadratically or cubically scaling behaviour of kernel methods in the number of data samples.

Related Work

We distinguish two main directions in domain adaptation. On the hand, many of the existing approaches try to find weights for the samples that account for a mismatch in distribution of a target and a source domain. This is especially useful under the so call covariate shift assume. Here, we assume that the distribution of the labels given a sample is the same for both target and source domain. Via the weights, a sample selection bias shall be corrected. This means, we assume that the source domain is sampled from the target distribution applied a certain weighting mechanism. Many previous approaches learn such weights such that the weighted source distributions is most similar to the target distribution.

For instance [8] propose density estimators that incorporate sample selection bias to adapt two distribution, [13] do this by matching the distributions in an RKHS, [14] find the optimal weights by solving least squares problem and [23] minimize the Kullback-Leibler divergence of the target distributions and the weighted source distribution, to name only a few. A theoretical analysis of this adaptation can be found in [6] and [5].

In contrast to these approaches, several other works try to extract a subspace or feature representations in the data space that covers invariant parts across the target and the source distribution. Within such a subspace or feature representations, transferring knowledge between the source and target domain is expected to be more effective than in the whole ambient space.

In [18], Transfer Component Analysis is introduced to find low dimensional representations in a kernel defined Hilbert space. In this representation the target and source domain are more similar than before. The authors in [22] learn a linear subspace that is suitable for transfer learning by minimizing Bregman divergence of the target and source distribution in this subspace, [21] transform the target points such that they are a linear combination of a basis in the source domain, [25] propose to transfer knowledge in a Hilbert space by aligning a kernel with the target domain, [17] learn domain invariant data transformation to minimize differences in source and target domain distributions while preserving functional relations of the data with possible label information. Further, in [9] the authors propose to create subspaces that aligns to the eigenspaces of the target and source domain.

Background

In this section, we introduce the background on kernel methods, subspaces in Hilbert Spaces and distances of distributions. The presented information are crucial for our proposed strategy in the next sections.

Kernel Methods and RKHS

Kernel methods accomplish to apply linear methods on non-linear representations of data. Any kernel method uses a map $X \rightarrow \phi(X)$ from a compact input space X , for instance \mathbb{R}^n , into a so called Reproducing Kernel Hilbert Space (RKHS). In this space, linear methods are applied to the mapped elements like Linear Regressions or Support Vector Machines. The RKHS is a space of functions $f(y) = \phi(x)(y) \forall x \in X$ that allows point evaluations by an inner product, hence $f(y) = \phi(x)(y) = \langle \phi(x), \phi(y) \rangle$. $\phi(x)$ is a function and $\phi(x)(y)$ means the function value at y .

Subspace Methods A subspace in an RKHS H is a closed subset $H' \subset H$. We identify this subspace by a projection P that maps all elements of H into H' . In this work, we concentrate only on subspaces that are spanned by the given data points in the RKHS. This means each element in the subspace can be written as linear combination of all data points in the RKHS, hence $v = \sum_{x \in H} \alpha_i \cdot \phi(x_i)$ for all $v \in H'$. This is important since we only need to consider kernel evaluations and not infinite dimensional elements of the RKHS. Kernel PCA for instance can be used to find an appropriated projection matrix onto such a subspace. See [20] for further details.

Distance Measures

As proposed by [12] the maximum mean discrepancy (MMD) can be used to estimate the difference of two distributions p_s and p_t . For the unit ball H in an RKHS induced by a universal kernel k , the MMD and its empirical estimate are defined as:

$$MMD(H, p_s, p_t)^2 = \|\mu[p_s] - \mu[p_t]\|_H^2$$

respectively

$$\begin{aligned} MMD(H, S, T)^2 &= \frac{1}{|S|^2} \sum_{x_i, x_j \in S} k(x_i, x_j) \\ &\quad - \frac{1}{|S||T|} \sum_{x_i \in S, x_j \in T} k(x_i, x_j) + \frac{1}{|T|^2} \sum_{x_i, x_j \in T} k(x_i, x_j). \end{aligned} \quad (1)$$

Random Features

To avoid large computational and storage complexity of kernel methods, approximations of the kernel can be used. Random features for instance approximate the feature maps in Hilbert spaces by low dimensional random projections. The expectation of the inner products of these random features evaluate to corresponding kernel values. Any shift-invariant kernel (as for example the Gaussian kernel) can be represented as expectation of random features $\cos(\omega x + b)$ for an appropriate distribution $p(\omega)$ and b uniformly drawn from $[0, 2\pi]$, see [19]. For Gaussian kernels, ω is drawn from the distribution: $p(\omega) = (2\pi)^{-k/2} e^{-\|\omega\|^2/2}$. An unbiased estimate of the expectation is $z_\omega(x_i)' z_\omega(x_j)$ for $z_\omega(x) = \frac{\sqrt{2}}{k} [\cos(\omega_1 x), \dots, \cos(\omega_k x), \sin(\omega_1 x), \dots, \sin(\omega_k x)]$.

The deviation of the inner product of the random features of dimension k to the true kernel value is bounded by a tail bound using Hoeffding's inequality. Since $z_\omega \in [-\sqrt{2}, \sqrt{2}]$, we have $z_\omega(x_i)' z_\omega(x_j) \in [-2, 2]$. This and $E_\omega[z_\omega(x_i)' z_\omega(x_j)] = k(x_i, x_j)$ justifies the following bound:

$$P(|z_\omega(x_i)' z_\omega(x_j) - k(x_i, x_j)| \geq \epsilon) \leq 2e^{-k\epsilon^2/8}$$

Domain Adaptation

In a domain adaptation task, we try to use information about a data set S for a classification task on data from set T . For instance, in online reviews about products we might have reviews and information about the sentiment of the reviews about lots of electronic products. Now, the people also start reviewing books. A company might for instance broaden their offers. Now, the new reviews of books shall also be classified by their sentiment. Instead of starting from scratch and labelling all book reviews, we want to leverage the information from all the reviews about electronics that have already been classified by their sentiment. Using this information, a classifier can be learned on a transformed representation of the electronic reviews and be applied to transformed book reviews.

Domain Adaptation via Subspaces

We assume that both data sets lie in the same Hilbert space H by using the same kernel and that their distributions have the same support. Further, we have for each element a probability distribution over a label l that is the same for both data sets. This is the so called Covariate Shift assumption. This means, given an element from H the probability of label l depends not on the set the elements is in, but only on the element.

To transfer knowledge, we project all data onto a low dimensional subspace that captures the structure of the source data and the target data. This is important since otherwise we might not be able to train a good classifier or even project all data points onto a single point. In this case the distributions are the same but we can not train a good classifier.

The simplest way to find a projection onto a subspaces that captures most of the structure is using kernel PCA. We have two data sets that should not loose too much of its structure after projection. The structure of the source domain must be kept to train a good classifier, but the target domain is the actual data we are interested in. Further, we expect that not all information from the source is useful. The idea is now to keep the structure of the target data completely, but for the source data only those parts such that the source and target distributions are close on the subspace that covers only this structure.

Having found a suitable subspace for domain adaptation we project all data orthogonally onto this space. An orthogonal projection onto a low dimensional subspace retracts all data points and makes the distributions of the two data sets more similar. This is true since $\|P \cdot \mu_t - P \cdot \mu_s\| \leq \|P\| \cdot \|\mu_t - \mu_s\|$ and $\|P\| = 1$ for an orthogonal projection P and the mean functional μ_t of the target distribution and μ_s of the source distribution.

Further, the expected distance between classification models on source and target domain decreases. Since we concentrated on linear classifiers in an RKHS, we write any classifier from the source, respectively the target domain as: $h_s(\cdot) = \sum \alpha_i \cdot \langle \phi(x_i^s), \cdot \rangle$ and $h_t(\cdot) = \sum \beta_j \cdot \langle \phi(x_j^t), \cdot \rangle$. Hence, we identify the classifier by weight vectors $w_s = \sum \alpha_i \cdot \phi(x_i^s)$ respectively $w_t = \sum \beta_j \cdot \phi(x_j^t)$. After projecting all elements onto the subspace via P , the corresponding weight vectors are $w_s^P = \sum \alpha_i \cdot P \cdot \phi(x_i^s)$ respectively $w_t^P = \sum \beta_j \cdot P \cdot \phi(x_j^t)$. The distance of any of these classifiers can be bounded in the following way:

$$\begin{aligned}
& \int |w_s^P(x) - w_t^P(x)| p_t(x) dx & (2) \\
&= \int | \sum \alpha_i \cdot P \cdot \phi(x_i^s) - \sum \beta_j \cdot P \cdot \phi(x_j^t) | p_t(x) dx \\
&\leq \|P\| \cdot \int | \sum \alpha_i \cdot \phi(x_i^s) - \sum \beta_j \cdot \phi(x_j^t) | p_t(x) dx \\
&= \int | \sum \alpha_i \cdot \phi(x_i^t) - \sum \beta_j \cdot \phi(x_j^t) | p_t(x) dx \\
&= \int |w_s(x) - w_t(x)| p_t(x) dx
\end{aligned}$$

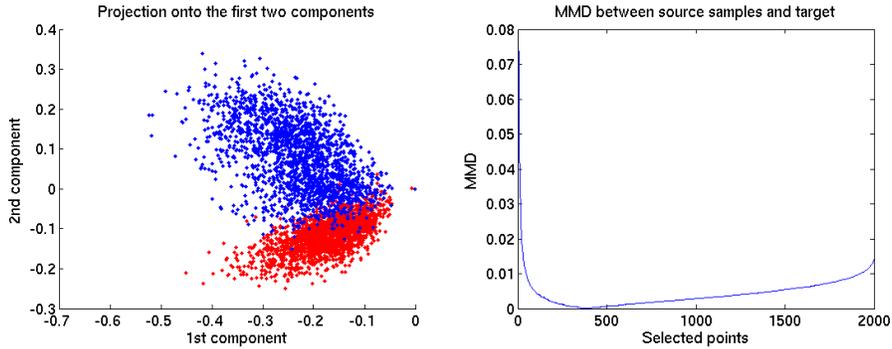


Fig. 1. Illustration of the samplings. Left: Source (electronic reviews in red) and target (DVD reviews in blue) data plotted in the space of the first two components of both of them together. Right: MMD of the selected samples from the source data by Herding based sampling.

Here, we use the fact that the norm of the orthogonal projection is 1, hence $\|P\| = 1$. The bound shows that the expected distance of the linear classifiers in the subspace is less than in the original Hilbert space. The inequality cannot become an equality since we project always on lower dimensional subspace. This shows that these projections decrease the expected error on the target domain of any classifier trained on source domain with a different distribution, see (CF-[2]).

Greedy Selection

To find the most promising data points from the source domain for the domain adaptation, we propose a greedy strategy to efficiently select them. The sampled data points shall be close to the target domain to prevent too much influence of the source domain. On the other hand, the samples must keep enough structure of the source domain such that a good classifier can be trained on the source domain data. The proposed strategy is based on the distance of the source domain distribution to the target domain distribution. The picture in Figure 1 illustrates our idea on electronic (red) and DVD (blue) reviews. We assume the reviews of electronics as target domain and the reviews about DVDs as source domain. The reviews seem to be more similar on one direct than on the other. The idea now is to prefer points from the source domain that are more prominent in this direction for the domain adaptation.

Distribution Based Sampling We propose a sampling strategy that is based on the data distribution. In the Hilbert space we iteratively select mapped samples from the source domain that are most similar to the target distribution. For μ_{p_t} the expectation functional for the target domain in an RKHS, the difference $\|\mu_{p_t} - \frac{1}{n} \sum_{x \in S' \subset S} \phi(x)\|_H^2$ estimates the difference of the target distribution and a subset of samples from the source distribution. Similar approaches are proposed by [4]. The authors showed that the sampling strategy introduced by [24] can be used to match empirical and true distributions in an

RKHS. Equation 3 shows the selection strategy based on matching distributions in an RKHS.

$$\begin{aligned} x_{t+1} &= \operatorname{argmax}_{x \in S - \{x_1, \dots, x_t\}} \langle w_t, \phi(x) \rangle \\ w_{t+1} &= w_t + E_{p_t}[\phi(x)] - \phi(x_{t+1}) \end{aligned} \quad (3)$$

For deciding when to stop the sampling, we monitor $\max_{x \in S - \{x_1, \dots, x_t\}} \langle w_t, \phi(x) \rangle$. As soon as we have only data points from the source data set left that make the distance in distribution no longer decreasing, we stop. By this, we sample only those points such that the empirical distributions of samples and the target data are minimal. The picture on the right of Figure 1 shows an example of the course of the MMD of the samples from the source domain (electronic reviews) and the target domain (DVD reviews). We sample as long as the MMD decreases to find all points that make the distribution similar. This beware us to sample points that make the two distribution dissimilar.

Analysis of Distribution Based Sampling For $\mu_{p_t} = \frac{1}{n_t} \sum_{x_i \in T} \phi(x_i)$, our sampling strategy minimizes:

$$E = \|\mu_{p_t} - \frac{1}{T} \sum_{x_j \in S'} \phi(x_j)\|_H^2.$$

To see this we rewrite

$$E = \langle \mu_{p_t}, \mu_{p_t} \rangle - \frac{2}{T} \sum_{x_j \in S'} \langle \mu_{p_t}, \phi(x_j) \rangle + \frac{1}{T^2} \sum_{x_i, x_j \in S'} \langle \phi(x_i), \phi(x_j) \rangle.$$

Since $\langle \mu_{p_t}, \mu_{p_t} \rangle$ is constant, minimizing E is the same as maximizing

$$\frac{2}{T} \sum_{x_j \in S'} \langle \mu_{p_t}, \phi(x_j) \rangle - \frac{1}{T^2} \sum_{x_i, x_j \in S'} \langle \phi(x_i), \phi(x_j) \rangle.$$

Multiplying the last expression by T results in the greedy sampling as defined above when we set $w_0 = \mu_{p_t}$. This means the strategy matches the empirical distribution of the target samples with the empirical distribution of the subset of the samples from the source distribution.

Random Feature Sampling Our proposed sampling strategy can still result in a large number of points from the source distribution. We further propose to combine the selection strategy and the domain adaptation on a subspace by random features of dimension k . This enables us to perform the domain adaptation task in the linear space spanned by the random Fourier bases of the random features as defined above.

We define MMD_ω similar as MMD in Equation 2 except that the kernel evaluations are replaced by the inner products of the random features. Since $MMD_\omega \in [-8, 8]$, we can apply Hoeffding's inequality to bound the difference to the true MMD by:

$$P(|MMD_\omega^2 - MMD^2| \leq \epsilon) \leq 2e^{-k\epsilon^2/128}.$$

Due to linearity of the expectation we have: $E_\omega MMD_\omega^2 = MMD^2$ and from the definition of the random features we have: $k(x_i, x_j) = E_\omega[z_\omega(x_i)'z_\omega(x_j)]$. All together results in the bound.

Further, we need to estimate how much the components for the random features deviate from the true components the source samples in the RKHS. For this it suffices to investigate the expected difference of the true kernel matrix K for n data points and the matrix of the inner products of the random features K_ω . An appropriate bound is proposed by [16]:

$$E[\|K_\omega - K\|] \leq \sqrt{\frac{2n^2 \log n}{k}} + \sqrt{\frac{2n \log n}{k}}.$$

Experiments

We test our proposed method to find projections onto subspaces for domain adaptation on three standard benchmark data sets that have been used in previous domain adaptation experiments.

As first data set, we use the Amazon reviews [3] about products from the categories books (B), DVDs (D), electronics (E) and kitchen (K). The classification task is to predict a given document as being written in a positive or negative context. We use stop word removal and keep only the words that appear less than 95% and more often than 5% of the time on all documents. The reviews of a certain product will be used as target domain and all the others as source domain.

The second data set is the Reuters-21578 [15] data set. It contains texts about categories like organizations, people and places. For each two of these categories a classification task is set up to distinguish texts by category. Each category is further split into subcategories and different subcategories are used as source and target domains. The exact configuration of the tasks is given by [7].

The third data set is the 20 Newsgroup data set¹. We use the four top-categories (comp,rec,sci and talk) in the same configuration and splits as in [1]. For each two of these top-categories a classification task is set up to distinguish texts by category. Each category is further split into subcategories and different subcategories are used as source and target domains.

For the subspace for domain adaptation, we simply extract the first 100 principle components from the kernel matrix K for all samples from the sampled source domain data and the target domain. This means, for each $x_i, x_j \in \{T \cup S'\}$ we have $K = (k(x_i, x_j))_{i,j}$. We project all data samples (all source and training data) onto the subspace spanned by the extracted components and train a classifier on the source domain in this subspace. Next, we apply this classifier on the target domain in the

¹ <http://qwone.com/~jason/20Newsgroups/>

Method	org vs. places	places vs. org	places vs. people	people vs. places	comp vs. rec	comp vs. sci	comp vs. talk	rec vs. sci	rec vs. talk	sci vs. talk
KMM	60.1	56.8	58.5	56.2	96.9	84.4	98.5	91.2	98.5	95.4
TCA	85.4	80.5	76.5	76.5	94.5	87.8	96.2	90.2	94.1	88.9
GFK	72.9	66.1	68.7	66.4	84.1	74.7	91.9	72.5	86.6	79.02
Sampling	90	82	83.5	79.2	99.1	92	99.2	98.3	99	96.2
Sampling+RF	84.7	82.9	85.5	77.3	98	88.4	98.7	91.7	98	93.7

Table 1. Accuracies on the Reuters and 20 news groups data sets. We compare our proposed greedy sampling methods (without and with random features) and projection with Kernel Mean Matching (KMM) and Transfer Component Analysis (TCA), Gradient Flow Kernel (GFK).

Method	$\{D \cup B \cup K\} \rightarrow E$	$\{E \cup B \cup K\} \rightarrow D$	$\{E \cup D \cup K\} \rightarrow B$	$\{E \cup D \cup B\} \rightarrow K$
KMM	81.0	75.2	72.5	83.9
TCA	81.4	77.8	74.7	84.9
GFK	68.7	66.3	62.2	70.7
Sampling	82.4	79.15	77.25	85.25
Sample+RF	81.3	79.7	77.65	84.85

Table 2. Accuracies on Amazon reviews using one product as target domains and all the other domains as source domain. We compare our proposed greedy sampling methods (without and with random features) and projection with Kernel Mean Matching (KMM) and Transfer Component Analysis (TCA), Gradient Flow Kernel (GFK) and the Landmark method (LM) with projection for domain adaptation.

subspace. We compare the sampling strategies without and with random features (Sampling, Sampling+RF) with Transfer Component Analyses (TCA) [18], Kernel Mean Matching (KMM) [13] and Gradient Flow Kernel (GFK) [11]. For TCA we also use 100 components. We use Gaussian kernels with optimized width parameter σ . For the classification we train an SVM with optimized error weight C . For the random features, the results are mean values over 10 runs with random features of dimension 10,000.

The method by [10] has the same objective as our sampling methods. They find those source domain points that minimize the MMD to the target domain. Compared to our method, the points are extracted by solving a quadratic optimization problem with constraints. This is computationally challenging when we have large source domains. Further, they do not directly select the points, they propose to learn weights of the points and remove those points that have weights below a threshold. This threshold has to be chosen by hand. In the experiments we use the same threshold as they have done in their experiments.

The results of the first experiment are shown in Tables 1 and 2. The projections onto the components result in the best performances for all the domains. The subspace obviously covers the important invariant parts of the data very well. Using random features to approximate the kernel values results in the second best accuracies compared to the other methods.

We now explore how many source domain points have been chosen from which domain.

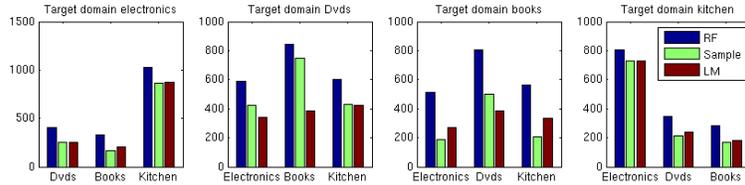


Fig. 2. Histograms of the selected points from Amazon reviews.

Figure 2 shows histograms of the selected data points from the source domain for the different methods. The sampling strategy without and with random features and the GFK method uses a similar amount of samples from the source domains. The histograms show that for each target domain the methods have always one domain in the mixture of source domain where most of the samples are drawn from. For sampling there is always on clear domain from which the method samples most from.

To investigate this further we calculate the Maximum Mean Discrepancy as defined in Equation 2 to estimate the difference of the distributions of the target and source domains. Table 3 shows the MMD values using reviews from the domains. For the electronics reviews (E), the reviews about kitchens (K) are closest in distributions. Comparing this result with the accuracies from above, on the target domain with reviews about electronics, source domain kitchen performs best for domain adaptation. Similar results can be seen for the other domains. Comparing the MMD of the domains with the sampled points from the last experiments, we see that the sampling method chooses the source domain points that results in low MMD best.

MMD	E	D	B	K
E	0	0.0177	0.0207	0.0067
D	0.0177	0	0.0174	0.0173
B	0.0207	0.0174	0	0.0200
K	0.0067	0.0173	0.0200	0

Table 3. Maximum Mean Discrepancy (MMD) measure on the different domains from the categories from the Amazon reviews.

Finally, we investigate the influence of the random features on the quality of the domain adaptation. We perform several runs using different feature sizes.

The plots in Figure 3 show a fast convergence already after some thousand random features. Experiments with random features of dimension less than one thousand has let to poor performance. This might be due to the slower convergence of the kernel matrix to the matrix of the inner products of the random features in the norm. In the future we will investigate this further.

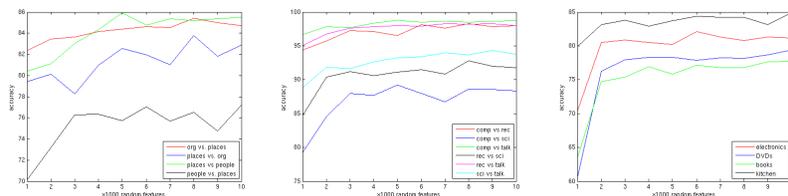


Fig. 3. The classification accuracies using different numbers of random features. Left: the Reuters data set; middle: 20 news groups data set; right: Amazon reviews

Conclusion and Future Work

We proposed a selection strategy on samples from a source domain that are best suited for domain adaptation to a target domain with a different data distribution. The samples are selected to keep the structure of the target domain points while adding some structure from the source domain points. Projecting onto the subspace of the selected samples and the target samples results in a subspace that is well suited for domain adaptation from the source to the target domain. To apply this approach also on large scale data sets, we use random features to approximate kernel values. On benchmark data sets, we showed that our methods perform well on domain adaptation tasks. In the future we want to investigate domain adaptation across different feature spaces. In this context, we want to look at the connections to MKL and domain adaptation using multiple sources.

References

1. Yang Bao, Nigel Collier, and Anindya Datta. A partially supervised cross-collection topic model for cross-domain text classification. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management, CIKM '13*, pages 239–248, New York, NY, USA, 2013. ACM.
2. Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, May 2010.
3. John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
4. Yutian Chen, Max Welling, and Alex J. Smola. Super-samples from kernel herding. *CoRR*, abs/1203.3472, 2012.
5. Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *NIPS*, pages 442–450. Curran Associates, Inc., 2010.
6. Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory, ALT '08*, pages 38–53, Berlin, Heidelberg, 2008. Springer-Verlag.

7. Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 210–219, New York, NY, USA, 2007. ACM.
8. Miroslav Dudík, Robert E. Schapire, and Steven J. Phillips. Correcting sample selection bias in maximum entropy density estimation. In *NIPS*, 2005.
9. Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Subspace alignment for domain adaptation. *CoRR*, abs/1409.5241, 2014.
10. Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML (1)*, volume 28 of *JMLR Proceedings*, pages 222–230. JMLR.org, 2013.
11. Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073. IEEE, 2012.
12. Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample problem. *CoRR*, abs/0805.2368, 2008.
13. Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In Bernhard Schölkopf, John Platt, and Thomas Hoffman, editors, *NIPS*, pages 601–608. MIT Press, 2006.
14. Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.*, 10:1391–1445, December 2009.
15. David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, December 2004.
16. D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf. Randomized nonlinear component analysis. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning, W and CP 32 (1)*, pages 1359–1367. JMLR, 2014.
17. Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. *CoRR*, abs/1301.2115, 2013.
18. Sinno Jialin Pan, I.W. Tsang, J.T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, 22(2):199–210, Feb 2011.
19. Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. 2007.
20. Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. Advances in kernel methods. chapter Kernel Principal Component Analysis, pages 327–352. MIT Press, Cambridge, MA, USA, 1999.
21. Ming Shao, Carlos Castillo, Zhenghong Gu, and Yun Fu. Low-rank transfer subspace learning. *Data Mining, IEEE International Conference on*, 0:1104–1109, 2012.
22. Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, 2010.
23. Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bnau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *NIPS*. Curran Associates, Inc., 2007.
24. Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1121–1128, New York, NY, USA, 2009. ACM.
25. Kai Zhang, Vincent Zheng, Qiaojun Wang, James Kwok, Qiang Yang, and Ivan Marsic. Covariate shift in hilbert space: A solution via surrogate kernels. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 388–395. JMLR Workshop and Conference Proceedings, May 2013.

Challenges with Continuous Deployment of NoSQL-backed Database Applications

Meike Klettke¹, Stefanie Scherzinger², Uta Störl³,
Stephanie Sombach², and Katharina Wiech²

¹ University of Rostock, Germany

² OTH Regensburg, Germany

³ Darmstadt University of Applied Sciences, Germany

We address a practical challenge with the continuous deployment of database applications, which actually constitutes a data integration problem: Upon a new deployment of the application code, entities already persisted in the production database no longer match what the application code expects. Apart from migrating all legacy entities *eagerly* at the time of the release, *lazy* migration is an alternative popular with NoSQL data stores: A schema-flexible database stores entities with legacy structure, as well as up-to-date entities. When a legacy entity is loaded into the application, all pending structural changes are applied. Thus, from the viewpoint of the application, entities are always up-to-date.

Yet lazily migrating legacy data from several releases back, involving more than one entity at-a-time, is not a trivial task. At LWA 2015 [3], we presented our vision of a schema management unit for NoSQL data stores that carries out schema evolution lazily: This involves an internal, Datalog-based model for reading, writing, and migrating data [2]. However, we use Datalog not only to specify the semantics of schema evolution operations, but Datalog is our actual vehicle for carrying out data migrations: In this overview talk, we introduce *Datalution* [1], a tool that alternatively evaluates our Datalog rules bottom-up (for eager data migration) or top-down (for lazy data migration). In particular, our tool allows for an easy comparison of both approaches in terms of the number of physical writes to the data store.

We demonstrate *Datalution*, provide insight into its mechanics, and outline our next steps in integrating the *Datalution* engine with an industrial-strength NoSQL data store.

Keywords: Schema evolution, NoSQL data stores, Datalog

References

1. Scherzinger, S., Sombach, S., Wiech, K., Klettke, M., Störl, U.: Datalution: A Tool for Continuous Schema Evolution in NoSQL-backed Web Applications. In: Proceedings QUDOS'16 (2016), Tool Demo.
2. Scherzinger, S., Störl, U., Klettke, M.: A Datalog-based Protocol for Lazy Data Migration in Agile NoSQL Application Development. In: Proceedings DBPL'15 (2015)
3. Störl, U., Klettke, M., Scherzinger, S.: Kontrolliertes Schema-Evolutionsmanagement für NoSQL-Datenbanksysteme. In: Proceedings LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. (2015)

Case Completion of Workflows for Process-Oriented Case-Based Reasoning

Gilbert Müller and Ralph Bergmann

Business Information Systems II
University of Trier
54286 Trier, Germany
[muellerg] [bergmann]@uni-trier.de,
<http://www.wi2.uni-trier.de>

Abstract. Cases available in real world domains are often incomplete and sometimes lack important information. Using incomplete cases in a CBR system can be harmful, as the lack of information can result in inappropriate similarity computations or incompletely generated adaptation knowledge. Case completion aims to overcome this issue by inferring missing information. This paper presents a novel approach to case completion for process-oriented case-based reasoning (POCBR). In particular, we address the completion of workflow cases by adding missing or incomplete dataflow information. Therefore, we combine automatically learned domain specific completion operators with generic domain-independent default rules. The empirical evaluation demonstrates that the presented completion approach is capable of deriving complete workflows with high quality and a high degree of completeness.

Keywords: process-oriented case-based reasoning, workflows, workflow completion, case completion, completion operators, completion rules

Resubmission of Müller G., Bergmann R.: Case Completion of Workflows for Process-Oriented Case-Based Reasoning. In: Proceedings of the 24th International Conference on Case-Based Reasoning, ICCBR 2016, Atlanta (Georgia), USA. Springer (2016)

Acknowledgements. This work was funded by the German Research Foundation (DFG), project number BE 1373/3-1.

Human Categorization Learning as Inspiration for Machine Learning Algorithms

Christina Zeller and Ute Schmid

Cognitive Systems Group, University of Bamberg
An der Weberei 5, 96045 Bamberg, Germany
{Christina.Zeller,Ute.Schmid}@uni-bamberg.de

Empirical observations of humans learning to categorize inspired the development of early machine learning algorithms (cf. Unger & Wysotzki, 1981). For example, Hunt, Marin, and Stone (1966) developed a decision-tree learning algorithm based on experiments by Bruner, Goodnow, and Austin (1956). However, nowadays the focus of machine learning lies on efficient categorization and not on cognitive plausibility of the underlying learning algorithms.

Recently Lafond, Lacouture, and Cohen (2009) modeled human categorization behavior with decision-trees, but they did not address the question of how these decision-trees are constructed from training trials. We analyzed their data and could show that a measure of incremental information gain can be an appropriate feature selection criterion (Zeller & Schmid, accepted).

Empirical data imply that humans use (meta-)strategies while learning to categorize (cf. Unger & Wysotzki, 1981). Typically, humans focus first on single features as categorization criteria and only later use conjunctions or disjunctions. As a next step, we plan to conduct case studies where participants, while learning correct categorization with a trial by trial feedback, shall explain their decisions. Based on the results, we intend to design experiments where the material is constructed in such a way that cues enhance or hinder specific strategies. We hope that a deeper understanding of the human process of categorization learning can inspire cognitively plausible machine learning algorithms.

References

- Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking: With an appendix on language by Roger W. Brown*. New York, NY: Wiley.
- Hunt, E. B., Marin, J., & Stone, P. J. (1966). *Experiments in induction*. New York, NY: Academic Press.
- Lafond, D., Lacouture, Y., & Cohen, A. L. (2009). Decision-tree models of categorization response times, choice proportions, and typicality judgments. *Psychological Review*, *116*, 833–855.
- Unger, S., & Wysotzki, F. (1981). *Lernfähige Klassifizierungssysteme (Classification Systems Being Able to Learn)*. Berlin, Germany: Akademie-Verlag.
- Zeller, C., & Schmid, U. (accepted). Rule learning from incremental presentation of training examples: Reanalysis of a categorization experiment. In *13th Biannual Conference of the German Cognitive Science Society (Bremen, Germany, Sept. 2016)*.

Understanding online financial communities: What constitutes a valuable information exchange for users?

Ann-Kathrin Hirzel¹, Michael Leyer², Nick Russell³, Alistair Barros³,
Jürgen Moormann¹

¹Frankfurt School of Finance & Management, Frankfurt, Germany
{a.hirzel;j.moormann}@fs.de

²University of Rostock, Rostock, Germany
michael.leyer@uni-rostock.de

³Queensland University of Technology, Brisbane, Australia
{n.russell;alistair.barros}@qut.edu.au

Abstract. Financial service providers continually struggle to attract and maintain customer interest in their company-hosted virtual communities. These are expected to improve the low level of individual involvement and emotional attachment to financial products and services. Based on Bandura's (1986) social cognitive theory, we argue that the content and type of user interaction associated with a created environment influences the level of user issue involvement, which embodies the user's interest in and valuation of the content associated with a virtual community. In particular we examine whether the range of topics, types of contribution and responsiveness of interactions are contributors to the overall level of interest of users in virtual financial communities. Our results, derived from an inductive content analysis based on 8,855 posts from 1,447 users across three virtual financial communities, show that specific topics discussed in a virtual community can have a significant positive influence on a user's overall topic interest. Moreover, a significant positive relationship was found between the type of contribution (e.g., questions, statements or answers) and the extent of a user's topic interest. Furthermore, our results reveal that the timeliness of responses influences a users topic interest in a positive way. However, the overall number of responses related to a specific topic does not play a significant role in any of the communities analysed. This research contributes to a better understanding of virtual communities in the service industry and provides evidence of the importance of content as key driver for user involvement.

Keywords: content analysis, issue involvement, virtual community, financial services

Resubmission of Hirzel, A.-K., Leyer, M., Russell, N., Barros, Alistair, Moormann, J. (2016) Understanding online financial communities: What constitutes a valuable information exchange for users?, European Conference of Information Systems, in: Benbasat, I./Bjorn-Andersen, N./Sencer, A. (Hrsg.), Proceedings of the European Conference on Information Systems 2016, 12.-15.06.2016, Istanbul, Türkei, Paper 1498.

Model-Driven Integration of Compression Algorithms in Column-Store Database Systems

Juliana Hildebrandt, Dirk Habich, and Wolfgang Lehner

Technische Universität Dresden, Database Systems Group,
{juliana.hildebrandt,dirk.habich,wolfgang.lehner}@tu-dresden.de
WWW home page: <https://wwwdb.inf.tu-dresden.de>

Abstract. Modern database systems are very often in the position to store their entire data in main memory. Aside from increased main memory capacities, a further driver for in-memory database systems was the shift to a decomposition storage model in combination with lightweight data compression algorithms. Using both mentioned storage design concepts, large datasets can be held and processed in main memory with a low memory footprint. In recent years, a large corpus of lightweight data compression algorithms has been developed to efficiently support different data characteristics. In this paper, we present our novel model-driven concept to integrate this large and evolving corpus of lightweight data compression algorithms in column-store database systems. Core components of our concept are (i) a unified conceptual model for lightweight compression algorithms, (ii) specifying algorithms as platform-independent model instances, (iii) transforming model instances into low-level system code, and (iv) integrating low-level system code into a storage layer.

1 Introduction

With an ever increasing volume of data in the era of *Big Data*, the storage requirement for database systems grows quickly. In the same way, the pressure to achieve the required processing performance increases, too. For tackling both aspects in a consistent and uniform way, data compression plays an important role. On the one hand, compression drastically reduces storage requirements. On the other hand, compression also is the cornerstone of an efficient processing capability by enabling in-memory technologies. This aspect is heavily utilized in modern in-memory database systems [1, 23] based on a decomposition storage model (DSM) [7]. For the compression of sequences of integer values like applied in DSM compression or in the compression of posting lists in the context of information retrieval, a large corpus of lightweight data compression algorithms has been developed to efficiently support different data characteristics especially of sequences of integer values. Examples of lightweight compression techniques are: frame-of-reference (FOR) [8, 23], delta coding (DELTA) [15, 19], dictionary compression (DICT) [1, 23], bit vectors (BV) [22], run-length encoding (RLE) [1, 19], and null suppression (NS) [1, 19]. These algorithms achieve good compression rates and they provide fast compression as well as decompression. The

corpus evolves further because it is impossible to design an algorithm that always produces optimal results for all kinds of data. As shown, e.g., in [1, 23, 24], the query performance gain for analytical queries is massive.

From the database system architecture perspective, the most challenging task is now to define an approach allowing us to integrate this large and evolving corpus of lightweight compression algorithms in an efficient way. The naïve approach would be to natively implement the algorithms in the storage layer of an in-memory database system as done today. However, this naïve approach has several drawbacks, e.g., (1) massive effort to implement every possible lightweight compression algorithm as well as (2) database developers or even users are not able to integrate a specific compression algorithm without a deep system understanding and implementing the algorithm on a system level. Generally, users know their data and could design appropriate compression algorithms, however, database developers have to preserve the generality of the system and cannot implement many algorithms for a small range of applications. To overcome this situation, we propose our novel model-driven approach to easily and automatically integrate almost every lightweight data compression algorithm in an in-memory DSM storage layer. In detail, we address the following points:

1. We start with a brief solution overview in Section 2. As we are going to show, our solution consists of four components: (i) unified conceptual model for lightweight compression algorithms, (ii) description approach for algorithms as model instances, (iii) transforming model instances to executable code and (iv) integration of generated code into the storage layer.
2. Based on this solution overview, we summarize our conceptual model for lightweight compression algorithms in Section 3. This specific conceptual model helps in describing, understanding, communicating, and comparing the algorithms on a conceptual level.
3. Then, we introduce our description language enabling database developers to specify algorithms in a platform-independent form based on our model.
4. These platform-independent model instances are the foundation for database system integration as described in Section 5. Here, we introduce our approach to transform the platform-independent instances to executable algorithms.
5. We conclude with related work and a summary in Section 6 and 7.

2 Solution Overview

Fundamentally, the model-driven architecture (MDA) is a software design approach for the development of software systems [14, 21]. In the MDA approach, the system functionality is defined with a platform-independent model (PIM) using an appropriate domain-specific language [14, 21]. Then, the PIM is translated into one or more platform-specific models (PSMs) that can be executed [14, 21]. The MDA paradigm is widely used in the area of database applications for database creations. On the one hand, the model-driven data modeling and the generation of normalized database schemas should be mentioned. On the

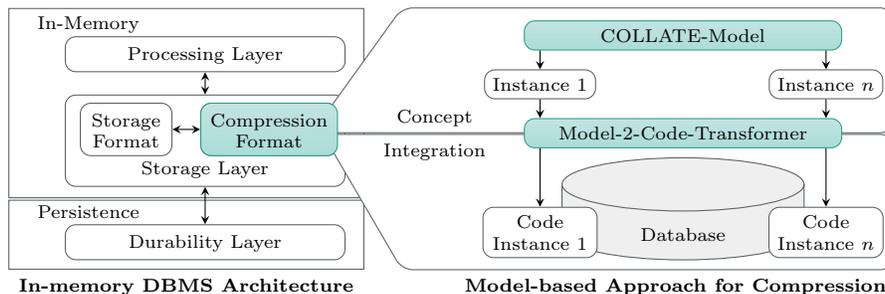


Fig. 1: Model-driven approach for the integration of lightweight data compression algorithms.

other hand, there is the generation of full database applications, including the data schema as well as data layer code, business logic layer code, and even user interface code [10, 18]. Furthermore, the MDA approach has been successfully applied in the area of data warehouse schema creation [17], as well as the modeling and generation of data integration processes [2, 3]. However, there is little to no work on the utilization of MDA for database system-internal components.

As depicted on the left side in Fig. 1, the storage layer of an in-memory database system usually consists of two important components. The first component is the storage format, thereby several formats are proposed. One well-known format is the N-ary storage model (NSM) storing tuples coherently [13]. That means, the tuple unit is preserved in this storage format. The decomposition storage model (DSM), proposed in 1985 [7], is a second widely utilized format. The DSM partitions an n-attribute relation vertically into n sub-relations. Each sub-relation contains two attributes, a logical record id (surrogate) and an attributed value [7]. In most cases, the surrogate can be neglected due to the order of the tuples. That means, sub-relation storage equals to a value-based storage model in form of a sequence of values. While the NSM storage format is used in disk-based database systems, the DSM is the preferred layout of in-memory databases [1, 4, 23, 24]. Compression is the second component of the storage layer, thereby a large and evolving corpus of lightweight data compression algorithms has been developed to efficiently support different data characteristics [1, 15, 23].

For this compression component, we have developed an MDA-based approach as illustrated on the right side in Fig. 1. According to the MDA paradigm, the lightweight data compression algorithms have to be defined in a platform-independent model. To achieve this, we developed a conceptual model for lightweight compression algorithms called *COLLATE* for this specific class of algorithms. The aim of *COLLATE* is to provide a holistic and platform-independent view of necessary concepts including all aspects of data (structure), behavior (function), and interaction (component interaction). In particular, *COLLATE* consists of only five main concepts and we can specify a basic functional arrangement of these concepts as a blueprint for every lightweight data compression algorithm for DSM (see Section 3). That means, the platform-independent

model of a lightweight data compression algorithm is a specific model instance of *COLLATE* expressed in an appropriate language (see Section 4). Then, the platform-specific model can be transformed into a platform-specific executable code as presented in Section 5. The generated code can be used in the database system in a straightforward way. In this paper, we focus on data compression algorithms. The same can be done for the decompression algorithms with a slightly adjusted conceptual model.

3 Conceptual Model

One of our main challenges is the definition of a conceptual model for the class of lightweight data compression algorithms. This is the starting point and anchor of our approach, since all algorithms can be consistently described with this unified and specific model in a platform-independent manner. In [12], we have proposed an appropriate model called *COLLATE* and the development of this model in detail. In the remainder of this section, we briefly summarize its main aspects.

The input for *COLLATE* is a sequence of uncompressed (integer) values due to the DSM storage format. The output is a sequence of compressed values. Input and output data have a logical representation (semantic level) and a physical representation (bit or encoding level). Through the analysis of the available algorithms, we have identified three important aspects. First, there are only six basic techniques which are used in the algorithms. These basic techniques are parameter-dependent and the parameter values are calculated within the algorithms. Second, a lot of algorithms subdivide the input data hierarchically in subsequences for which the parameters can be calculated. The following data processing of a subsequence depends on the subsequence itself. That means, data subdivision and parameter calculation are the adjustment points and the application of the basic techniques is straightforward. Third, for an exact algorithm description, the combination and arrangement of codewords and parameters have to be defined. Here, the algorithms differ widely.

Based on a systematic algorithm analysis, we defined our conceptual model for this class of algorithms. The *COLLATE* model consists of five main concepts—or building blocks—being required to transform a sequence of uncompressed values to a sequence of compressed values:

Recursion: Each model instance includes a **Recursion** per se. This concept is responsible for the hierarchical sequence subdivision and for applying the included concepts in the **Recursion** on each data subsequence.

Tokenizer: This concept is responsible for dividing an input sequence into finite subsequences of k values (or single values).

Parameter Calculator: The concept **Parameter Calculator** determines parameter values for finite subsequences or single values. The specification of the parameter values is done using parameter definitions.

Encoder: The third concept determines the encoded form for values to be compressed at bit level. Again, the concrete encoding is specified using functions representing the basic techniques.

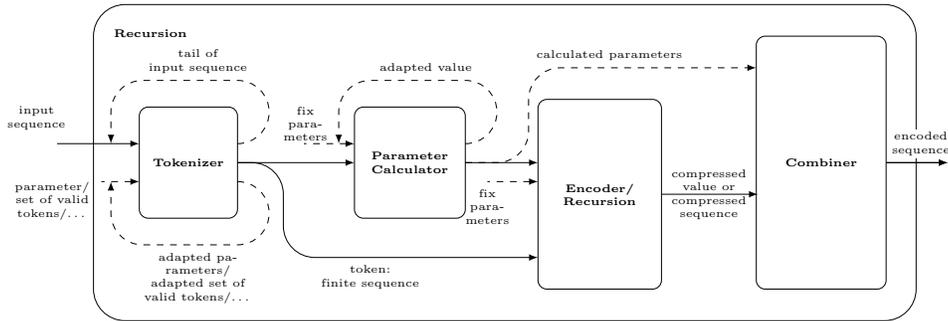


Fig. 2: Interaction and data flow of *COLLATE*.

Combiner: The **Combiner** is essential to arrange the encoded values and the calculated parameters for the output representation.

In addition to these individual concepts, Fig. 2 illustrates the interactions and the data flow through our concepts. In this figure, a simple case with only one pair of **Parameter Calculator** and **Encoder** is depicted and can be described as follows. The input data is first processed by a **Tokenizer**. Most **Tokenizers** need only a finite prefix of a data sequence to decide how many values to output. The rest of the sequence is used as further input for the **Tokenizer** and processed in the same manner (shown with a dashed line). Moreover, there are **Tokenizers** needing the whole (finite) input sequence to decide how to subdivide it. A second task of the **Tokenizer** is to decide for each output sequence which pair of **Parameter Calculator** and **Encoder** is used for the further processing. Most algorithms process all data in the same way, so we need only one pair of **Parameter Calculator** and **Encoder**. Some of them distinguish several cases, so that this choice between several pairs is necessary. The finite **Tokenizer** output sequences serve as input for the **Parameter Calculator** and the **Encoder**.

Parameters are often required for the encoding and decoding. Therefore, we defined the **Parameter Calculator** concept, which knows special rules (parameter definitions) for the calculation of several parameters. Parameters can be used to store a state during data processing. This is depicted with a dashed line. Calculated parameters have a logical representation for further calculations and the encoding of values as well as a representation at bit level, because on the one hand they are needed to calculate the encoding of values, on the other hand they have to be stored additionally to allow the decoding.

The **Encoder** processes an atomic input, where the output of the **Parameter Calculator** and other parameters are additional inputs. The input is a token that cannot or shall not be subdivided anymore. In practice the **Encoder** mostly gets a single integer value to be mapped into a binary code. Similar to the parameter definitions, the **Encoder** calculates a logical representation of its input value and an encoding at bit level using functions. Finally, the **Combiner** arranges the encoded values and the calculated parameters for the output representation.

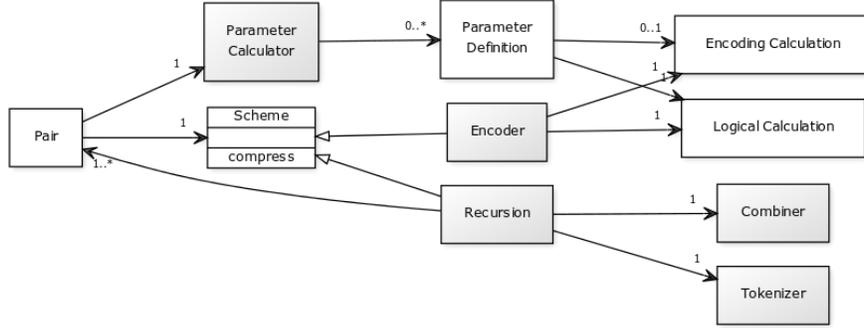


Fig. 3: Class diagram of the *COLLATE* model and data types in the domain-specific language Octave

4 Description Language for Model Instances

Based on our conceptual model, we are able to specify lightweight data compression algorithms in a platform-independent way [12]. For the specification and our overall MDA-based solution, we further require an appropriate language approach. While we introduce our language concept in Section 4.1 in general, we describe an example in Section 4.2.

4.1 Language Approach

Instead of developing a completely new language, we decided to use the GNU Octave¹ high-level programming language as a foundation for the functional behavior of *COLLATE*. A *COLLATE* instance is always a **Recursion** concept. A **Recursion** consists of (1) a **Tokenizer**, (2) a set of pairs of **Parameter Calculator** and a **Scheme**, and (3) a **Combiner** concept. A **Scheme** is either an **Encoder** or a **Recursion**. This relationship can be summarized using a set-oriented notation as follows:

$$\begin{aligned}
 ModelInstance &\in Recursion, \\
 Recursion &= Tokenizer \times \mathcal{P}^{Pair} \times Combiner, \\
 Pair &= ParameterCalculator \times Scheme, \\
 Scheme &= Recursion \cup Encoder.
 \end{aligned}
 \tag{1}$$

Fig. 3 depicts this organization of the *COLLATE* concept as a class diagram in more detail. As described in the previous section, each **Parameter Calculator** can contain several parameter definitions. A parameter definition consists of a logical mapping to calculate a semantic parameter for a subsequence and a bit level mapping (physical level) to calculate the encoding for the semantic parameter in case the parameter has to be encoded. Likewise an **Encoder** consists of a logical mapping and a physical mapping.

¹ <https://www.gnu.org/software/octave/>

As mentioned in the Section 3, concepts contain functions for data processing. Therefore, we combine functional and object-oriented programming for the specification of an algorithm in order to preserve the component interaction. According to the class diagram of Fig. 3 and the set-oriented notation, we are able to specify a *COLLATE* model instance as an object of the class **Recursion** in the Octave high level programming language. The **Recursion** class contains (i) a **Tokenizer** object which is characterized by a function handle, (ii) an array of pairs and (iii) a **Combiner** object which is also defined by a function handle. A function handle is an anonymous function with the syntax `@(argument-list) expression`. The input for a **Tokenizer**'s function handle is (1) a sequence **inp** of values and (2) a structure **par**. This structure contains one field resp. attribute for each valid parameter. These attributes might be global parameters that are valid for each subsequence or each single value, or other calculated parameters. The output of a **Tokenizer**'s function handle returns a triple of values. The first element is the number of output values, the second element is a finite subsequence of the input sequence and the third element is a number that indicates, which of the pairs of **Parameter Calculator** and **Scheme** is chosen for the further data processing. A **Combiner**'s function handle has two input values. The first one is an array **inp** of tuples of a compressed sequence or a compressed value **inp.enc** which is the output of a **Scheme** and the corresponding output of the **Parameter Calculator**, **inp.par**. The second one is the structure **par** that contains all other parameters that are valid for the whole array **inp**. The output of a **Combiner** is a concatenation of compressed sequences or values.

Each pair consists of a **Parameter Calculator** object which is defined by several parameter definition objects and an **Encoder** resp. a **Recursion** object. Each parameter definition contains a function handle for the calculation of logical parameters and a function handle for the calculation of the encoding of the logical parameter. The input for the logical calculation is (1) the subsequence **inp** that is an output of the **Tokenizer** and (2) a structure **par** of known and valid parameters. The output is a logical parameter. Its type is not fix. Often it is one single value, but it can be a more complex parameter, i.e., a mapping like a dictionary. The input of the physical calculation is (1) the logical parameter **inp** and (2) all other known and valid parameters **par**. Its output is the encoding for the parameter. The function composition of both function handles maps a finite subsequence to a bit level representation of a parameter that is valid for the sequence **inp** and the parameters **par**. It is the same for the logical and physical function handles of the encoder.

4.2 Example algorithm

To illustrate our Octave language approach, we now present a simple example lightweight data compression algorithm: *frame-of-reference with binary packing for n values (forbp)*. Here, an input sequence of arbitrary length is subdivided in subsequences of n values. The minimum is calculated for each subsequence of n values as the reference value. So, each value of the subsequence can be mapped to its difference to the reference value at the logical data level. This technique is

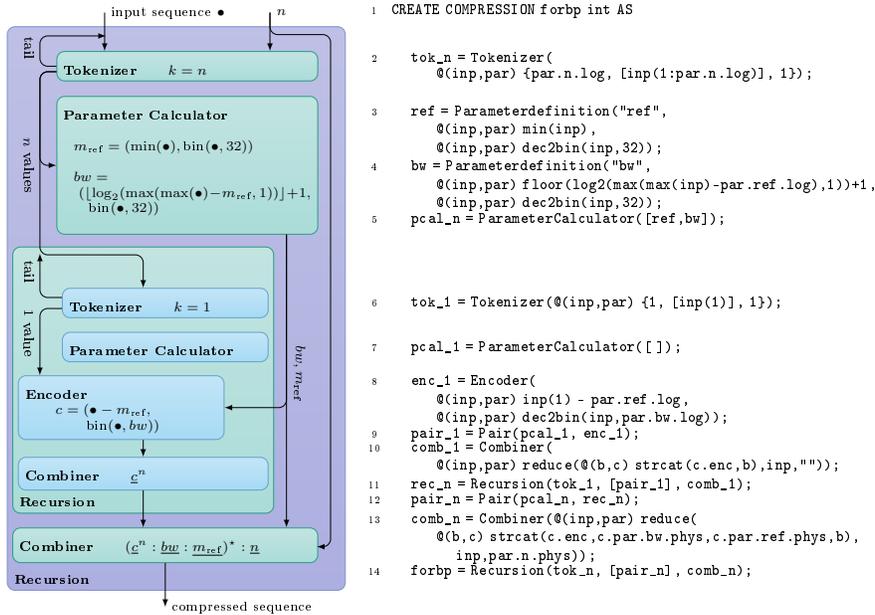


Fig. 4: Graphical representation and Octave code representation example

called frame-of-reference (FOR) and it is used to get smaller values, which can be encoded with a smaller bit width than 32 resp. 64 bits at the physical data level. The bit level representations of all n deltas can be encoded with a common bit width. Because the deltas are smaller than the original values, a possible common bit width might be smaller than 32 resp. 64 bits. This technique is called binary packing (BP). To guarantee the decodability, the reference value and the used bit width have to be stored additionally to every sequence of n encoded values.

Fig. 4 depicts the graphical representation of the corresponding *COLLATE* instance at the left side and the Octave code representation at the right side. The input (\bullet) is an arbitrary sequence of 32-bit integer values. The block size n , a further input, serves as a global parameter to design the algorithm more generic. The whole model instance is an instance of a *Recursion* concept. It consists of a *Tokenizer*, one pair of a *Parameter Calculator* and a *Recursion* and a *Combiner*. The *Tokenizer* outputs the first $k := n$ values. For these n values the *Parameter Calculator* determines the minimum as the reference value $m_{\text{ref}} = \min(\bullet)$ and an appropriate bit width $bw = \lfloor \log_2(\max(\max(\bullet) - m_{\text{ref}}, 1)) \rfloor + 1$ at the logical data level. In the end, both values are encoded in the output with 32 bits each at the physical data level (denoted by the function $\text{bin}(\bullet, 32)$). So each parameter definition consists of one function that addresses the logical data level and one function that considers the encoding resp. physical data level. The next *Recursion* consists of a *Tokenizer*, one pair of *Parameter Calculator* and *Encoder*, and a *Combiner*. This *Tokenizer* outputs single values ($k := 1$). The next *Parameter Calculator* has no task in this instance. Just as parameter

definitions, the **Encoder** consists of one function that processes an input value at the logical level and one function that encodes the value at the physical data level. The **Encoder** determines the difference of the single input value and the reference value at the logical data level (denoted by $c = \bullet - m_{\text{ref}}$) and the encoding of this value with the calculated bit width at the physical data level (denoted by $\text{bin}(\bullet, bw)$). The inner **Combiner** concatenates the physical representations of the n values. This is denoted by \underline{c}^n . Physical representations of values are underlined in the graphical model and power n indicates the concatenation of n values. The outer **Combiner** concatenates the n encoded values with the physical representation of their reference value and the physical representation of their bit width for all blocks and the bit level representation of the number n .

The Octave code at the right side contains the same information. In line 2, a **Tokenizer** object is defined. A function handle is input for the constructor. Its return value is the triple consisting of the number n , the first n values of the input sequence and the number resp. address 1. The whole algorithm as the **Recursion** object **forbp** (line 12) is constructed with the **Tokenizer** (line 2), an array with one pair (line 12) of a **Parameter Calculator** (line 5) and **Recursion** (line 11) and a **Combiner** (line 13). The first calculated parameter is the reference value m_{ref} . The parameter definition is constructed in line 3 with one function handle for the logical calculation and a second function handle for the physical calculation. Each concept can be expressed with one or few lines of Octave code.

5 System Integration

The result of the previous two sections is that we are now able to specify lightweight compression algorithms in a platform-independent way by defining model instances with an Octave notation. As depicted in Fig. 4, the first system integration is done using a **CREATE COMPRESSION** statement to register algorithms in the database system under a user-defined name, e.g., **forbp** as in this example. This system integration continues with (i) an approach to transform model instances to executable code and (ii) the specification of the application of the compression algorithm. For this application, we extended the **CREATE TABLE**-syntax in a straightforward way to allow the specification of the compression algorithm which should be used for each attribute separately:

```
CREATE TABLE Test (attribute_a int compress forbp(
    struct("num", struct("log", 128, "phys", dec2bin(128, 32))));
```

In order to execute model instance specifications inside a database system (e.g. MonetDB [4]), we require an approach to transform model instances to executable code or platform-specific code. Fig. 5 depicts our developed overall approach for this challenge. Generally, we follow a generator approach in our **Model-2-Code Transformer**. At the moment, the input is (i) an Octave specification of a model instance and (ii) code templates for our model concepts. On the **COLLATE** model level, we have 5 specific concepts. That means, we require one code template for each model concept to generate executable code. The code

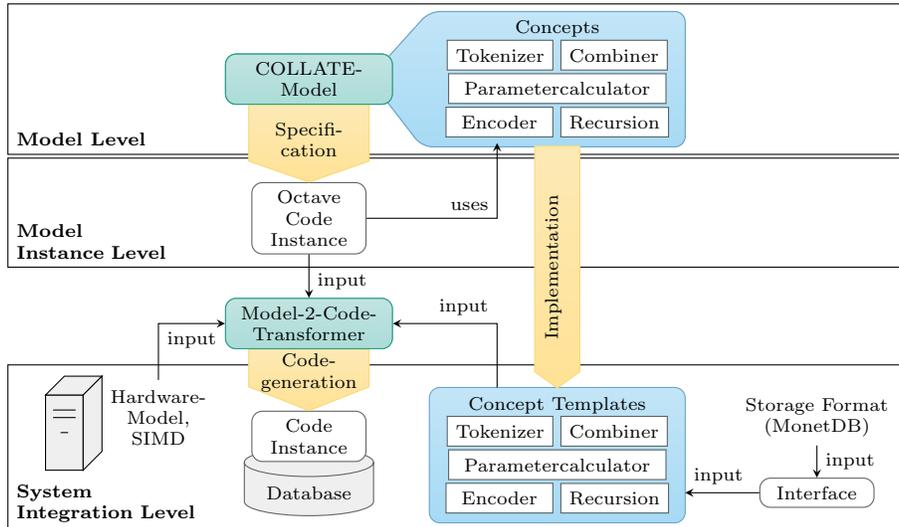


Fig. 5: Transformation of model instances to executable code.

templates have to be implemented once for each specific database system, e.g., MonetDB [4]. This is necessary to get access to data on the specific storage layer implementation.

Based on this input, our **Model-2-Code Transformer** generates the executable code using a replacement and optimization strategy. In this case, the Octave specification is parsed and a corresponding arrangement of the code templates is constructed. The code templates are enriched with the corresponding parameter calculations or encoding functions. Then, the code template arrangement is optimized by applying typical compiler techniques like loop unrolling or load constant replacement [11]. The goal of this optimization is to generate code with less instructions to save as many CPU cycles as possible. This is particularly crucial for reducing the compression overhead.

The generated code can be seamlessly integrated into the specific database system since the templates are implemented for the particular system. At the moment, we are implementing and integrating the whole approach for MonetDB[4], whereas our **Model-2-Code Transformer** is implemented using the LLVM compiler infrastructure. Unfortunately, the current status does not allow a meaningful evaluation with regard to performance.

6 Related Work

In the previous sections, we explained our overall approach for the integration of different data compression algorithms in a column-store database system. With our approach, an in-memory column-store system is extendable with regard to

the bit-level data representation on storage layer level. Generally, extensibility of DBMS was a research field in the late 80's and early 90's [5, 6, 20]. Here, the focus was the integration of new user defined data types, storage methods and indexes, but not the extensibility with regard to new algorithms on top of storage formats. Furthermore, there is also work available focussing on code generation in database systems. For example, Neumann has proposed a completely new query processing architecture [16]. Here, a code generator is used to build specialized query execution plans on demand based on operator templates. He also used the LLVM framework to merge operations into one machine code function. In our work, we follow a similar approach for lightweight data compression algorithms. An interesting research direction would be to combine both approaches for a compression-aware query processing [9].

7 Conclusion

The efficient storage and processing of large datasets is a challenge in the era of *Big Data*. To tackle this challenge, data compression plays an important role from a system-level perspective. Aside from drastically reducing the storage requirement, data compression also enables an efficient processing using "in-memory" technologies. In order to do justice to that, we have presented a model-based approach to integrate a large and evolving corpus of lightweight data compression algorithms in column-store database systems like MonetDB [4] in this paper. Generally, our approach consists of four components: (i) unified conceptual model for lightweight compression algorithms, (ii) description approach for algorithms as model instances, (iii) transforming model instances to executable code and (iv) integration of generated code into the storage layer. In this paper, we concentrated on data compression. The same is also possible for data decompression. In our further research activities, we want to shift our focus to different storage formats to cover not only column-store database systems, but also regard other storage formats, for example row-stores, but our overall approach works.

Acknowledgments

This work was partly funded (1) by the German Research Foundation (DFG) in the context of the project "Lightweight Compression Techniques for the Optimization of Complex Database Queries" (LE-1416/26-1) and (2) by the German Federal Ministry of Education and Research (BMBF) in EXPLOIDS project under grant 16KIS0523.

References

1. Abadi, D.J., Madden, S.R., Ferreira, M.C.: Integrating compression and execution in column-oriented database systems. In: In SIGMOD. pp. 671–682 (2006)
2. Böhm, M., Wloka, U., Habich, D., Lehner, W.: Model-driven generation and optimization of complex integration processes. In: ICEIS. pp. 131–136 (2008)

3. Böhm, M., Wloka, U., Habich, D., Lehner, W.: GCIP: exploiting the generation and optimization of integration processes. In: EDBT. pp. 1128–1131 (2009)
4. Boncz, P.A., Kersten, M.L., Manegold, S.: Breaking the memory wall in monetdb. *Commun. ACM* 51(12), 77–85 (2008)
5. Carey, M., Haas, L.: Extensible database management systems. *SIGMOD Rec.* 19(4), 54–60 (Dec 1990)
6. Carey, M.J., DeWitt, D.J., Frank, D., Muralikrishna, M., Graefe, G., Richardson, J.E., Shekita, E.J.: The architecture of the exodus extensible dbms. In: OODS. pp. 52–65 (1986)
7. Copeland, G.P., Khoshafian, S.N.: A decomposition storage model. *SIGMOD Rec.* 14(4), 268–279 (May 1985)
8. Goldstein, J., Ramakrishnan, R., Shaft, U.: Compressing relations and indexes. In: ICDE. pp. 370–379 (1998)
9. Habich, D., Damme, P., Lehner, W.: Optimierung der Anfrageverarbeitung mittels Kompression der Zwischenergebnisse. In: BTW. pp. 259–278 (2015)
10. Habich, D., Richly, S., Lehner, W.: Gignomda - exploiting cross-layer optimization for complex database applications. In: VLDB (2006)
11. Hänsch, C., Kissinger, T., Habich, D., Lehner, W.: Plan operator specialization using reflective compiler techniques. In: BTW. pp. 363–382 (2015)
12. Hildebrandt, J., Habich, D., Damme, P., Lehner, W.: COLLATE - a conceptual model for lightweight data compression algorithms. Tech. rep., Technische Universität Dresden, Database Systems Group (2016), <https://goo.gl/SgXm5z>
13. Kim, W., Chou, H.T., Banerjee, J.: Operations and implementation of complex objects. In: ICDE. pp. 626–633 (1987)
14. Kleppe, A., Warmer, J., Bast, W.: MDA Explained. The Model Driven Architecture: Practice and Promise. Addison-Wesley (2003)
15. Lemire, D., Boytsov, L.: Decoding billions of integers per second through vectorization. *Softw., Pract. Exper.* 45(1), 1–29 (2015)
16. Neumann, T.: Efficiently compiling efficient query plans for modern hardware. *Proc. VLDB Endow.* 4(9), 539–550 (Jun 2011)
17. OMG: Common Warehouse Metamodel (CWM), Version 1.0 (2001)
18. Richly, S., Habich, D., Lehner, W.: Gignomda - generation of complex database applications. In: Grundlagen von Datenbanken (2006)
19. Roth, M.A., Horn, S.J.V.: Database compression. *SIGMOD Record* 22(3), 31–39 (1993)
20. Schwarz, P., Chang, W., Freytag, J.C., Lohman, G., McPherson, J., Mohan, C., Pirahesh, H.: Extensibility in the starburst database system. In: OODS. pp. 85–92 (1986)
21. Thomas, D., Barry, B.M.: Model driven development: the case for domain oriented programming. In: OOPSLA (2003)
22. Williams, R.: Adaptive Data Compression. Kluwer international series in engineering and computer science: Communications and information theory, Springer US (1991)
23. Zukowski, M., Heman, S., Nes, N., Boncz, P.: Super-scalar RAM-CPU cache compression. In: ICDE. p. 59 (2006)
24. Zukowski, M., Nes, N., Boncz, P.A.: DSM vs. NSM: CPU performance tradeoffs in block-oriented query processing. In: DaMoN. pp. 47–54 (2008)

Min-Hashing for Probabilistic Frequent Subtree Feature Spaces^{*}

Pascal Welke¹, Tamás Horváth^{1,2}, and Stefan Wrobel^{1,2}

¹ Dept. of Computer Science, University of Bonn, Germany

² Fraunhofer IAIS, Schloss Birlinghoven, Sankt Augustin, Germany

Abstract. We propose a fast algorithm for approximating graph similarities. Here, the similarity between two graphs is defined by the Jaccard-similarity of their images in a binary feature space spanned by the set of frequent subtrees generated for some training dataset. While being an adequate choice for many similarity based learning tasks, this approach suffers from severe computational limitations. In particular, mining frequent trees in arbitrary graph databases cannot be done in output polynomial time and embedding a graph in the above space is NP-hard.

To overcome these limitations, we represent each graph by k of its spanning trees generated uniformly at random. In this way, we reduce the frequent subgraph mining, as well as the embedding of a graph into the feature space to problems involving only trees and forests. Clearly, the output of this probabilistic technique is always *sound* (any tree found to be frequent by this algorithm is a frequent subtree with respect to the original dataset), but *incomplete* (the algorithm may miss frequent subtrees). Similarly, the embedding of a given graph in the feature space spanned by the above trees is computed with a one-sided error. We improve the speed and space consumption of the above method by applying min-hashing for the embedding step. Each graph is represented by a small sketch vector that can be used to approximate Jaccard-distances. We show that the partial order on the feature set defined by subgraph isomorphism allows for a fast calculation of the min-hash sketch, without explicitly performing the feature space embedding.

Our experimental results demonstrate that the proposed technique can dramatically reduce the number of subtree isomorphism tests, compared to an algorithm performing the embedding explicitly. We also show that even for a few random spanning trees per chemical compound, remarkable precisions of the active molecules can be obtained in a highly imbalanced chemical dataset by taking the i nearest neighbors of an active compound. Finally, we show that the predictive power of support vector machines using our approximate similarities compares favorably to that of state-of-the-art related methods.

A long version of this extended abstract appears in [1].

- [1] P. Welke, T. Horváth, and S. Wrobel. Min-Hashing for Probabilistic Frequent Subtree Feature Spaces. To appear in: Proceedings of the 19th International Conference on Discovery Science, DS 2016, Springer LNAI, 2016.

SEMAFLEX - Semantic Integration of Flexible Workflow and Document Management

Lisa Grumbach¹, Eric Rietzke², Markus Schwinn², Ralph Bergmann¹ and Norbert Kuhn²

¹ University of Trier, Department of Business Information Systems II,
54286 Trier, Germany

² Trier University of Applied Sciences, Location Birkenfeld, Campusallee,
55761 Birkenfeld, Germany

Abstract. Small and medium-sized enterprises need support by process-aware information systems (PAIS) that offer a high degree of flexibility in workflow execution. Current agile workflow approaches lack acceptance as they introduce a significant overhead for workflow control. Therefore, we propose a new approach for flexible PAIS, based on workflow enactment flexibility by deviation. We enable the potential deviation of the factual workflow from the ideal workflow, while keeping track of the workflow execution. In this paper we describe the proposed SEMAFLEX architecture, which semantically integrates flexible workflow and knowledge-based document management, as well as selected use cases for illustrating the interrelationships among the architectural components.

1 Introduction

Process-Aware Information Systems [1] are essential for efficiency in operational processes for most enterprises in today's business. Large corporations benefit the most, as their business processes and documents are standardized and of large volume. Small and medium-sized enterprises (SME) have different requirements for support concerning their processes. The amount of transactions is much smaller, processes are less standardized [11] and most times weakly structured. The conducted workflows differ significantly due to specific surrounding conditions, peculiar events or coincidences, which implies a need for flexibility. This flexibility might be a competitive advantage concerning large enterprises as ideally a much faster and customized processing of business cases is achieved [7]. Often, current software systems, e.g. Enterprise Resource Planning (ERP) Systems, which are established in large enterprises, cannot satisfy the requirements of SMEs for support of value adding processes, especially due to their lack of flexibility. Another drawback is a significant limitation concerning document management. Business data and documents are managed, but this does not include a semantic analysis and an automatic classification. Thus the content of documents can hardly be used to control the status of processes [2]. In this paper

we present a new approach currently developed within the SEMAFLEX³ project. Its objective is a more efficient supervision of customized business processes on the basis of a semantic integration of processes and business documents. Dependencies between processes and documents will be identified automatically and used for the automatic identification of the actual workflow. A basic idea of this approach is that the ideal workflow, defined previously, and the actual enacted workflow will be distinguished. Hereby, a new approach for flexible workflow enactment, called Flexibility by Deviation [9] (see also Sect. 2), will be provided, which will allow for deviating from the predefined ideal workflow, but without losing control. Detected deviations from the predefined workflow, will be logged, rated and considered for further control. Deficiency management in construction will serve as application scenario, as we expect a great benefit of the presented approach [6]. In this paper, we present the overall idea as well as a proposal for an architecture for implementing flexibility by deviation based on the semantic integration of document content and workflow execution information. After presenting the basic foundations, the SEMAFLEX concept and architecture are presented. The following description of detailed uses cases will illustrate the components of the architecture as well as their interrelationships.

2 Foundations

Flexible approaches concerning workflow management are discussed since about ten years [10]. Four different concepts are distinguished [9]: “*Flexibility by Design* is the ability to incorporate alternative execution paths within a process model at design time.” A major drawback of this approach is that only predictable events might be included in the workflow. The second approach *Flexibility by Change* describes the “ability to modify a process model at runtime”. Since every deviation in the workflow requires a manual intervention and remodeling before continuing with the workflow, the acceptance of this approach in practice is rather low. The same applies to *Flexibility by Underspecification*, which enables to postpone the definition of certain unclear parts of the workflow from design time to runtime. The fourth approach, which is implemented in the presented system, is called *Flexibility by Deviation*. It represents the ability to deviate from the prescribed workflow definition at runtime without manually modifying the workflow. Flexibility by deviation has rarely been explored in research so far. Only FLOWer [2] is an implementation of this approach, which is limited to skipping, undoing, and redoing tasks during enactment.

Workflow management is tightly connected with the exchange of documents, which must be organized systematically. Knowledge-based document management allows the semantic analysis and management of documents with the help of various kinds of background knowledge [4]. Semantic technologies enable to regard documents in the entire context of the knowledge base of the enterprise [5]. For example, it is possible to annotate documents semantically and arrange them as semantic net. Previous work in this research area rarely considers the

³ SEMAFLEX is funded by Stiftung Rheinland-Pfalz für Innovation, grant no. 1158

application context of analysed documents and an explicit context representation is missing. For documents that emerge within a business process controlled by a workflow system, the process context is relevant but also easily available. Business process-oriented knowledge management [3] focusses exactly on the manifold relations between business process and knowledge management, though the latter is regarded predominantly compared to the flexible process enactment of single cases. The virtual office prototype [12] is one of few systems that additionally uses information of process instances for the document analysis.

3 SEMAFLEX Concept and Architecture

The SEMAFLEX concept combines flexible workflow management and knowledge-based document management. As illustrated in Fig. 1 both approaches are semantically integrated on the basis of an ontology, which stores knowledge about documents and workflows. With the help of the document management, incom-

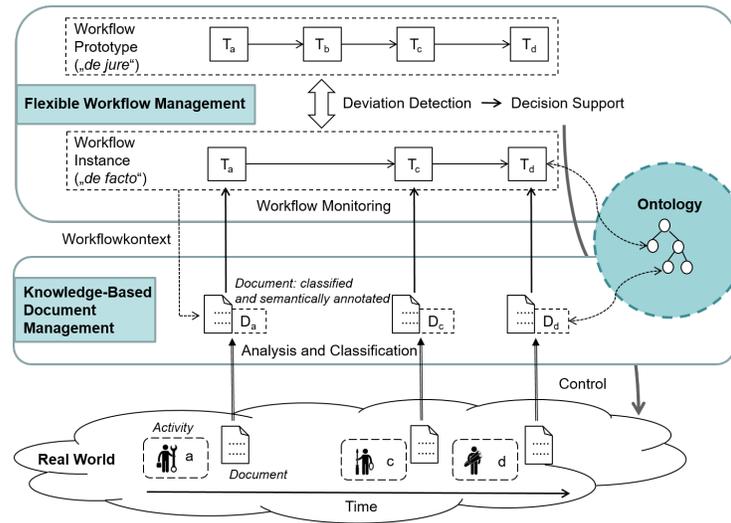


Fig. 1. Concept of SEMAFLEX

ing documents are classified and relevant information is extracted automatically leading to semantic annotations of the documents. This process is based on the conceptual knowledge of the ontology. From the semantic annotation we aim to derive by abductive inference which workflow tasks may have been executed in the real world that caused the observed documents to be present. While specific information in the document might enable to easily determine the workflow instance to which the document belongs to, the semantic description of tasks

will provide knowledge for abductively deriving a hypothesis about actually executed task. Deviations concerning actually enacted tasks (de facto workflow) and defined ideal workflow (de jure workflow) are detected and used for further workflow control. The deviations will be classified w.r.t. their criticality and if necessary, warnings can be issued. In particular, the workflow engine must consider the deviation when determining the further progress of the workflow. For example, it must decide whether a skipped task can be omitted or whether it must be caught up. This requires additional domain specific knowledge about execution constraints on the level of workflow definitions.

To implement this new approach, we propose an architecture on the basis of a semantic integration of knowledge based- document and workflow management. Figure 2 shows the architecture including its single components and their interrelationships. The architecture consists of three main parts. The green

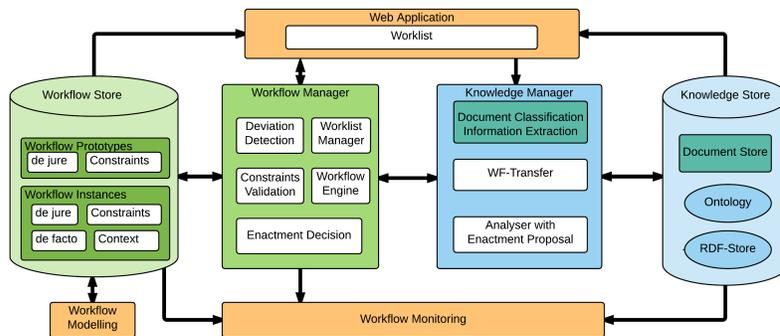


Fig. 2. Architecture of SEMAFLEX

modules are responsible for the general workflow functionality, which basically includes the realization of the approach of Flexibility by Deviation. The blue components realize the semantic integration of documents to represent external process contribution. Every module coloured in orange belongs to a certain user interface. Not only the colours illustrate associated components, but also the layout points out different layers. Whereas the outer parts, i.e. the workflow and the knowledge store, represent the storage layer, the centered components, i.e. the workflow and the knowledge manager, act as application layer. The lower and upper parts incorporate the presentation layer. Smaller modules, depicted in the main modules, take care of specific functionality. The connections between the core components represent the most important kinds of interaction, data flow or activities. In the following every main module and its functionality is described in detail.

Knowledge Manager: The knowledge manager includes all modules that create, extract and process knowledge. With its access to the Knowledge Store

it will expand a simple document store to a knowledge based document management system. The core components combine- extracted data with the predefined conceptual knowledge of the ontology to extract process relevant information within the context of already available knowledge, provided by former documents and by the workflow manager directly. Its main goal is to provide enactment hypotheses to the Workflow manager.

Knowledge Store: The knowledge store contains the stored data like documents, conceptual knowledge, and instantiated knowledge. The document store preserves all documents which were uploaded by the user and offers access to all kind of documents for the purpose of extraction and visualization. The user is granted access to certain documents if they are necessary to enact a task. Another substantial part of the knowledge store is the ontology, which can be divided into two main parts. On the one hand there is knowledge about general concepts, like processes and documents, called upper ontology, which can be applied universally for this workflow-approach. On the other hand a domain ontology is necessary to provide information about domain concepts and specific workflows, with relations between documents and tasks, which is essential for the overall mapping process that incorporates the semantic integration. The instances determined by the extraction module or from notifications of the workflow manager are stored as triples in the RDF store. Beside the metadata and extracted data from documents the RDF Store also reflects the state of all workflow tasks as well as additional information delivered by the workflow manager.

Workflow Manager: The workflow manager covers the application layer concerning workflow functionality and is responsible for the realization of flexibility by deviation. Core components take charge of workflow execution, deviation detection, constraints validation, and managing the task suggestion. The previously mentioned modules are elucidated in the subsequent section with respect to the use cases.

Workflow Store: The workflow store contains all data concerning the workflows, which involves general workflow knowledge, as well as concrete instantiated workflows and their state. Data structures represent the definitions (workflow prototypes), which is the ideal workflow enactment, as well as the executed, traced instances (workflow instances) including the data context. The workflow prototypes comprise the de jure workflow, which is regarded as ideal flow of activities in the context of SEMAFLEX, and constraints which might be constructed additionally. Constraints describe dependencies or requirements between tasks, which should not or must not be violated. As soon as a workflow starts, a new workflow instance will be created on the basis of the corresponding workflow prototype. The current de jure workflow and the constraints of the workflow instance are duplicated and stored in the workflow prototype, as they might be modified over time. The de facto workflow is built step by step as a simple sequence of activities by means of the logged task enactments. The context contains information about the data which is used or generated during the workflow execution. Thus, the user has access to relevant data, while working on tasks.

User Interfaces: There are different types of users, who interact with the system using different interfaces. There will be a graphical user interface for operational users who complete their work following the workflow. Required documents and information, which are necessary to complete certain tasks, are accessible, either loaded from the document or the workflow store. Furthermore, another graphical user interface provides workflow modelling, which should be enabled for users, who have permission to create or adopt workflow prototypes or associated constraints. The third user interface component, which is called Workflow Monitoring, will offer a central overview of all running and terminated workflows. It combines the data from the workflow store and the knowledge store to provide an overview for the real and deviated processes and the impact of and connection to context relevant knowledge.

4 Use Cases

The following use cases derived from the application field of deficiency management in construction [6] will be used to explain the components of the architecture.

The first use case (see Fig. 3, left side) represents a task enactment triggered by the user through the choice of a proposed task like in a standard WFMS. A significant situation in practice might be a deficiency, which is not caused by the enterprise itself, but has to be reported to a subcontractor for remedial actions. The user chooses to send this notification and thus activates the task enactment. Through interaction with the web application, the user selects a task in the worklist which he wants to execute (see 1). The workflow engine is notified about the enactment and updates the de facto workflow, either creating a new instance with the specific task and possibly a corresponding data object, if a new workflow has been started, or appending the new nodes to an existing de facto workflow. This updated information is stored in the workflow store (see 2a). Furthermore the worklist manager updates the suggested tasks in the web application, which might be executed next by the user (see 2b). The selection of these tasks is based upon the data of the de jure workflow and the currently executed task. Besides, the user interfaces are updated concerning the changed workflow status (see 3a). Additionally the knowledge store, specifically the RDF store, is notified about the changing workflow data (see 2c). Another impact of the task enactment, triggered by the workflow engine is the activation of the module deviation detection. This module logs deviations concerning the task enactment with regard to the de jure workflow. Furthermore the deviation detection sets the de jure and the de facto workflow of the workflow instances into relation.

The second use case (cf. right side of Fig. 3) covers the semantic integration of documents and tasks. In deficiency management this might be a received document, which reports the state of the deficiency, with attached picture. Because of extracted information, like customer, contract site, etc. it can be identified as a deficiency acquisition task. As for each deficiency there is one running workflow and as there might be several deficiencies for one contract site, it might be

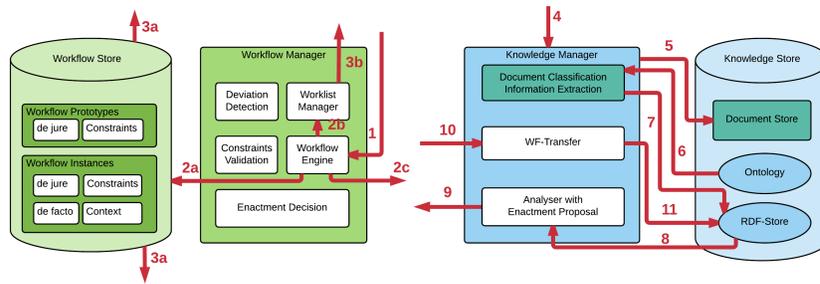


Fig. 3. Processing of Use Case 1 and 2

difficult to determine the corresponding running workflow. Therefore, the user has to manually assign the correct instance. The user transfers a document (see 4), which may be an uploaded pdf-document or a mailed picture, to the web application, which is afterwards processed by the knowledge manager. The first step is to store the document for later accessibility (see 5). Afterwards the module Document Classification and Information Extraction classifies the document based on the definition from the ontology (see 6). The extracted data is stored in the RDF-Store and is linked semantically to available knowledge (see 7). As a new triple is added to the RDF-Store the module Analyser with Enactment Proposal activates an ontology reasoner, which uses modelled inference rules, e.g. property chains of the ontology to check, if this new information can re-enact or start new tasks (see 8). These task candidates are proposed to the workflow manager (see 9). If there are several possible task candidates, the module enactment decision is activated, which involves the user in the mapping process. He might be able to choose the right task instance out of the proposed candidates, as he might be aware of missing information while viewing the corresponding document. An example would be a picture, which represents relevant information. As the extraction module is currently only able to process textual content, visual information is inaccessible and cannot be utilized automatically. If the decision is completed, the chosen task is executed, resulting in an activation of the workflow engine. Constraints are now validated, as the mapping process might have resulted in an undesired state of task enactments. If any constraint is violated, a warning will be send to the user. The warnings are sent to the web application as well as to the workflow monitoring interface. Once an enactment is determined, this new state is send to the knowledge manager (see 10) and subsequently stored in the RDF store (see 11). Such changes will cause another change detection and will start the Analyser again, which results in a loop which ends as soon as no new enactment proposals can be found. The examples reveal that any information might be a contribution to the state of knowledge, whether they come from semantic integrated documents or directly through a user interaction.

5 Conclusion

We presented an architectural concept for a new workflow management approach, which utilizes semantic integration of knowledge-based document and workflow management for the realization of flexibility by deviation. Document classification and information extraction modules are used for the semantic integration, which are ontological-based and use inferencing mechanisms. The workflow component combines imperative and declarative [8] approaches to offer flexibility to the user while preventing him from doing something undesirable. The workflow designer will be able to construct a workflow in an appropriate manner within a range of total flexibility to tight restrictions. Future work will focus on the implementation of the presented approach, followed by an evaluation with business partners, representing the target group of SMEs.

References

1. van der Aalst, W.M.P.: Business process management: A comprehensive survey. ISRN Software Engineering (2013)
2. van der Aalst, W.M.P., Weske, M., Grünbauer, D.: Case handling: a new paradigm for business process support. *Data Knowl. Eng.* 53(2), 129–162 (2005)
3. Abecker, A., Hinkelmann, K., Maus, H., Müller, H.J.: Geschäftsprozessorientiertes Wissensmanagement. Xpert.press (2002)
4. Bläsius, K., Grawemeyer, B., John, I., Kuhn, N.: Knowledge-based document analysis. In: ICDAR'97, Ulm, Germany, Proceedings. pp. 728–731 (1997)
5. Dengel, A.: Semantische Technologien. Springer Spektrum (2012)
6. Gessinger, S., Bergmann, R.: Potentialanalyse des prozessorientierten Wissensmanagement für die Baubranche. In: LWA 2013. Workshop Proceedings Bamberg, 7.-9. October 2013. pp. 195–202 (2013)
7. Levy, M., Powell, P.: SME flexibility and the role of information systems. *Small Business Economics* 11(2), 183–196 (1998)
8. Pesic, M., Schonenberg, H., van der Aalst, W.M.P.: Declarative workflow. In: Modern Business Process Automation - YAWL and its Support Environment, pp. 175–201 (2010)
9. Schonenberg, H., Mans, R., Russell, N., Mulyar, N., van der Aalst, W.M.P.: Process flexibility: A survey of contemporary approaches. In: Advances in Enterprise Engineering I, 4th International Workshop CIAO! and 4th International Workshop EOMAS, held at CAiSE 2008, Montpellier, France, June 16-17, 2008. Proceedings. pp. 16–30 (2008)
10. Schonenberg, H., Mans, R., Russell, N., Mulyar, N., van der Aalst, W.M.P.: Towards a taxonomy of process flexibility. In: Proceedings of the Forum at the CAiSE'08 Conference, Montpellier, France, June 18-20, 2008. pp. 81–84 (2008)
11. Supyuenyong, V., Islam, N., Kulkarni, U.R.: Influence of SME characteristics on knowledge management processes: The case study of enterprise resource planning service providers. *J. Enterprise Inf. Management* 22(1/2), 63–80 (2009)
12. Wenzel, C., Maus, H.: Leveraging corporate context within knowledge-based document analysis and understanding. *IJDAR* 3(4), 248–260 (2001)

Network Analysis with NetworKit – Interactive *and* Fast

Henning Meyerhenke, Elisabetta Bergamini,
Moritz von Looz, and Christian L. Staudt

Faculty of Informatics, Karlsruhe Institute of Technology (KIT), Germany

Network science methodology is increasingly applied to study various real-world phenomena. Consequently, large network data sets comprising millions or billions of edges are more and more common. In order to process and analyze such massive graphs, we need algorithms whose running time is nearly linear in the number of edges. Many analysis methods have been pioneered on small networks, where speed was not the highest concern. Developing a scalable analysis tool suite thus often entails replacing them with suitable faster variants.

Here we present NetworKit (<http://networkkit.itl.kit.edu>), an open-source software suite for analyzing the structure of large networks. We describe our methodology to develop scalable solutions to network analysis problems, including parallelization, fast heuristics for computationally expensive problems, efficient data structures, and modular software architecture. NetworKit is implemented as a hybrid: It combines performance-critical parts in C++ (using OpenMP for parallelism) with a Python user interface, enabling interactive workflows with a scripting language and integration into the Python ecosystem.

Our goal for the software is to put our algorithm engineering efforts into the hands of domain experts. The package provides a wide and growing range of functionality, including common and novel analysis algorithms and graph generators. Focus areas for novel analysis algorithms have been community detection, structure-preserving sparsification, and the ranking of vertices and/or edges based on their structural importance (so-called centralities). For scaling studies and benchmarking purposes, our fast generators can create graphs that exhibit typical complex network structure (e. g. random hyperbolic graphs) or scaled and obfuscated replicas of networks that could otherwise not be shared due to privacy concerns. Also, NetworKit can exploit the correspondence between graphs and matrices (GraphBLAS, <http://graphblas.org>) and has been used for supporting probabilistic range queries in large spatial data sets. In experiments with typical analysis and generation tasks on networks with millions or billions of edges, the ratio of graph size (number of edges) and running time (in seconds) is usually between 10^6 and 10^8 for NetworKit's nearly-linear time algorithms. Compared to the closely related software packages graph-tool and igraph, NetworKit shows consistently the highest speed. Our relevant publications can be found on the NetworKit website (<https://networkkit.itl.kit.edu/publications.html>).

Applying Topic Model in Context-Aware TV Programs Recommendation

Jing Yuan¹, Andreas Lommatzsch¹, Mu Mu²

¹ DAI-Labor, Technische Universität Berlin, Germany
{jing.yuan, andreas.lommatzsch}@dai-labor.de

² The University of Northampton, UK – mu.mu@northampton.ac.uk

Abstract. In IPTV systems, users' watching behavior is influenced by contextual factors like time of day, day of week, Live/VOD condition etc., yet how to incorporate such factors into recommender depends on the choice of basic recommending model. In this paper, we apply a topic model in Information Retrieval (IR)–Latent Dirichlet Allocation (LDA) as the basic model in TV program recommender. What makes employing such approach meaningful is the resemblance between user watching frequency as the entry in user-program matrix and term frequency in term-document matrix. In addition, we propose an extension to this user-oriented LDA by adding a probabilistic selection node in this probabilistic graphical model to learn contextual influence and user's individual inclination on different contextual factors.

The experiment using the proposed approach is conducted on the data from a web-based TV content delivery system “Vision”, which serves the campus users in Lancaster University. The experimental results show that both user-oriented LDA and context-aware LDA converge smoothly on perplexity regarding both iteration epoch and topic numbers under inference framework Gibbs Sampling. In addition, context-aware LDA can perform better than user-based LDA and baseline approach on both precision metrics and diversity metrics when the number of topic is over 50. Aside from that, programs with highest probability distribution within top 10 topics represent the natural clustering effect of applying this topic model in TV recommender.

Keywords: TV recommender, context-awareness, Latent Dirichlet Allocation.

Resubmission of J. Yuan, F. Sivrikaya, F. Hopfgartner, A. Lommatzsch, and M. Mu. Context-aware LDA: Balancing relevance and diversity in TV content recommenders. In Proc. of the 2nd Workshop on RecsysTV, Sept. 2015.

Acknowledgement The work of the first author has been continuously funded by China Scholarship Council (CSC).

Correlated Variable Selection in High-dimensional Linear Models using Dual Polytope Projection

Niharika Gauraha and Swapan K. Parui

Indian Statistical Institute, India

We consider the case of high dimensional linear models ($p \gg n$) with strong empirical correlation among variables. The Lasso is a widely used regularized regression method for variable selection, but it tends to select a single variable from a group of strongly correlated variables even if many or all of these variables are important. In many situations, it is desirable to identify all the relevant correlated variables, examples include micro-array analysis and genome-wide association studies. We propose to use Dual Polytope Projections (DPP) rule, for selecting the relevant correlated variables which are not selected by the Lasso.

We consider the usual linear model setup, that is given as $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. Let $\lambda \geq 0$ be a regularization parameter. Then the Lasso estimator(see [1]) is defined as: $\hat{\beta}(\lambda) = \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$. Let $\lambda_{max} = \max_{1 \leq j \leq p} |\mathbf{X}_j^T \mathbf{Y}|$, then for all $\lambda \in [\lambda_{max}, \infty)$, we have $\hat{\beta}(\lambda) = 0$. It has been shown that the screening methods based on DPP rule are highly effective in reducing the dimensionality by discarding the irrelevant variables (see [2]). Suppose we want to compute Lasso solution for a $\lambda \in (0, \lambda_{max})$, the (global strong) DPP rule discards the j^{th} variable whenever $|\mathbf{X}_j^T \mathbf{Y}| < 2\lambda - \lambda_{max}$ (variables having smaller inner products with the response).

Exploiting the above property, we propose a two-stage procedure for variable selection. At the first stage, we perform Lasso using cross-validation and we choose the regularization parameter λ_{Lasso} , that optimizes the prediction. At the second stage, we select all the variables for which $|\mathbf{X}_j^T \mathbf{Y}| \geq 2\lambda_{Lasso} - \lambda_{max}$. Though, the Lasso solution at λ_{Lasso} does not include all the relevant correlated variables, but these correlated variables have the similar magnitude for their inner products with the response. Hence, all the relevant correlated predictors also get selected at the second stage.

Keywords: Correlated Variable Selection, Lasso, Dual Polytope Projection, High-dimensional Data Analysis

References

- [1] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *J. R. Statist. Soc* 58 (1996), 267–288.
- [2] Jie Wang et al. "Lasso Screening Rules via Dual Polytope Projection". In: *NIPS* (2013).

Ähnlichkeitsbasiertes Retrieval von BPMN-2.0-Modellen

Maximilian Pfister, Florian Fuchs und Ralph Bergmann

Universität Trier, Wirtschaftsinformatik II,
Universitätsring, 54286 Trier, Germany
[s4mapfis] [s4flfuch] [bergmann]@uni-trier.de,
<http://www.wi2.uni-trier.de>

Zusammenfassung. Business Process Model and Notation (BPMN) 2.0 gilt als einer der vielversprechendsten Standards zur Prozessmodellierung, jedoch ist die Modellierung auch für Experten mit großem Aufwand verbunden, sodass heute versucht wird, die Modellierung durch Wiederverwendung von Prozessmodellen zu unterstützen. Hierzu werden Repositories von qualitativ hochwertigen Geschäftsprozessmodellen aufgebaut, aus denen wiederverwendbare Modelle ausgewählt werden können. In dieser Arbeit wird ein Ansatz aus dem prozessorientierten Fallbasierten Schließen für das ähnlichkeitsbasierte Retrieval von BPMN-2.0-Modellen erweitert und empirisch erprobt. Anhand eines existierenden BPMN-Repositories wird überprüft, welchen Mehrwert hierbei die Einbeziehung der Semantik in die Ähnlichkeitsberechnung gegenüber einer rein lexikalischen Bewertung erbringt.

1 Einleitung

Modellierung und Analyse von Geschäftsprozessen sind wichtige Aufgaben für moderne Unternehmen. Es existieren heute verschiedene Notationen zur Prozessmodellierung. Nach Allweyer [3] unterscheidet man folgende Standards: XPD (XML Process Definition Language) und BPEL (Business Process Execution Language) sind Prozessausführungssprachen, die hauptsächlich zur Definition von automatisierten Prozessen genutzt werden. EPC (Event-driven Process Chain) ist eine Notation mit einem stärkeren Fokus auf Geschäftsprozessmodellierung. Diese Notation ist jedoch kein offener Standard. Bei der Business Process Model and Notation (BPMN) steht die Modellierung und Visualisierung von Geschäftsprozessen im Vordergrund. “Die Notation wird sowohl für fachliche Prozessmodelle [(Geschäftsprozesse), Anm. d. Verf.] verwendet als auch für detaillierte Ablaufspezifikationen [(Workflows), Anm. d. Verf.], die durch Process Engines ausgeführt werden” [2].

BPMN hat in der aktuellen Version 2.0 eine weitere Verbreitung erfahren, was dazu geführt hat, dass umfangreichere Prozessrepositorien in diesem Format entstehen [11]. Die Wiederverwendung von Prozessmodellen stellt eine vielversprechende Möglichkeit dar, um die Produktivität von Prozessgestaltern zu steigern, den Designprozess zu beschleunigen und Redundanzen in Prozessdatenbanken zu vermeiden [10]. Um ein Retrieval von gespeicherten Geschäftsprozessen zu ermöglichen, ist es notwendig, Ähnlichkeitsmerkmale für Prozesse zu definieren, um diese vergleichbar zu machen.

Idealerweise wird das Prozessmodell aus dem Repository ausgesucht, das die Anforderungen der aktuellen Suchanfrage am besten erfüllt, sodass anschließend nur noch wenige Anpassungen vorgenommen werden müssen, um ein geeignetes Prozessmodell zu erhalten. Es werden neue Methoden zur Ähnlichkeitsberechnung und zum Retrieval von Prozessmodellen benötigt [4], um Experten dahingehend zu unterstützen, dass sie nicht nur strukturell, sondern auch semantisch verwandte Prozesse finden können. Prozessorientiertes Fallbasiertes Schließen stellt einen Ansatz dar, um die Wiederverwendung von Prozessmodellen zu ermöglichen [14].

Fallbasiertes Schließen (Case-based Reasoning, CBR) [1,6,15] ist eine Technik, die erfahrungsbasiertes Problemlösen unterstützt. Neue Probleme werden dadurch gelöst, dass die Lösungen ähnlicher Probleme aus einer Fallbasis abgerufen und angepasst werden. Prozessorientiertes Fallbasiertes Schließen (Process-oriented Case-based Reasoning, POCBR) befasst sich mit der Verbindung von CBR mit prozessorientierten Informationssystemen [14]. POCBR ermöglicht die Erstellung neuer Prozessmodelle durch Auswahl (Retrieval) von bestehenden, zu einer Anfrage ähnlichen, Prozessmodellen aus der Fallbasis (Repository). Hierbei geht man davon aus, dass als Anfrage bereits ein partielles Prozessmodell erstellt wurde, das einige wichtige Bestandteile enthält, aber noch nicht vollständig alle notwendigen Prozesselemente zu einem konsistenten Prozessmodell verbindet. Bergmann und Gil [7] haben einen Ansatz zum ähnlichkeitsbasierten Retrieval für das POCBR entwickelt, der auf blockorientierten und datenflussbasierten Workflows basiert. Dieses Paper hat zum Ziel, diesen Ansatz auf BPMN 2.0 zu übertragen und dadurch auszuweiten. Hierzu ist es notwendig, sowohl die BPMN-Elemente als auch die Charakteristik von BPMN als graphenorientierte Modellierungssprache bei der Ähnlichkeitsberechnung zu berücksichtigen. Die Ähnlichkeitsberechnung für Prozessmodelle erfordert darüber hinaus ein lokales Ähnlichkeitsmaß zum Vergleich von Knoten und Kantenlabels der Prozessmodelle. Dies ist für bestehende BPMN-Modelle schwierig, da diese Labels überwiegend aus Freitexten bestehen. Ein sorgfältig modelliertes Ähnlichkeitsmaß ist entscheidend für eine gute Retrievalqualität, jedoch ist dessen Entwicklung mit großem Aufwand verbunden. In dieser Arbeit werden daher die Ergebnisse eines einfachen lexikalischen Ähnlichkeitsmaßes mit denen eines semantischen Ähnlichkeitsmaßes im Rahmen einer empirischen Untersuchung verglichen.

2 Block- vs. graphenorientierte Prozessmodelle

Blockorientierte Modellierungssprachen definieren den Kontrollfluss über strukturierte Aktivitäten, durch deren Verschachtelung komplexe Abläufe mit Sequenzen, Alternativen und Schleifen erstellt werden können [12]. Ein Beispiel für blockorientierte Modellierungssprachen ist BPEL [12]. Graphenorientierte Modellierungssprachen wie BPMN definieren den Kontrollfluss über Kanten, die die zeitlichen und logischen Abhängigkeiten zwischen Knoten repräsentieren [12]. Knoten können Start- und Endzustände sowie Aktivitäten oder Konnektoren sein. Konnektoren lassen sich wiederum in *AND*, *OR* und *XOR* unterteilen, welche den Kontrollfluss verzweigen (*SPLIT*) oder vereinigen (*JOIN*) können. Graphenorientierte Modellierungssprachen gelten als ausdrucksstärker als blockorientierte, da sie Abfolgen von Aktivitäten durch die Verkettung von Knoten mit gerichteten Kanten darstellen können [13]. Dies ermöglicht beispielsweise die Kon-

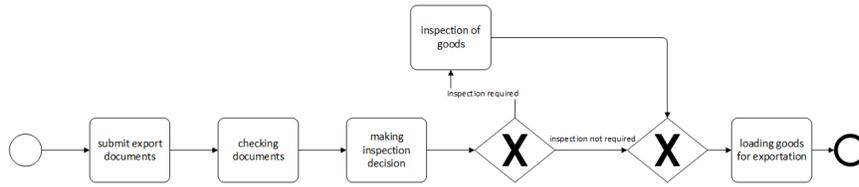


Abb. 1. BPMN-Prozessmodell aus der BPMAI

struktur von zyklischen Kantenfolgen mit mehreren Ein- und Ausstiegspunkten, was in blockorientierten Sprachen nicht realisierbar ist [12]. Abbildung 1 zeigt ein BPMN-Prozessmodell aus der BPM Academic Initiative (BPMAI) [11]. Das Prozessmodell beschreibt einen Exportvorgang im Güterverkehr. Das Beispiel zeigt die Steuerung des Kontrollflusses über die Kanten des Prozessmodells (“inspection required” und “inspection not required”). BPMN besitzt verschiedene grafische Elemente, welche sich in *Flow Objects*, *Connecting Objects*, *Pools/Swimlanes* und *Artifacts* unterteilen [3]. *Flow Objects* sind die Knoten des Prozessmodells; dazu zählen Aktivitäten, Konnektoren und Ereignisse. Aktivitäten werden als Rechteck dargestellt und beschreiben die Aufgaben eines Geschäftsprozesses. Konnektoren werden als ein auf der Spitze stehendes Quadrat dargestellt. Ereignisse (wie z. B. Start des Exportprozesses) haben unterschiedliche Symbole und werden kreisförmig dargestellt. Da das Konzept von Bergmann und Gil [7] ebenfalls graphenbasiert ist und somit die Definition von Knoten und Kanten unterstützt, lässt sich die grundlegende Struktur von BPMN Diagrammen ohne Probleme abbilden.

3 Repräsentation von BPMN-Modellen als Prozessgraphen

Wir beschreiben nun eine graphbasierte Repräsentation von BPMN-Modellen, die sich am Konzept von Bergmann und Gil [7] orientiert, jedoch einige BPMN 2.0-spezifische Erweiterungen hinzufügt. Ein Geschäftsprozess wird als semantisch annotierter gerichteter Graph dargestellt. Ein solcher *semantischer Prozessgraph* W ist ein Quadrupel $W = (N, E, S, T)$, bei dem N die Knoten und $E \subseteq N \times N$ die Kanten des Prozessmodells repräsentieren. $S : N \cup E \rightarrow \Sigma$ assoziiert zu jedem Knoten und jeder Kante eine semantische Beschreibung aus einer semantischen Metadatensprache Σ . Die Verwendung einer bestimmten Sprache für Σ ist nicht festgelegt; einzige Bedingung ist, dass ein Ähnlichkeitsmaß für diese Sprache erstellt werden kann. In dieser Arbeit wird für Σ eine Taxonomie der Knoten und Kantenbeschreibungen verwendet. $T : N \cup E \rightarrow \Omega$ weist jedem Knoten und jeder Kante einen der folgenden Typen aus Ω zu:

- Jeder Prozess besitzt genau einen *Prozessknoten*. Dieser enthält allgemeine Informationen, die den gesamten Prozess betreffen, wie z. B. eine semantische Beschreibung in Form von Tags, einer Einordnung in eine Ontologie oder Kennzahlen zur Performance, Qualität und Zuverlässigkeit des Prozessmodells.
- Jede Aufgabe eines Prozesses wird durch einen *Aufgabenknoten* repräsentiert, dessen Aufgabe durch die semantische Beschreibung näher spezifiziert wird. Der Aufgabenknoten bietet außerdem die Möglichkeit, einen Adressaten zur Ausführung

der Aufgabe festzulegen. Dies ist eine Übertragung der Rollenverteilung, die in BPMN üblicherweise über Pools und Lanes realisiert wird.

- Datenobjekte werden durch *Datenknoten* repräsentiert. Die semantische Beschreibung dieser Knoten dient der Klassifizierung, z. B. durch eine Datentypenontologie. Datenmodellierung steht jedoch nicht im Mittelpunkt von BPMN 2.0, was sich auch an der überschaubaren Anzahl von Datenobjekten (einfaches Datenobjekt, Datenlistenobjekt, Dateninput, Datenoutput, Datenspeicher) zeigt. Außerdem werden die Zuordnung von Verantwortlichkeiten repräsentiert, um Pools und Lanes zu berücksichtigen.
- Kontrollflussobjekte, wie zum Beispiel das XOR-Gateway oder das Parallele Gateway, werden durch *Kontrollflussknoten* repräsentiert.

Zusätzlich sind folgende Kantentypen möglich:

- Der Prozess-Knoten ist mit jedem anderen Knoten mit einer *Part-Of-Kante* verbunden. Die semantische Beschreibung einer solchen Kante beschreibt die Rolle des betreffenden Knotens im Prozessmodell.
- Der Kontrollfluss zwischen Aufgabenobjekten wird durch *Kontrollflusskanten* repräsentiert. Kontrollflusskanten verbinden entweder zwei Aufgabenknoten oder einen Aufgabenknoten mit einem Kontrollflussknoten. Kontrollflusskanten legen die Reihenfolge fest, in der Knoten ausgeführt werden. Der Startknoten einer Kante muss immer vor dem Endknoten der Kante ausgeführt werden.
- Der Datenfluss zwischen Aufgaben- und Datenobjekten wird durch *Datenflusskanten* repräsentiert. In BPMN 2.0 existieren verschiedene Elemente, um Datenfluss zu modellieren, wie zum Beispiel gerichtete und beidseitige Assoziation oder Nachrichtenfluss. Die semantische Beschreibung der Kante legt fest, um welches Element es sich handelt. Für gerichtete Datenflüsselemente gilt: Ist der Startknoten der Kante ein Datenknoten und der Endknoten ein Aufgabenknoten, liegt ein lesender Zugriff vor, im umgekehrten Fall ist es ein schreibender Zugriff.

Abbildung 2 zeigt das Prozessmodell aus dem vorherigen Abschnitt, dargestellt als Prozessmodellgraph. Aufgabenknoten sind eckig dargestellt, Kontrollflussknoten haben eine ovale Form, *n1* ist der Prozess-Knoten.

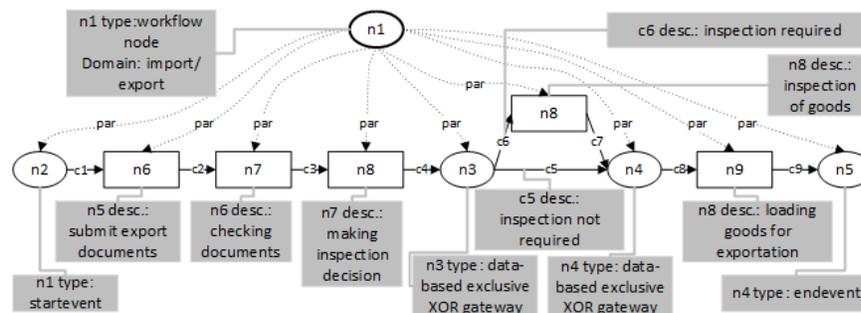


Abb. 2. Beispielhafter Prozessgraph

4 Ähnlichkeit von BPMN-Modellen

Eine Fallbasis (Repository), der Kern eines POGBR-Ansatzes, entspricht einer Menge von Prozessmodellgraphen $CB = \{CW_1, \dots, CW_n\}$, die dieselbe semantische Metadaten-sprache Σ besitzen. Auch eine Anfrage ist ein solcher Prozessmodellgraph, der aber nicht vollständig spezifiziert sein muss, d. h. es können nur einige Aktivitäten oder Datenknoten angegeben werden; auch die Verknüpfung dieser Kanten muss nicht vollständig sein oder kann auch ganz fehlen. Aufgabe des Retrievals im POGBR ist es nun, die ähnlichsten Prozessmodellgraphen zur Anfrage zu finden.

Um die Ähnlichkeit eines Prozesses aus der Fallbasis (Case) $CW = (N_c, E_c, S_c, T_c)$ zu einer Anfrage (Query) $QW = (N_q, E_q, S_q, T_q)$ beurteilen zu können, müssen sowohl die einzelnen Prozesselemente als auch die Verknüpfungsstruktur berücksichtigt werden. Wir verwenden in dieser Arbeit das Ähnlichkeitsmodell von Bergmann und Gil [7], das eine Erweiterung des Lokal/Global-Ansatzes für strukturelles CBR [8,6] darstellt. Das lokale Ähnlichkeitsmaß bestimmt die Ähnlichkeit zwischen zwei Knoten oder Kanten. Die globale Ähnlichkeit wird durch eine Aggregationsfunktion berechnet, die die lokalen Ähnlichkeitswerte unter Berücksichtigung eines geeigneten Mappings zwischen Query und Prozessmodell zusammenfasst.

In dieser Arbeit werden zwei Ansätze zur Berechnung der lokalen Ähnlichkeiten verglichen. Zum einen wird eine lexikalische Ähnlichkeitsfunktion auf Basis der Levenshtein-Distanz der Knoten- und Kantenbeschreibungen verwendet und zum anderen kommt die semantische Ähnlichkeitsfunktion, basierend auf den semantischen Beschreibungen der Knoten und Kanten, zum Einsatz. Die lexikalische Ähnlichkeit ist domänenunabhängig, wohingegen die semantische Ähnlichkeit für den jeweiligen Gegenstandsbereich spezifisch modelliert werden muss. In dieser Arbeit wird für die Knoten und Kanten jeweils eine Taxonomie der Knoten- und Kantenbeschreibungen (Ontologie) genutzt, die händisch für den untersuchten Teilbereich des Repositories modelliert wurden. Abbildung 3 zeigt einen Teilausschnitt der Knotentaxonomie. Die Ähnlichkeit basiert nun wie folgt auf dieser Taxonomie: Die Ähnlichkeit zweier Beschreibungen hat den Wert 1, wenn diese identisch sind; ansonsten wird die Ähnlichkeit dadurch bestimmt, dass der gemeinsame Oberknoten in der Taxonomie ermittelt wird. Dieser wird mit einem festen Ähnlichkeitswert annotiert, der diese Ähnlichkeit auf einfache Weise repräsentiert (z. B. $sim(\text{“start investigation”}, \text{“x-ray inspection”}) = 0.8$, siehe [5]). Liegen zwei Knoten unterschiedlichen Typs vor, so ist die Ähnlichkeit grundsätzlich 0.

Bei der Berechnung der Kantenähnlichkeit werden nicht nur die semantischen Beschreibung der betrachteten Kanten berücksichtigt, sondern auch die Ähnlichkeit der Knoten, die diese verbinden. Zwei Kontrollflusskanten sollten nur ähnlich sein, wenn diese auch ähnliche Aufgabenknoten als Start- und Endknoten verbinden. Die Funktion $F_E(S_e, S_l, S_r)$ ist eine Aggregationsfunktion, die die semantische Ähnlichkeit der Kanten S_e sowie der Startknoten S_l und Endknoten S_r zu einem Ähnlichkeitswert zusammenfasst. Für F_E wird folgende Funktion verwendet: $F_E(S_e, S_l, S_r) = S_e \cdot 0,5 \cdot (S_l + S_r)$. Die Kantenähnlichkeit und die Ähnlichkeiten der verbundenen Knoten gehen also jeweils zur Hälfte in die Berechnung der aggregierten Kantenähnlichkeit ein.

Die Grundlage der Ähnlichkeitsberechnung zwischen einer Query QW und einem Prozessmodell CW bildet ein zulässiges Mapping m [7]. Ein zulässiges Mapping ist

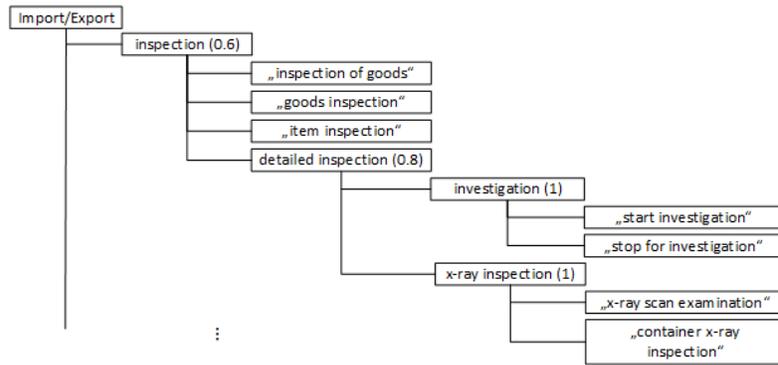


Abb. 3. Teilausschnitt der Import/Export-Taxonomie

eine typerhaltende, partielle Abbildung $m : N_q \cup E_q \rightarrow N_c \cup E_c$. Ein solches Mapping muss die Bedingungen erfüllen, dass ein Knoten oder eine Kante aus QW jeweils nur auf einen Knoten oder eine Kante aus CW abgebildet wird. Außerdem können Kanten nur zugewiesen werden, wenn die von der Kante verbundenen Knoten auch entsprechend als Start- und Zielknoten der Zielkante zugewiesen wurden. Für ein gegebenes Mapping können somit die Elemente der Query auf die Elemente im Prozessmodell abgebildet werden, so dass nun eine Ähnlichkeitsberechnung der zugeordneten Elemente möglich ist. Die resultierenden lokalen Ähnlichkeitswerte können dann zur Gesamtähnlichkeit aggregiert werden. Um die Ähnlichkeit zu bestimmen, ist somit die Kenntnis eines Mappings erforderlich. Diese kann gemäß des von Bergmann und Gil [7] beschriebenen heuristischen Suchverfahrens so bestimmt werden, dass die resultierende Ähnlichkeit maximiert wird.

5 Empirische Evaluation

Ziel der Evaluation ist es herauszufinden, ob das Ranking von Prozessmodellen durch die beschriebene Ähnlichkeitsberechnung vergleichbare Ergebnisse liefert, wie das Ranking von Experten. Des Weiteren werden die Ergebnisse des semantischen Ähnlichkeitsmaßes mit denen des lexikalischen Ähnlichkeitsmaßes verglichen. Zur Erstellung der Fallbasis wurden die Prozessmodelldatenbanken der BPMMAI [11] analysiert. Um eine einheitliche Fallbasis von hoher Qualität zu erreichen wurden nur Prozessmodelle betrachtet, die eine detaillierte Knotenbeschreibung in englischer Sprache beinhalten und vollständig sind (Start- und Endpunkt sind vorhanden, es gibt keine offenen Abzweigungen). Eine detaillierte Beschreibung der Aktivitätsknoten ist von großer Bedeutung, da die Ähnlichkeitsberechnung auf diesen aufbaut. Die gefundenen Prozessmodelle wurden nach Domänen gruppiert. Die Fallbasis besteht aus den ausgewählten Domänen *Import/Export*, *Abrechnungen* und *Bestellprozesse*, wobei jede Domäne 15-20 Prozessmodelle enthält.

Zu jeder Domäne wurden fünf plausible Suchanfragen händisch formuliert. Die Anfragen sind meist kleine, leicht abgeänderte Teilabschnitte der Prozessmodelle aus der

Fallbasis. Zu jeder dieser Anfragen wurden von Experten (Studierende der Wirtschaftsinformatik) die fünf ähnlichsten Prozessmodelle aus der Fallbasis ausgesucht und nach Ähnlichkeit sortiert. Die Ranglisten der Experten wurden mit den Ergebnissen der Ähnlichkeitsberechnungen verglichen. Die Bewertung der Ähnlichkeit erfolgt mit Hilfe der Evaluationskriterien *Korrektheit* und *Vollständigkeit* nach der Ranglistenberechnung von Cheng et al. [9]. Der Wert für die Korrektheit liegt im Intervall [-1,1]. Sofern beide Ranglisten übereinstimmen, ist der Wert 1. Wenn beide Ranglisten gegensätzlich sind, liegt der Wert bei -1. Der Wert für die Vollständigkeit bewegt sich im Intervall [0,1] und gibt an, wie viele Ordnungen des Expertenrankings durch das Ähnlichkeitsmaß ebenfalls geordnet wurden. Die Tabelle 1 zeigt die Ergebnisse der Evaluation für die ausgewählten Domänen. Die dargestellten Ergebnisse für Korrektheit und Vollständigkeit sind jeweils das arithmetische Mittel aus den Einzelergebnissen der fünf Anfragen pro Domäne.

Ähnlichkeitsmaß	Import/Export		Bestellungen		Abrechnungen	
	Korr.	Voll.	Korr.	Voll.	Korr.	Voll.
lexikalisch	0,662	1,000	0,596	1,000	0,351	1,000
semantisch	0,751	1,000	0,796	0,793	0,770	0,860

Tabelle 1. Evaluation der Ähnlichkeitsmaße

Die Korrektheit des semantischen Ähnlichkeitsmaßes liegt in allen Domänen über dem des lexikalischen Maßes, am deutlichsten in der Domäne *Abrechnungen*, in der das lexikalische Ähnlichkeitsmaß mit einem Wert von 0.351 kaum zu brauchbaren Ergebnissen führt. Die Vollständigkeit ist hingegen beim lexikalischen Ähnlichkeitsmaß in allen Domänen maximal, was darin begründet ist, dass aus dem Levenshtein-Vergleich von zwei Knotenlabels sehr differenzierte Ähnlichkeitswerte resultieren, wohingegen beim semantischen Ähnlichkeitsmaß nur die in der Taxonomie hinterlegten Ähnlichkeitswerte vorkommen. Somit bewertet das lexikalische Maß nur selten zwei Prozessmodelle in der Fallbasis mit der gleichen Ähnlichkeit, sodass grundsätzlich zwischen allen Fällen eine Ordnungsrelation besteht. Auch wenn diese nicht korrekt ist, führt dies zu einer Vollständigkeit von 1. Beim semantischen Ähnlichkeitsmaß kommt es hingegen vor, dass zwei Prozessmodelle als gleich ähnlich eingestuft werden, obwohl Experten dies nicht so bewerten. Ein stärker differenzierendes semantisches Ähnlichkeitsmaß könnte die Ergebnisse für die Vollständigkeit verbessern. Insgesamt ist zur Einschätzung des praktischen Nutzens aber die Korrektheit ein wichtigeres Kriterium als die Vollständigkeit. Bei Verwendung des semantischen Ähnlichkeitsmaßes wird daher eine deutlich bessere Übereinstimmung mit der Einschätzung der Experten erzielt.

6 Fazit und Ausblick

Diese Arbeit setzt an der aktuellen Problemstellung an, Prozessdatenbanken mittels Anfragen nach ähnlichen Prozessmodellen zu durchsuchen, um die Modellierung von neuen Prozessmodellen zu erleichtern. Der vorgestellte Ansatz auf Basis von prozessorientiertem Fallbasierten Schließen eignet sich zur Ähnlichkeitsberechnung und zum Retrieval von Prozessmodellen im BPMN-2.0-Format. Die Einbeziehung der Semantik

von Prozessmodellen hat sich als gute Möglichkeit erwiesen, um die Retrievalqualität des Systems gegenüber einer rein syntaktischen Auswertung zu steigern. Dies wird jedoch erkauft durch den höheren Aufwand zur Modellierung der Ähnlichkeitsmaße mittels Taxonomien. Die manuelle Erstellung von Taxonomien könnte sich bei großen Repositorien auf Grund des Modellierungsaufwands als impraktikabel erweisen. Als Ansatzpunkt für weitere Forschungsarbeiten bietet es sich daher an, alternative Verfahren zu untersuchen, die keine manuelle Modellierung erfordern, um die semantischen Ähnlichkeiten zwischen Labels in Prozessmodellen zu bestimmen. Es könnte beispielsweise ein semantisches Vektorraummodell zur Ähnlichkeitsberechnung genutzt werden, das zusätzlich durch Fachtexte domänenspezifisch angereichert werden könnte.

Literatur

1. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications* 7(1), 39–59 (1994)
2. Allweyer, T.: BPMN-Prozessmodelle und Unternehmensarchitekturen. Untersuchung von Ansätzen zur Methodenintegration und ihrer Umsetzung in aktuellen Modellierungstools. Forschungsbericht, Hochschule Kaiserslautern, <http://www.kurze-prozesse.de/blog/wp-content/uploads/2014/11/BPMNProzessmodelle-und-Unternehmensarchitekturen.pdf>
3. Allweyer, T.: BPMN 2.0: introduction to the standard for business process modeling. BoD–Books on Demand (2010)
4. Becker, M., Laue, R.: A comparative survey of business process similarity measures. *Computers in Industry* 63(2), 148 – 167 (2012)
5. Bergmann, R.: On the Use of Taxonomies for Representing Case Features and Local Similarity Measures. In: Gierl, L., Lenz, M. (eds.) *Proceedings of the 6th German Workshop on Case-Based Reasoning (GWCBR'98)* (1998)
6. Bergmann, R.: *Experience management: foundations, development methodology, and internet-based applications*. Springer-Verlag (2002)
7. Bergmann, R., Gil, Y.: Similarity assessment and efficient retrieval of semantic workflows. *Information Systems* 40, 115–127 (2014)
8. Burkhard, H.D., Richter, M.M.: On the notion of similarity in case based reasoning and fuzzy theory. In: *Soft computing in case based reasoning*, pp. 29–45. Springer (2001)
9. Cheng, W., Rademaker, M., De Baets, B., Hüllermeier, E.: Predicting partial orders: ranking with abstention. In: *Machine Learning and Knowledge Discovery in Databases*, pp. 215–230. Springer (2010)
10. Koschmider, A., Fellmann, M., Schoknecht, A., Oberweis, A.: Analysis of process model reuse: Where are we now, where should we go from here? *Decision Support Systems* 66, 9 – 19 (2014)
11. Kunze, M., Berger, P., Weske, M.: BPM academic initiative - fostering empirical research. In: *Proceedings Demonstration Track - 10th International Conference on Business Process Management*. CEUR Workshop Proceedings, vol. 940, pp. 1–5. CEUR-WS.org (2012)
12. Mendling, J., Lassen, K.B., Zdun, U.: On the transformation of control flow between block-oriented and graph-oriented process modeling languages. *International Journal of Business Process Integration and Management* 3(2), 96–108 (2006)
13. Mendling, J., Reijers, H.A., van der Aalst, W.M.: Seven process modeling guidelines (7pmg). *Information and Software Technology* 52(2), 127–136 (2010)
14. Minor, M., Montani, S., Recio-Garcia, J.A.: Process-oriented case-based reasoning. *Information Systems* 40, 103 – 105 (2014)
15. Richter, M.M., Weber, R.O.: *Case-Based Reasoning - A Textbook*. Springer (2013)

Assessing the Quality of Unstructured Data: An Initial Overview

Cornelia Kiefer

Graduate School of Excellence Advanced Manufacturing Engineering
Nobelstr. 12, 70569 Stuttgart, Germany
cornelia.kiefer@gsame.uni-stuttgart.de
<http://www.gsame.uni-stuttgart.de>

Abstract. In contrast to structured data, unstructured data such as texts, speech, videos and pictures do not come with a data model that enables a computer to use them directly. Nowadays, computers can interpret the knowledge encoded in unstructured data using methods from text analytics, image recognition and speech recognition. Therefore, unstructured data are used increasingly in decision-making processes. But although decisions are commonly based on unstructured data, data quality assessment methods for unstructured data are lacking. We consider data analysis pipelines built upon two types of data consumers, human consumers that usually come at the end of the pipeline and non-human / machine consumers (e.g., natural language processing modules such as part of speech tagger and named entity recognizer) that mainly work intermediate. We define data quality of unstructured data via (1) the similarity of the input data to the data expected by these consumers of unstructured data and via (2) the similarity of the input data to the data representing the real world. We deduce data quality dimensions from the elements in analytic pipelines for unstructured data and characterize them. Finally, we propose automatically measurable indicators for assessing the quality of unstructured text data and give hints towards an implementation.

Keywords: quality of unstructured data, quality of text data, data quality dimensions, data quality assessment, data quality metrics

1 Introduction

In recent years the methods for knowledge extraction from unstructured data have improved and unstructured data sources such as texts, speech, videos and pictures have gained importance. Nowadays, sentiment analysis of social media data leads to decisions in marketing campaign design, images are classified automatically and unstructured information can be retrieved easily using search engines [6, 19]. But methods which determine the quality of the data are lacking. To be able to make good decisions, the quality of the underlying data must be determined. Similar to the concepts, frameworks and systems developed for structured data we need means to ensure high quality of unstructured data. We

focus on data consumers of unstructured data and define them as humans or non-humans / machines (e.g. algorithms) that are using or processing data. The quality of the data consumed by the final consumer such as a human who needs to derive a decision from the data, depends on the quality assessed for earlier consumers. This is especially true for unstructured data, which is analyzed in a pipeline.

The remainder of this paper is organized as follows: First we motivate research in assessing the quality of unstructured data in section 2. In section 3 we define data quality of unstructured data. Furthermore, we describe the data quality dimensions interpretability, relevance and accuracy. Based on this, in section 4 we present data quality indicators for unstructured text data. In section 5 we discuss related work and finally conclude the work and highlight future work in section 6.

2 Motivation

Low data quality is dangerous because it can lead to wrong or missing decisions, strategies and operations. It can slow down innovation processes, and losses for organizations caused by low data quality are estimated to lie over billions of dollars per year [8]. Bad data is a huge problem: 60% of enterprises suffer from data quality issues, 10-30% of data in organizational databases is inaccurate and individual reports of incomplete, inaccurate and ambiguous organizational data are numerous [13, 18].

The most important information sources in organizations, such as the workers, managers and customers produce unstructured data. About 90% of all data outside of organizations and still more than 50% inside are estimated to be unstructured [20]. In the era of Big Data the amount of data is increasing immensely and filtering relevant and high quality data gets more and more important. Organizations need to leverage the information hidden in unstructured data to stay competitive [14]. Therefore, the quality of texts, pictures, videos and speech data needs to be ensured. But while the need for data quality assessment and improvement strategies for unstructured data was recognized (e.g. [2, 23]) no concrete approach to assessing the quality of unstructured data was suggested yet. We fill this gap and provide data quality dimensions and executable indicators for unstructured data. By focusing on automatically calculable indicators of data quality, we aim to support real time analytics of stream data (such as social media data) with real time data quality assessment techniques, both running concurrently.

3 Definition of Data Quality and of Data Quality Dimensions for Unstructured Data

The definitions of data quality in [24, 30] focus on structured data which is consumed by humans. They define data quality via the similarity of the data

D to the data set D' which is expected by the data consumer [24] and via the fitness for use by the data consumer [30]. We extend the meaning of these existing definitions by pointing out that machine consumers and many different consumers in a pipeline need to be considered as well as human end consumers in the case of unstructured data. Furthermore, data quality needs to be defined in terms of accuracy. Accuracy describes the similarity between the input data and the data which would be representing the real world. This definition of Accuracy is equal to exiting ones, e.g. [11].

The quality of data has a multi-faceted nature and many lists of data quality dimensions and indicators for structured data exist (see 5). All of the dimensions that were found to be relevant in the literature, such as completeness, timeliness and accuracy are relevant to structured as well as unstructured data. From these dimensions we selected three dimensions which are relevant to mining processes on unstructured data.

We deduce the dimensions from the elements involved in mining processes on unstructured data: The input data, the real world, data consumers, a task and the knowledge extracted. Based on these elements, the quality of data D can be determined by comparing it to three classes of ideal data sets: the data as expected by the current data consumer D_C (we will call this the Interpretability dimension), the data as it would be optimal for the task D_T (Relevancy) and the data set which is representing the real world D_W (Accuracy). The deduced dimensions are also in line with the data quality definitions stated above. In Fig. 1, we illustrate the three data sets in the context of an ideal mining process on unstructured data. Ideally, D would match the real world D_W and would be exactly the same as the data expected by the first data consumer. Since unstructured data is analyzed in a pipeline, the output of the first data consumer is input to the second and should therefore match the data expected by the second data consumer and so on (as indicated in Fig. 1 with the analysis pipeline). An ideal result of the mining process can be D_T (which is still bound to D , D_W and D_C and is usually equal to the data expected by the final consumer). By basing the data quality dimensions on the elements involved in a mining process on unstructured data, we focus on the quality of unstructured data which is analyzed automatically in analytics pipelines.

In the following, we describe the deduced data quality dimensions in more detail:

Interpretability can be assessed as the degree of similarity between D and D_C . For example, consider a statistical preprocessor which is used to segment a text into sentences. If it was trained on Chinese texts and is used to segment English texts, D and D_C are not similar and data quality is low. Since often many different data consumers are involved in interpreting unstructured data, this dimension is crucial for unstructured data.

Relevancy can be assessed as the similarity between D and D_T . Usually D_T will be very similar to the D_C of the end consumer (which we will call D_{CE}) who wants to use the data to accomplish the task. While differences between D_T and the data expected by the end consumer D_{CE} indicate problems, these

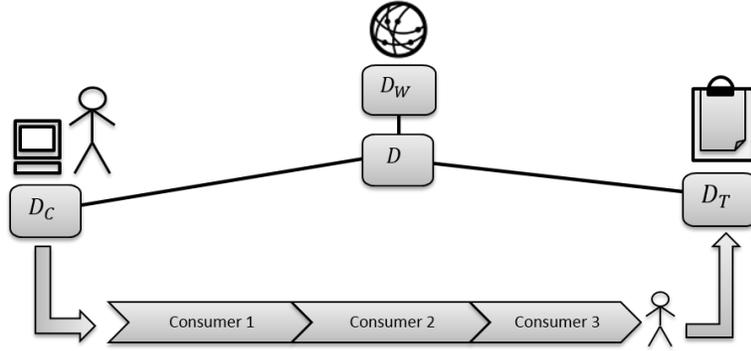


Fig. 1. The three ideal data sets D_C , D_T and D_W in the context of an ideal mining process on unstructured data

are not related to data quality and we will therefore assume D_T and D_{CE} to be equivalent. As an example for relevancy, consider a worker on the shop floor who is searching for a solution for an urgent problem with a machine in a knowledge base. If he only finds information on the price of the machine, the data quality of the result is low because it does not help him with his task of solving the problem.

We assess the **Interpretability and Relevancy** of a data set D by its similarity to the data set D_C and D_{CE} which is expected by the data consumers. Expectations differ from human to machine consumers. What a human data consumer expects, depends on factors such as his knowledge, experiences and goals. Expectations of machine consumers are very precise and depend on the algorithm, training data, statistical models, rules and knowledge resources available. This holds for all types of unstructured data. As illustrated in Fig. 2, unstructured data such as textual documents may be consumed by machines or humans and the data set D_C or D_{CE} depends on factors such as the native language of the human and the statistical language models available to the machine. For example, a human data consumer expects a manual for a machine to be in his native language or in a language he knows. He also expects the manual to explain the machine in a way he understands with his technical expertise. When a machine consumes unstructured data, similar factors influence the interpretability and more precisely the similarity of the input data and the data expected. The knowledge of a machine consumer can be represented by machine-readable domain knowledge encoded in semantic resources (such as taxonomies), by training data, statistical models or by rules. As an example, imagine a machine consumer that uses a simple rule-based approach to the extraction of proper names from German text data, where all uppercased words are extracted. This machine consumer expects a data set D_C with correct upper and lowercased words. If D is

all lower-cased, D_C and D are not similar and the data is not fit for use by that data consumer.

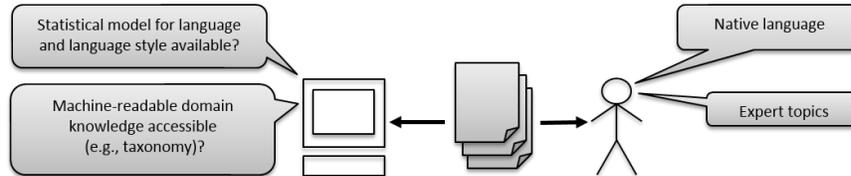


Fig. 2. Machine and human data consumer and factors that influence the data expected

Unstructured data is usually consumed by many different data consumers with many different data sets D_C expected. In an analytics pipeline, the raw data is consumed and processed by several consumers in a row and the output of the previous consumer is the input to the next consumer and so on. Data quality problems at intermediate consumers may be automatically propagated to following consumers. By considering all intermediate (machine and/or human) consumers, the exact points for data quality improvement can be determined. In Fig. 3 we illustrate an analytics pipeline involving three machine consumers and one human end consumer of the data. Machine consumers are in this illustration represented by three high level machine consumers which are present in many analytic pipelines of unstructured data: preprocessors, classifiers and visualizers. For example, as depicted in Fig. 3, the output of the preprocessor is input to automatic classification and the results are then visualized. The visualizations are finally the input to a human consumer of the data, who e.g., derives decisions from it.

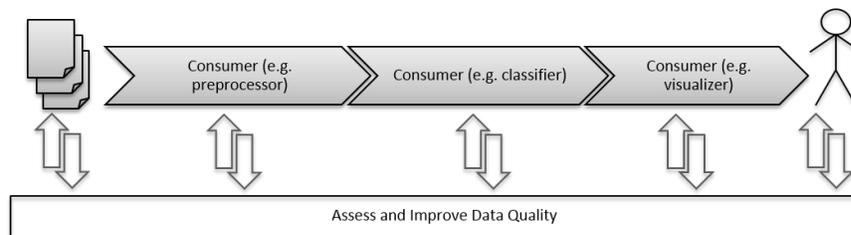


Fig. 3. Assessing and Improving data quality for each data consumer on the way from e.g., raw text documents to final consumer

As for structured data, the **Accuracy** of data and information is a very important data quality dimension. It is hard to measure, because the data set

D_W , which represents the real world, is often not known and creating it involves the work of human experts, is time-consuming, costly or even impossible. The solution is usually to abstract away from details e.g., by using rules to check general conformance of data points with expected patterns (e.g., e-mail addresses containing an @ sign) or to build D_W manually for a part of the data set only (see [28, 29]). D_W may be represented by a so-called gold standard data set with the accurate values annotated manually by human experts. For example, statistical classifiers are evaluated by comparing the prediction of the statistical classifier with those in a gold standard with manually annotated classes. Since D_W is not known for all data sets D , many statistical classifiers can not be evaluated and the number of problems with accuracy in big data bases can only be approximated.

4 Data Quality Indicators for Unstructured Text Data

A data quality dimension can be measured by exploitation of data quality indicators. Data quality indicators must be transferable to a number in the interval $[0,1]$ where 0 indicates low data quality and 1 indicates high data quality (this is similar to the standard characterizations of data quality metrics, such as in [1]). Therefore, indicators can e.g., be represented by yes/no-questions, proportions of data items which have a certain characteristic or by evaluation metrics. The standard approaches to more concrete indicators for the quality of structured data involve counting the number of missing values, wrong values or the number of outliers. For the case of unstructured data, different indicators are needed. We compiled an extensive list of indicators for all three dimensions. The definition of indicators is based on the dimensions discussed in the previous section and on related work in natural language processing, information retrieval, automated assessment and machine learning (see section 5.2). Here, we limit the indicators presented to those which are (1) automatically measurable and (2) applicable to unstructured text data. Furthermore, we selected indicators, which we already implemented or which are straightforward to implement (since libraries with good documentations are available), so that the indicators can be verified in experiments in near future work. In table 1, we describe each dimension with these more concrete indicators of data quality.

While the concept behind the indicators *confidence*, *precision*, *accuracy* and *quality of gold annotations* are applicable to all types of unstructured data which are processed by statistical machine learning components, the remaining indicators are text specific. With a different definition of *noisy data* and *fit of training data*, the concepts may be transferred to other data types as well, e.g. measuring the similarity between input pictures and training data pictures or measuring the percentage of noisy data, defined as the percentage of background noise, in speech.

In the following we describe the indicators in more detail and give hints towards possible implementations:

Table 1. Indicators for the quality of unstructured text data

Dimension	Indicator
Interpretability	Fit of training data
	Confidence
	Noisy data
Relevancy	Frequent keywords
	Specificity
Accuracy	Precision
	Accuracy
	Quality of gold annotations

The first indicator *fit of training data* directly follows from the definition for **Interpretability** we gave in section 3, when considering statistical classifiers as data consumers. The quality of text data with respect to a machine consumer, can be measured by calculating the similarity of the input text data and the data expected by the data consumer. In the case of statistical classifiers such as a part of speech tagger (which automatically assigns parts of speech to each token such as a word in a text) or sentiment classifier (which automatically detects opinions in texts and assigns e.g., the classes positive, negative and neutral to texts), D_C may be represented by the training data. For the case of unstructured text data the similarity can be measured using text similarity measures. For example, consider the situation where Twitter data is consumed by a statistical classifier such as a part of speech tagger that was trained on newspaper texts. By the definition of interpretability used in this work, data quality is lower than for another tagger that was trained on text data from Twitter as well. Examples for measures for this indicator are text similarity measures such as Cosine Similarity and Greedy String Tiling which are e.g. implemented in the DKPro Similarity package (see [7]). Using the DKPro Similarity library in Java two lists of tokens can be easily compared and a similarity score in the interval [0,1] can be calculated, following the instructions on the web site¹.

The second indicator, *confidence*, also focuses on data quality of text data as perceived from the point of view of a statistical classifier. A statistical classifier estimates the probabilities for each class from a fixed list of classes, given the data. These probabilities are also called confidence values (for more details, see [12]). If the probability of a classification decision is very high, confidence of the statistical classifier is said to be high. Confidence is expressed as a number in the interval [0,1] and may be used for measuring data quality. For example, confidence measures are available and can be retrieved for the natural language processing tools in OpenNLP² (such as the tokenizer and part of speech tagger), a Java library for natural language processing which is heavily used in industry applications because it has an Apache license. To get these confidence values, follow the documentation of the OpenNLP library (see footnote 2, e.g., for the

¹ <https://dkpro.github.io/dkpro-similarity/>

² <https://opennlp.apache.org/>

part of speech tagger, just call the *probs* method which will return an array of the probabilities for all tagging decisions).

The third indicator in the interpretability dimension is the percentage of *noisy data*. This is a relevant indicator for human and machine consumers, since reading a text is more difficult for a human if it is full of misspelled words, non-grammatical sentences and abbreviations. Since most machine consumers of text data expect clean text data such as newspaper texts, the degree of noisy data also measures data quality from the viewpoint of such standard machine consumers. The percentage of noisy data may be measured as the percentage of sentences which cannot be parsed by an automatic syntax parser, unknown words, punctuation, very long/short sentences, incorrect casing, special signs, urls, mail addresses, emoticons, abbreviations, pause filling words, rare words or by the percentage of spelling mistakes (the latter as already suggested by [26]). Non-parsable sentences can be identified using an automatic syntax parser such as the parser implemented in natural language processing libraries such as OpenNLP (see footnote 2) or the Natural Language Processing Tool Kit NLTK³. The number of punctuation and of unknown words (e.g., defined as words unknown to a standard part of speech tagger) may be e.g., calculated using the standard part of speech tagger implemented in NLTK (which has individual classes for punctuation and unknown words). Very long/short sentences can be identified using a tokenizer and a sentence segmenter from a natural language processing library and by counting the automatically determined tokens and sentences. Incorrect casing may be detected using supervised machine learning methods, such as suggested in [17]. Regular expressions can be used to automatically identify the percentage of special signs, urls, mail addresses, emoticons, abbreviations and pause filling words in texts. Rare words can be identified internally by counting all words that occur less than a specified number of times in the text corpus, by counting words that are not found in a standard dictionary or a generated dictionary (such as a dictionary generated from a very encompassing text corpus from the domain). The number of spelling mistakes in a text corpus may be calculated using the Python implementation PyEnchant⁴ or any other spelling correction module. Most of the measures suggested for the indicator noisy data can be implemented using the NLTK library which comes with very good documentation and an active community (see footnote 3).

But it is not sufficient if data is interpretable only. Interpretable data, which is not relevant to the end data consumer and his goal is of low quality. Therefore, it's **Relevancy** need to be calculated. For text data this can be done following approaches already developed for information retrieval systems. The relevance metric used in information retrieval systems determines the relevance of search results with respect to the information need of the searcher. The information need is captured via keywords or documents first and can then be compared e.g., to the *frequent keywords* in the input texts (see [16] for the relevance metric in information retrieval). Again, textual similarity measures such as cosine

³ <http://www.nltk.org/>

⁴ <http://pythonhosted.org/pyenchant/>

similarity are used to determine the similarity of the information need and a text (as implemented in [7] and accessible via the well-documented DKPro Similarity library, see footnote 4). Besides the *frequent keywords*, also specificity can indicate the relevance of unstructured text data for the task a certain end consumer wants to accomplish. The *specificity* of language in texts and speech can be determined via the coverage of a domain-specific semantic resource which contains all relevant technical terms. In the simplest version this would be a text file with all domain words listed which is used to determine the percentage of domain words in a corpus. Coverage of domain specific taxonomies may be e.g., calculated with a concept matcher such as the one presented in [22].

If the data is interpretable and relevant, the remaining question is whether it reflects the real world or not, that is whether it is accurate. The **Accuracy** of unstructured text data may be indicated by evaluation metrics such as precision and accuracy. These metrics compare the automatically annotated data to parts of the data which represent the real world, such as manually annotated gold standard corpora. Statistical classifiers are evaluated by comparing them to gold standards and by determining how many of the classified entities really belong to a class (*precision*) and the percentage of classification decisions that were correct (*accuracy*), see [16]. The metrics precision and accuracy were already suggested as indicators for text data quality by [26] and [23]. Furthermore, the *quality of gold annotations* of training and test data is an indicator in the accuracy dimension. These can be calculated according to [10] by measuring the inter-rater agreement which measures the number of times one or more annotators agree. Evaluation metrics and inter-rater metrics are e.g. implemented in NLTK (see footnote 3).

In this section we presented automatically measurable indicators for text data which are executable. Not all indicators presented here are relevant and applicable in all cases. Only few out of the many statistical tools give access to the confidence metric and only with access to gold test data precision and accuracy can be calculated.

5 Related Work

While research on the quality of structured data is numerous, the quality of unstructured data has hardly been considered yet. We present related work in the field of data quality in section 5.1 and list isolated methods useful in assessing unstructured text data quality in section 5.2.

5.1 Related Work in Data Quality

Many frameworks and data quality dimensions dedicated to the quality of structured data have been suggested (e.g. [24, 30]) and also special frameworks and dimensions for social media data and big data were developed [5, 21]. In these works, data quality dimensions are defined from a human end consumer's point of view and no automatic measures for the assessment of unstructured data are

given. Several sources [2, 23, 26] address the need for data quality measures on unstructured data but none of them gives executable dimensions and indicators. In these works, interesting starting points for quality dimensions and indicators are defined, such as:

- The quality of technologies used to interpret unstructured data and the author’s expertise [23]
- Accuracy, readability, consistency and accessibility [2]
- Precision and spelling quality [26]

No hints towards possible implementations of these dimensions and indicators are suggested, though. As demanded in [26], we also support the view that textual data quality needs to be measured for both, human consumers and machine consumers. We have furthermore motivated the need to measure data quality at every stage. This is also demanded in [15, 27]. A closely related idea is also expressed in the concept of data provenance which aims at collecting the information on all data sources and transformation or merging steps of data (see [4]).

5.2 Isolated Methods for Data Quality Assessment of Unstructured Text Data

In the definition of the quality indicators in this article we focused on unstructured text data. Therefore, we limit the list of isolated methods to those relevant for the assessment of textual data. For example, quite some work in the field of natural language processing focuses on the interaction between textual data characteristics and the performance of Natural Language Processing (NLP) tools. In [3] the authors consider factors that affect the accuracy of automatic text-based language identification (such as the size of the text fragment and the amount of training data). Furthermore, work on correcting upper and lowercasing of words in texts (re-casing), spelling correction, abbreviation expansion and text simplification is related to our work (e.g., [17]). In the context of search engines, the quality of the search results and of the data basis is discussed as well [9]. In automated assessment, methods to automatically assess the quality of hand-written essays and short answers (e.g., student essays and answers to free text questions) are developed (for a good overview, see [31]). Work on training data selection in machine learning, which is on choosing subsets of training data which fit best to the domain of the test set (e.g. [25]) is also related to our work. The idea expressed in these works is similar to the idea behind the indicator *fit of training data*, which we added to our list of indicators for unstructured text data quality. However, we are the first to suggest the fit of training data as a data quality indicator. Furthermore, we do not suggest to use it for parts of training data, as suggested in these works, but to choose from different text corpora.

6 Conclusion and Future Work

We listed dimensions and indicators for determining the quality of unstructured data based on the basic elements of mining processes on unstructured data.

The indicators proposed are executable and easily transfer into a data quality metric in the interval $[0,1]$. In future work we will determine the most suitable implementations for the indicators and validate them in experiments. We will furthermore explore how indicators may be combined to measure the overall data quality of unstructured data and how the improvement of data quality as perceived by intermediate consumers influences data quality from a rather end consumer viewpoint.

Acknowledgments. The authors would like to thank the German Research Foundation (DFG) for financial support of this project as part of the Graduate School of Excellence advanced Manufacturing Engineering (GSaME) at the University of Stuttgart. Moreover, we thank B. Mitschang and L. Kassner for important feedback.

References

1. C. Batini, D. Barone, F. Cabitza, and S. Grega. A data quality methodology for heterogeneous data. *International Journal of Database Management Systems (IJDMIS)*, 3(1):60–79, 2011.
2. C. Batini and M. Scannapieco. *Data and Information Quality*. Springer International Publishing, Cham, 2016.
3. G. R. Botha and E. Barnard. Factors that affect the accuracy of text-based language identification. *Computer Speech & Language*, 26(5):307–320, 2012.
4. P. Buneman and S. B. Davidson. Data provenance – the foundation of data quality. 2010.
5. L. Cai and Y. Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14(0):2, 2015.
6. F. Camastra and A. Vinciarelli. *Machine learning for audio, image and video analysis: Theory and applications*. Advanced Information and Knowledge Processing. Springer, London, second edition edition, 2015.
7. Daniel Bär, Torsten Zesch, and Iryna Gurevych. Dkpro similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 121–126, Stroudsburg, PA, USA, 2013. Association for Computational Linguistics.
8. D. Dey and S. Kumar. Reassessing data quality for information products. *Management Science*, 56(12):2316–2322, 2010.
9. C. Feilmayr. Decision guidance for optimizing web data quality - a recommendation model for completing information extraction results. *24th International Workshop on Database and Expert Systems Applications*, pages 113–117, 2013.
10. Fleiss and Levin. The measurement of interrater agreement. In J. L. Fleiss, B. Levin, and M. C. Paik, editors, *Statistical methods for rates and proportions*, Wiley series in probability and statistics, pages 598–626. J. Wiley, Hoboken, N.J., 2003.
11. C. Fox, A. Levitin, and T. Redman. The notion of data and its quality dimensions. *Inf. Process. Manage.*, 30(1):9–19, 1994.

12. S. Gandrabur, G. Foster, and G. Lapalme. Confidence estimation for nlp applications. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(3):1–29, 2006.
13. J. Han, K. Chen, and J. Wang. Web article quality ranking based on web community knowledge. *Computing*, 97(5):509–537, 2015.
14. K. Hartl and O. Jacob. Determining the business value of business intelligence with data mining methods. *The Fourth International Conference on Data Analytics*, pages 87–91, 2015.
15. A. Immonen, P. Paakkonen, and E. Ovaska. Evaluating the quality of social media data in big data architecture. *IEEE Access*, (3):1, 2015.
16. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, New York, 2008.
17. C. Niu, W. Li, J. Ding, and R. K. Srihari. Orthographic case restoration using supervised learning without manual annotation. *International Journal on Artificial Intelligence Tools*, (13), 2003.
18. J. R. Nurse, S. S. Rahman, S. Creese, M. Goldsmith, and K. Lamberts. Information quality and trustworthiness: A topical state-of-the-art review. *International Conference on Computer Applications and Network Security (ICCANS 2011)*, 2011.
19. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.
20. P. Russom. Bi search and text analytics: New additions to the bi technology stack. 2007.
21. M. Schaal, B. Smyth, R. M. Mueller, and R. MacLean. Information quality dimensions for the social web. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pages 53–58. ACM, 2012.
22. M. Schierle and D. Trabold. Multilingual knowledge-based concept recognition in textual data. In A. Fink, B. Lausen, W. Seidel, and A. Ultsch, editors, *Advances in Data Analysis, Data Handling and Business Intelligence*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 327–336. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
23. A. Schmidt, C. Ireland, E. Gonzales, M. Del Pilar Angeles, and D. D. Burdescu. On the quality of non-structured data, 2012.
24. L. Sebastian-Coleman. *Measuring data quality for ongoing improvement: A data quality assessment framework*. Elsevier Science, Burlington, 2013.
25. Y. Song, P. Klassen, F. Xia, and C. Kit. Entropy-based training data selection for domain adaptation. *Proceedings of COLING 2012*, 2012.
26. D. Sonntag. Assessing the quality of natural language text data. In *GI Jahrestagung*, pages 259–263, 2004.
27. I.-G. Todoran, L. Lecornu, A. Khenchaf, and J.-M. Le Caillec. A methodology to evaluate important dimensions of information quality in systems. *Journal of Data and Information Quality*, 6(2-3):1–23, 2015.
28. T. Vogel, A. Heise, U. Draisbach, D. Lange, and F. Naumann. Reach for gold. *Journal of Data and Information Quality*, 5(1-2):1–25, 2014.
29. H. Wang, M. Li, Y. Bu, J. Li, H. Gao, and J. Zhang. Cleanix. *ACM SIGMOD Record*, 44(4):35–40, 2016.
30. R. Y. Wang and D. M. Strong. Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33, 1996.
31. R. Ziai, N. Ott, and D. Meurers. Short answer assessment: Establishing links between research strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, Montreal, Canada, 2012. Association for Computational Linguistics.

An Interactive e-Government Question Answering System

Malte Schwarzer¹, Jonas Düver¹, Danuta Ploch², and Andreas Lommatzsch²

¹ Technische Universität Berlin, Straße des 17. Juni, D-10625 Berlin, Germany
{malte.schwarzer, jonas.duever}@campus.tu.berlin.de}

² DAI-Labor, TU Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany
{danuta.ploch, andreas.lommatzsch}@dai-labor.de

Abstract. Services for citizens provided by the government are often complex and related with various requirements. Citizens usually have a lot of questions traditionally answered by human experts. In this paper, we describe an information retrieval-based question answering (QA) system for the e-government domain. The QA system is capable of giving direct answers to questions in German concerning governmental services. The system successfully handles ambiguous questions by combining retrieval methods, task trees and a rule-based approach. We evaluate our system in a scenario tailored to the needs of the administration of a big German city. The preliminary results show that our system provides high-quality answers for the most questions.

Keywords: e-government, direct question answering, interactive IR

1 Introduction

Government services challenge both, citizens and agencies. Agencies continuously work on new ways for improving the efficiency and quality of services. IT-based solutions combining information retrieval (IR) and machine learning (ML) technologies are promising approaches for supporting the administration in improving the access to offered services. The administration of the German city Berlin already operates an online platform allowing users to inform themselves about all services the administrative agencies provide. However, the platform currently only offers a basic search but it does not provide direct answers to specific questions. Citizens must read through comprehensive service descriptions and find the piece of information they are interested in on their own. In addition, citizens often are not familiar with officialese making the formulation of search queries a challenging task for most users. In order to support the citizens getting informed, a system is needed that analyzes the users' intentions and applies advanced retrieval methods for providing detailed information tailored to the specific questions. In this work we present an IR-based e-government question answering system for German capable of handling three major tasks:

1. **Customized Ranking:** The system retrieves service descriptions and ranks them applying a customized scoring function that is based on service popularity.
2. **Passage Retrieval:** The system provides direct answers to user questions instead of showing comprehensive full-text documents requiring much effort for reading. Users do not need to scan the entire document anymore; users now quickly find a concrete answer to their question.
3. **Interactive QA:** The system is able to handle ambiguous and unclear questions. It addresses the problem by asking additional questions. If a user question is too general, the system checks back to refine the question.

The remaining paper is structured as follows. In Sec. 2 we describe the fundamentals of QA systems including our dataset, the evaluation metric, and existing QA systems. We present our approach in Sec. 3. The evaluation results are discussed in Sec. 4. Finally, a conclusion and an outlook to future work are given in Sec. 5.

2 Related Work

This Section describes common approaches to question answering and compares related question answering systems in the e-government domain. In particular, we review two major online services offering information for Berlin citizens.

2.1 Question Answering

Question answering systems find the correct answer to a (natural language) question based on a set of documents [4]. In general, there are two paradigms for QA: information retrieval-based and knowledge-based approaches.

Information retrieval-based systems answer a user's natural language question "by finding short text segments" [5, p. 2] in a collection of documents. These systems typically consist of three main components, often integrated in a pipeline: question classification, information retrieval, and answer extraction.

Knowledge-based approaches answer "a natural language question by mapping it to a query over a structured database" [5, p. 9]. Hence, they rely on already structured data, for example in a relational database. Among the knowledge-based approaches there are rule-based and supervised methods. Concerning rule-based methods the rules must be defined by hand, which is feasible for very frequent information needs. Supervised methods build a semantic representation of the user's query and then map it to the structured data.

Interactive QA is the combination of QA and dialogue systems in which users find answers in an interactive way. QA systems initiate a dialogue with the user in order to clarify missing or ambiguous information or to suggest further topics for discussion [6].

Our system uses an IR-based approach. It derives the answer types applying a set of rules defined by experts and retrieves passages as answers.

2.2 Related Systems

There are only a few publicly available online systems that enable Berlin's citizens to inform themselves of governmental services.

One of them is the nationwide platform "Behördenfinder". The platform comes as an extended redirection service. It passes the entered search terms unmodified to the respective search pages of the federal states. For example, the service redirects Berlin's citizens to the service portal of Berlin, which performs the search.

At the service portal of the city Berlin the citizens may search in the database of governmental services by entering keywords. The server returns a list of documents that contain the entered query terms, sorted by relevance. No further filter options are provided. The user has to open the links and manually search the appropriate section.

The service portals of other federal states often work in a similar way. But, some portals extend the search component by functions like "related" sections, categorizations (e. g. *buergerservice.niedersachsen.de*, *service-bw.de*), or similar search terms (e. g. *muenchen.de/dienstleistungsfinder*).

To the best of our knowledge, there is no question answering system available for Berlin's citizens able to answer government-related questions interactively and as accurately as possible.

3 Approach

In this Section we describe the used data sources and the approach we index the data to make the content searchable. We explain our document retrieval strategy and present three ranking methods. Our approach to interactive question answering includes grouping search results by selected service features and offering additional filters to the users. In order to provide users with concrete answers we present a method to question categorization allowing the retrieval of appropriate document passages. A web GUI enables users to access the system and to ask questions about services offered by administrative agencies. Figure 1 shows the system components and their interaction.

3.1 Data Sources

The main data source of the implemented system is the *LeiKa* ("Leistungskatalog der öffentlichen Verwaltung" [1, p. 1]). The *LeiKa* is a catalog assembling and categorizing services in order to create a Germany-wide central information base with uniform descriptions of services offered by administration departments.

Each service is identified by a key and categorized using multiple levels:

1. "Leistungsobjekt", the *service object* the service description deals with, e. g. driver's license.
2. "Verrichtung", the *service action* to perform, e. g. whether the citizen applies for a driver's license or his license has to be replaced.
3. "Verrichtungsdetail", the *service action detail* to describe the *service action* more precisely, e. g. whether it is an EU or an international driver's license.

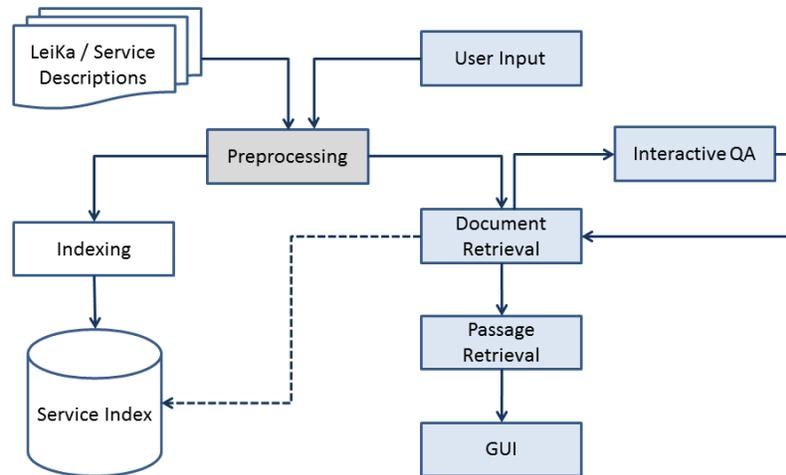


Fig. 1. The architecture of the presented interactive QA system consists of a service index, document and passage retrieval components, a component for enabling interactive searching, and a graphical user interface.

In addition, the LeiKa catalog provides for each service a set of standardized information such as a textual description, the costs for the service, the responsible authority and other necessary information [2, pp. 11-14].

The LeiKa catalog is already an important foundation of multiple projects, e. g. the uniform authority telephone service 115 (D115). In addition to the LeiKa service descriptions the D115 project provides popularity rankings for the top 100 services and commune-specific service descriptions [7, p. 5ff.].

Our system exploits a combination of the LeiKa and the D115 data of the Berlin government, including the ranking positions of the top 100 services.

3.2 Indexing

The data sources are provided in different formats. In order to make the data searchable we parse, aggregate, enrich and store it in an inverted index by using Elasticsearch [3]. This involves the annotation of the service documents with meta information, e. g. the popularity rankings (D115 top 100 services). We extend the services with additional keywords by applying NLP techniques. With the *Stanford* part-of-speech (POS) tagger we extract nouns and verbs from the title and the textual description of the services. Based on the extracted words, we determine additional keywords, i. e. synonyms and the stems of the words. We rely on the *Wortschatz* German dictionary web service [8] maintained by the University Leipzig.

3.3 Document Retrieval

The QA system is designed as an extension of a classical IR system. The retrieval of relevant documents builds the foundation of the question answering system. The retrieval part consist of two sub-tasks: (1) processing the user input and (2) formulating an appropriate query to search the document index. The user input is processed with the *Stanford* POS tagger and the *Wortschatz* web service in a similar manner as during the indexing process. We use a caching system to minimize response time and to reduce the number of required *Wortschatz*-queries. Based on the analyzed and enriched user input we formulate search queries. Our system implements the following three document scoring approaches to rank the retrieved documents:

Keyword scoring We assume that the description of relevant services contain at least one keyword of the extended user query. The more query keywords a service contains the more relevant it is. Therefore, we retrieve all services that contain at least one single keyword and sort them by the number of keyword occurrences within the service title or description.

TF-IDF scoring ElasticSearch supports full-text search by default, which is based on the term frequency-inverse document frequency (TF-IDF) scoring [9] from the Apache Lucene implementation³. For our retrieval task, we use the full-text search with the keywords from the user input on the complete service documents, whereby the weighting of document fields is normalized based on their content length.

Custom scoring Our custom scoring method (for ranking the documents) is an extension of the standard ElasticSearch scoring method based on the TF-IDF score. We modify the scoring function by adding a popularity factor to the formula, i. e. the rank of a service in the D115 top 100 list.

$$\text{score}(q, d) = \frac{1}{\text{D115-rank}(d)} \text{TF-IDF-score}(q, d) \quad (1)$$

Eq. 1 shows the custom scoring function of the query q and the service document d , where $\text{D115-rank}(d)$ is the rank in the D115 top 100 list and $\text{TF-IDF-score}(q, d)$ is the standard TF-IDF score.

3.4 Interactive Question Answering

The major challenge in interactive QA is the detection of ambiguities. Instead of detecting ambiguities in questions, we detect ambiguities in the retrieved services documents. We are capable of doing this, since the services descriptions are structured and not only plain texts.

The services are organized in objects, categories and actions. We extract these features from the retrieved services, group them and sort them by occurrence and popularity. We provide additional filters to ambiguous service results and let users interactively choose which object, category and action they are interested in.

³ http://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

Was meinten Sie?

Verrichtung:

Ausstellung Meldung Informationserteilung Befreiung

Objekte:

Personalausweis Ausweispflicht

Fig. 2. The GUI for interactive QA: A user enters an ID-related query. Filters for ID card (“Personalausweis”), obligatory identification (“Ausweispflicht”) and others are shown.

3.5 Passage Retrieval

Our application is an e-government QA system. We assume that users only enter questions related to governmental services. Therefore, the set of questions, our system needs to be able to answer, is limited. Based on the information passages (provided in the service documents), we manually pick four distinct types of questions, which can be answered: the costs of a service, the required documents, the opening time, and the location of the agency responsible for offering the service. For each type of question we define a static set of keywords. If a query contains such keywords (case-insensitive), we determine the type of the question and give the respective answer, i. e. we provide the corresponding excerpt of the service document directly in the search results or we provide several excerpts if multiple types match.

3.6 GUI

We design a simple and clear graphical user interface (GUI) that reflects the essential features of the QA system. The layout of the search result page is presented as a screen shot in Fig. 3. Users can input their questions, refine their questions (if they are ambiguous or too general), and read the search results including the corresponding text passages.

4 Evaluation

To allow a comparison to other QA systems and to judge whether algorithmic changes improve the system or not, we evaluate the QA system in a quantitative and distinct manner. Hence, we investigate the performance of our document and passage retrieval approach with two task-specific gold standard data sets.

4.1 Gold Standard

In order to evaluate our approach we need a data set containing German e-government questions along with the correct answers. To the best of our knowledge there is no gold standard data set that satisfies our needs. Thus, we create two data sets: one data set for measuring the performance of the document retrieval and one data set for the passage retrieval part, respectively.

Auf den Serviceseiten finden Sie, was Sie suchen.

Fig. 3. The GUI of the QA system: The user input is highlighted in the red box, the interactive QA in the green box. The search results are displayed with service titles (blue) and respective passages (yellow).

For the evaluation of the document retrieval process we develop a data set consisting of 6,700 questions, partly generated through permutation of synonyms or similar terms. Each question is associated with the correct set of answer documents.

For the passage retrieval evaluation we annotate 70 questions with their corresponding answers. The answers consist of the LeiKa- and D115-ID, the answer type for the passage retrieval, and the relevant LeiKa category information (service object, action, action detail). The set of questions and answers focuses on our four implemented answer types (see Sec. 3.5).

4.2 Document Retrieval

As our QA system follows an IR-based approach, our retrieval component provides a list of relevant services ordered by the relevance with respect to the user query. We assess the performance of our IR approach with the normalized Discounted Cumulative Gain (nDCG) measure. This measure considered both the relevance level and the order of the retrieved documents.

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)} \quad (2)$$

The standard Discounted Cumulative Gain (DCG) at a particular rank position p is computed as shown in Eq. 2. Our gold standard provides a binary relevance classification of service documents. Thus, the graded relevance of the result at position i is defined as $rel_i \in \{0, 1\}$.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (3)$$

In the e-government use-case of the QA system, we expect users to be only interested in the first few results and therefore we take only the top-k results with $k=10$ into account. We compare three document retrieval methods (Section 3.3) with a random baseline, i. e. documents retrieved in a random order. The scores of the random baseline are calculated as average over three runs.

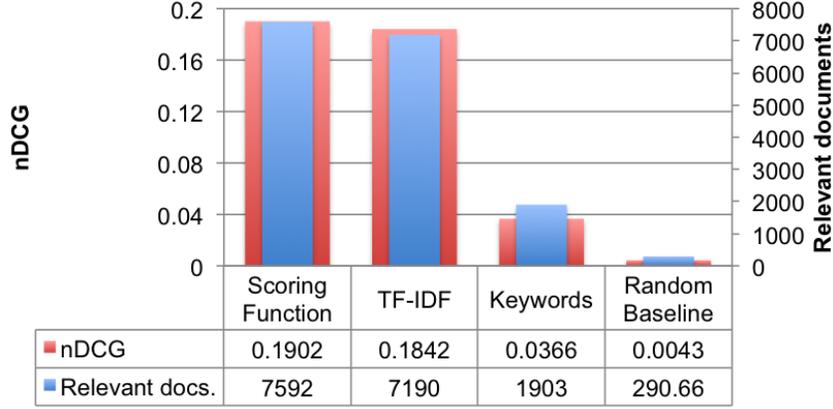


Fig. 4. Evaluation of query functions for document retrieval based on 6,700 questions and the top-10 documents of each approach.

The results in Fig. 4 show, that the custom scoring function leads to the best document retrieval performance in terms of nDCG and the total number of relevant documents. The inclusion of the popularity ranking affects the performance positively. The state-of-the-art TF-IDF method achieves lower but similar performance. On the contrary, the keyword-based approach yields rather poor results.

4.3 Passage Retrieval

We evaluate the passage retrieval component with a gold standard consisting of 70 questions and the corresponding passages.

Our QA system retrieves 54 passages (77%) correctly, while a random baseline achieves with 17.5 correct passages an accuracy of 25%.

5 Conclusion & Future Work

We developed a prototype of an e-government QA system applying a well-performing approach. The system is IR-based. It analyzes user questions, retrieves service documents from an inverted index and ranks them with a customized scoring function. We make use of the structured information encoded in the LeiKa catalog to provide a question categorization and passage retrieval feature and to interactively resolve ambiguous questions. A web-GUI enables users to interact with the system. We created two data consisting of 6,700 document retrieval and 70 passage retrieval questions. In a quantitative evaluation we showed that the use of the popularity ranking improves the retrieval quality.

The presented system is still work in progress. Hence, we propose two areas that require future work:

Incrementally optimized rankings Our nDCG evaluation already proves a good performance of the retrieved services. Anyhow, especially the ranking by relevance can be further improved by adjusting the ranking function to the domain of government services. We plan to evaluate additional user signals, e.g. clickstreams, to fine-tune this ranking function.

Improved QA A future goal is to provide QA features that do not rely on the structured information of the services or that answer only a limited set of question types. We aim for answering user questions based on knowledge and understanding and to optimize the handling of specific sub-tasks. In order to achieve this goal we plan to extend the data sets and to incorporate machine learning approaches.

References

1. Geschäfts- und Koordinierungsstelle LeiKa / BFD. Information der Geschäfts- und Koordinierungsstelle LeiKa. http://www.gk-leika.de/fileadmin/user_upload/gk-leika.de/dokumente/Startseite/News_der_GK_18082010.pdf, 2010. Accessed: 2016-03-30.
2. GK LeiKa. Handbuch: LeiKa-plus. http://www.gk-leika.de/uploads/media/Handbuch_LeiKa-plus_Stand_27.05.2014.pdf, 2014.
3. C. Gormley and Z. Tong. *Elasticsearch: The Definitive Guide*. O'Reilly Media, Inc., 2015.
4. P. Gupta and V. Gupta. A survey of text question answering techniques. *Intl. Journal of Computer Applications*, vol. 53, 2012.
5. D. Jurafsky and J. H. Martin. Speech and language processing. Chapter 28: Question answering. <https://web.stanford.edu/~jurafsky/slp3/28.pdf>, 2015.
6. N. Konstantinova and C. Orasan. Interactive question answering. <http://pers-www.wlv.ac.uk/~in0988/documents/iqa-chapter-final-21-07-2011.pdf>, 2011.
7. Projekt D115. Leitfaden: D115-informationsbereitstellung. http://www.115.de/SharedDocs/Publikationen/DE/Spezifikation_Leitfaeden/leitfaden_d115_informationsbereitstellung.pdf?__blob=publicationFile&v=1, 2011.
8. U. Quasthoff, M. Richter, and C. Biemann. Corpus portal for search in monolingual corpora. In *Procs. of the 5th Intl. Conf. on Lang. Resources and Eval.*, volume 17991802, 2006.
9. K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

From Cloud to Fog and Sunny Sensors

Position Paper

Hannes Grunert¹, Björn Butzin², Martin Kasparick², Andreas Heuer¹, and Dirk Timmermann²

¹ University of Rostock, Database Research Group, 18051 Rostock, Germany

² University of Rostock, Institute of Applied Microelectronics and Computer Engineering, 18051 Rostock, Germany

Abstract. Assistive systems collect large amounts of data in the internet of things and compute behavior and intentions of users in the cloud. Our approach is to push these computations (interpreted as database queries) as close as possible to the local sensors of the internet of things. We aim at replacing privacy-compromising cloud-based computations by fog- or edge-based computations or even by processing on the local sensors directly. Not only can this approach solve privacy problems, but also results in a better performance and energy-efficiency of the whole system: sensor-based computations are the (privacy-respecting) sunny side of the cloud. This position paper will give a short motivation and state of the art in different areas (from databases to wireless sensor networks) and will present our approach in combining modern interfaces to different sensors and concepts of database theory such as query rewriting and query containment.

Keywords: Cloud, Fog and Edge Computing, Internet of Things, Sensor Networks, Privacy, Performance, Query Rewriting, Query Containment

1 Motivation

Assistive systems support the users at work (Ambient Assisted Working) while they can remotely controls their homes (Ambient Assisted Living, AAL). Through various sensors, information about the current situations and the actions of the users are collected. This data is stored by the system and linked with other data from the web, for example the Facebook profiles of the users. By designing models for intention and activity recognition from the connected data, the smart environment can react autonomously to meet the needs of the users.

In assistive systems [12], significantly more information than required is collected in the cloud – which raises questions about privacy. The users usually have no or only a very small influence on the storage and processing of their personal data. If the cloud service is not located in their native country, the users cannot be sure that the same laws apply as in their home countries. As a result, their right to informational self-determination is violated.

The introduction of data privacy mechanisms in assistive systems is seen very skeptical by the developers. It is feared that the anonymization of the data hinders system development. Anonymization or pseudonymization of the data may lead to loss of detail, so that the results of analytic processes become inaccurate and, in extreme cases, unusable.

Our idea is to support assistive systems in performing the necessary behavior and intention recognition algorithms, but to automatically push the analysis operations as far as possible from the cloud to the (local) sensors. In an AAL environment, this results in a sensor-based or fog-based (edge-based) instead of a cloud-based computation. Besides privacy aspects, sensor-based or fog-based computing can also increase the performance and energy-efficiency of the system.

1.1 Cyber-physical systems and wireless sensor networks

The often cited law of Gordon Moore is used many times to argue that we do not need to increase efficiency, one just has to wait for the next hardware generation and thus a new boost of computational power. In wireless sensor networks (WSN), cyber-physical systems (CPS) and the internet of things (IoT), considerations are different. Here, the main driving factor is the reduction in energy consumption. Even newly developed moting platforms like UC Berkeley's Firestorm only have 512kB ROM and 64kB of RAM [1] which is not significantly more than in devices of 2005 in terms of absolute numbers, e.g., TelosB: 48kB ROM and 10kB RAM. Instead, the idle power consumption has been reduced, in this example by about 55 percent from $5.1 \mu A$ to $2.3 \mu A$. Hence, the focus on developing them is different. To save as much energy as possible, moting devices are kept asleep as long as possible. If awake, the transmission of data is the most costly operation in WSN, thus, it is tried to avoid them. Data should only be sent when requested by others or local memory is going to exceed. Additionally, through aggregation, preprocessing and compression, the time to send the next data-set can be extended. At this point computational power as well as memory capacities have to be utilized in an efficient way. After reducing the amount and frequency of transmissions required, the next level is to keep the middleware and transmission protocol energy-efficient. This means to reduce protocol overhead but also stack sizes and dynamic memory consumption need to be taken into account.

The idea of privacy through locality and energy saving constraints of wireless sensor networks harmonize with each other. However, other requirements of databases might be contradictory to the requirements of WSN e.g. in terms of latency, reliability, and consistency.

1.2 Privacy and Performance

By reducing and pre-aggregating raw sensor data to its minimal essence of required information, it is possible to protect privacy in smart systems. Additionally, the overall performance of the system can be enhanced when less data is analyzed on frequently busy nodes.

As part of our research project, privacy concepts for processing queries in assistive systems are designed. However, these concepts are not placed on top of the existing analysis functions, but are integrated in close cooperation during the development process.

The data-avoidant passing of the information regarding sensors and context to the analytical tools of the assistive system will not only improve the privacy-friendliness of the system. By pre-compressing the data by means of selection, aggregation, and compression operators on the sensor itself it is possible to increase the efficiency of the system. The privacy rights and the information requirements of the analysis tools can be implemented as integrity constraints in the database system. Through the integrity constraints, the necessary algorithms for anonymization and preprocessing can be run directly on the database. Thus, a transfer of the local data to external programs or modules, that might be located on different computing units, is omitted.

Instead of using hundreds of thousands of computers in the cloud (e.g. Google or Amazon), we can also use hundreds of billions of sensors or devices in the IoT to perform the necessary computations for the behavior and intention recognition of assistive systems. This results in fog or edge computing [11, 4] and even in local data processing on sensors.

2 State of the Art

We now give a short overview of the research areas cloud, Big Data, IoT (especially middleware and embedded database systems) and fog computing.

Cloud and Big Data: In the era of Big Data, more and more information is stored and processed in Cloud environments like IBM's Bluemix and similar platforms. Such systems offer a variety of services and possibilities for data storage, including services for the Internet of Things (e.g. APIs for REST and MQTT).

Unfortunately, privacy is often ignored or, at least, it is not guaranteed by cloud services. For example, nearly every service offered by IBM Bluemix states in the Terms of Use that the service “[...] does not comply with the US-EU [...] Safe Harbor Frameworks” (e.g. the Watson service “Driver Behavior” [7]).

Internet of Things — Middleware: IoT, CPS and WSN are distributed networks of small and heterogeneous applications. The service oriented architecture (SOA) approach has shown to be useful in such environments. SOA is well known for its capability to integrate different applications horizontally and vertically. Due to the restrictions of the devices in such scenarios, SOA approaches have been tailored to fit the needs for reduced overhead, descending memory footprint and less required computational power. Examples of those are the constrained application protocol (CoAP) and the devices profile for web services (DPWS). CoAP is a promising candidate for IoT applications, as it is a RESTful SOA

type with less overhead than HTTP and adds publish subscribe mechanisms. A comparison of different CoAP implementations can be found in [2]. Another competing protocol for networked embedded devices is Message Queue Telemetry Transport (MQTT). It is a pure publish subscribe protocol, that uses a centralized broker to manage the message flow. Often the cloud is used as broker. Thus, due to its centralized nature, MQTT is less optimal than CoAP for our proposed solution.

Data publishing of “traditional” database systems is performed on established interfaces like JDBC or ODBC. By these standardized interfaces, the programmer does not have to care about the actual implementation and can write code independent of the actual database system.

Embedded Databases: Besides standard database systems there exist several specialized databases like Berkeley DB [10] and TinyDB [9]. These systems are designed to run on resource limited devices like Raspberry Pis or even as embedded databases directly on the sensor. In [5], several approaches to a distributed database management on sensor networks are compared, TinyDB among them, here especially aiming at energy efficiency. Acquisitional query processing [5, 9] can push queries to sensors and select relevant sensors in a WSN to reduce the amount of sensors needed for a computation (sensor reduction). In the existing approaches mentioned in [5], this sensor reduction is completely done manually (by the programmer).

Fog Computing: What is missing in fog computing, is a database-centered approach to computation, that is, given a query representing a necessary computation on the sensor data of the IoT, how to automatically prevent to simply transfer the complete data sets to the cloud servers.

In [6] we introduced a framework for privacy aware query processing in layered networks of “traditional” databases. The query processor includes modules for query rewriting and transformation, detecting key-like combinations of attributes and different anonymization concepts like k-anonymity and slicing, which can be extended by including new concepts for querying modern hardware, e.g. by rewriting a SQL query to different data management layers, the lowest of these being only able to perform some simple filter operations (like selections against given attribute values).

3 Vision

We propose a layered architecture with four logically distinguishable layers (see Fig. 1). The *Sensor Layer* includes the sensors which are very resource constrained in terms of CPU, memory, and power. The *Personal Layer* consists of typically mobile devices or embedded systems with higher performance but also high power constraints, like smartphones or edge nodes of a WSN. Router, home automation control units, private servers, etc. build up the *Fog Layer*. As these

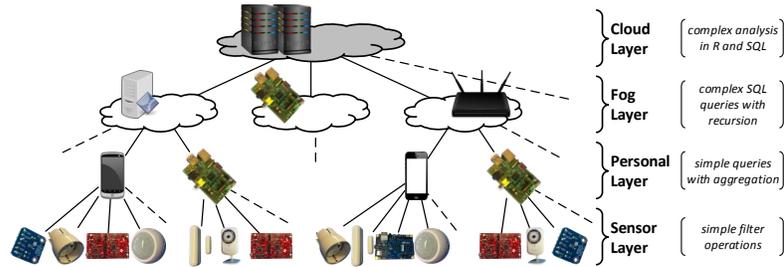


Fig. 1. Layered System Approach

devices have a wired power supply power saving is not as relevant as for the two lower layers. The *Cloud Layer* is built by powerful server farms without notable constraints according to power, CPU, or memory. Note that it is possible to have multiple layers within one of the four major layers, e.g. there could be several *Fog Layers* within a multi-tenancy or office building. From the top to the bottom layer resource constraints are increasing and the amount of possible (database related) functionalities and operations decreases. This layer approach has several advantages: In terms of privacy, each layer defines a strict transition where it can be defined which data is passed upwards and its granularity. This allows the fine-grained protection of critical personal data like health data, as the information can be stored and processed within the local parts of the system. Generally, the lower the layer, the higher is the ability of the user to control its own data.

As lower layers are more resource constrained than the upper ones, the middle layers provide functionalities for data processing as well as data transmission and additionally proxy functionalities. This enables optimized query execution according to the given resource constraints. The proxy functionality allows a reduction of the amount of communication. Especially within the *Sensor Layer* and between *Sensor* and *Personal Layer* this is a major aspect of power efficiency. The layered approach enables the power efficiency optimization of the overall system and not only for local nodes.

The constrained set of database operations at the lower layers can be compensated efficiently by the vertical fragmentation of queries into pushed-down and remainder queries. Thus, the privacy constraints of the users of these smart environments such as assistive systems are supported.

We apply our concept on machine learning algorithms to show how they will be transformed and pushed down in several steps resulting in simple filter operations.

A remaining open problem is to decide whether such queries can be performed on a resource-constrained device. If not, we have to check if the data can be sent to an upper layer without violating the privacy constraints of the users. This open problem results in a query containment problem that will be part of our future research.

Currently, only simple algorithms can be split up into their basic functions. The transformation of complex queries into simple fragments should be done automatically. By rewriting the complex query Q into Q_j and Q_δ , where Q_δ is executed outside of the protected system environment, we hope to only transfer data to the cloud that do not compromise privacy. We use extensions of the theory of query containment and query optimization for conjunctive queries [3, 8] to consider more complex queries (including complex statistical functions using aggregation and grouping) [6]. This approach can be extended to an IoT scenario with multiple layers, where the top layer is a cloud system while the bottom layer consists of embedded hardware.

The handling of data in IoT environments will be rethought fundamentally. Currently data is just pushed to the cloud while the layered approach enables new methods to store, process, and query data on the lower layers. To achieve this, IoT and database middleware have to be collated.

References

1. Andersen, M.P., Fierro, G., Culler, D.E.: System design for a synergistic, low power mote/BLE embedded platform. In: 2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). pp. 1–12
2. Butzin, B., Konieczek, B., Fiehe, C., Golatowski, F.: Applying the BaaS reference architecture on different classes of devices. In: 2nd International Workshop on Modelling, Analysis, and Control of Complex CPS (CPS Data). pp. 1–6 (Apr 2016), to be published
3. Chirkova, R.: Query containment. In: Encyclopedia of Database Systems, pp. 2249–2253. Springer US (2009)
4. Dastjerdi, A.V., Gupta, H., Calheiros, R.N., Ghosh, S.K., Buyya, R.: Fog computing: Principles, architectures, and applications. CoRR abs/1601.02752 (2016)
5. Diallo, O., Rodrigues, J.J.P.C., Sene, M., Mauri, J.L.: Distributed database management techniques for wireless sensor networks. *IEEE Trans. Parallel Distrib. Syst.* 26(2), 604–620 (2015)
6. Grunert, H., Heuer, A.: Datenschutz im PArADISE. *Datenbank-Spektrum* 16(2), 107–107 (July 2016)
7. IBM: IBM Watson IoT Driver Behavior Service. <http://www-03.ibm.com/software/sla/sladb.nsf/sla/bm-7328-01?Open>, last access: 09.06.2016
8. Kolaitis, P.G., Vardi, M.Y.: Conjunctive-Query Containment and Constraint Satisfaction. 17. Symposium on Principles of Database Systems, Seattle pp. 205–213 (1998)
9. Madden, S.R., Franklin, M.J., Hellerstein, J.M., Hong, W.: TinyDB: an acquisitional query processing system for sensor networks. *ACM Transactions on Database Systems (TODS)* 30(1), 122–173 (2005)
10. Oracle: Oracle Berkeley DB 12c. <http://www.oracle.com/technetwork/database/database-technologies/berkeleydb/overview/index.html>, last access: 09.06.2016
11. Shi, W., Dustdar, S.: The Promise of Edge Computing. *Computer* 49(5), 78–81 (May 2016)
12. Weiser, M.: The Computer for the 21st Century. *Scientific American* 265, 94–104 (1991)

Matrix Factorization for Near Real-time Geolocation Prediction in Twitter Stream

Nghia Duong-Trung, Nicolas Schilling, Lucas Rego Drumond, and Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Universitätsplatz 1, 31141 Hildesheim, Germany
{duongn, schilling, ldrumond, schmidt-thieme}@ismll.uni-hildesheim.de
<http://www.ismll.uni-hildesheim.de>

Abstract. The geographical location is vital to geospatial applications such as event detection, geo-aware recommendation and local search. Previous research on this topic has investigated geolocation prediction framework via conducting pre-partitioning and applying classification methods. These existing approaches target user’s geolocation all at once via concatenation of tweets. In this paper, we study a novel problem in geolocation. We aim to predict user’s geolocation at a given tweet’s posting time. We propose a geo matrix factorization model to address this problem. First, we map tweets into a latent space using a matrix factorization technique. Second, we use a linear combination in the latent space to predict exact latitude and longitude. However, we only use one individual tweet as the input instead of using a concatenation of all tweets of a user. Our experimental results show that the proposed model has outperformed a set of regression models and state-of-the-art classification approaches.

Keywords: Twitter, Near real-time Geolocation, Matrix Factorization

1 Introduction

In the past years, online social networking and social media sites, e.g. Twitter in general, have become an ubiquitous and constant mechanism for sharing and seeking information. Although a tweet’s length is limited to 140 characters, there is still a huge amount of information to explore. Its contents are inherently multifaceted and dynamic; consequently, representing people’s thoughts and public announcement at a temporal currency and vicinity. This causes Twitter data to become specifically interesting for multi-purpose investigations as they are tweeted in near real-time fashion. Understanding the near real-time user’s geographical location, e.g. latitude and longitude pairs or physical coordinates, enables providing policies and intervention aid strategies in a particular region such as localized aid [31,9], disaster response [27,21], event detection [28] and disease surveillance [3].

One of the early pioneer papers about geolocation in Twitter streams was published in 2010 [12]. In that work, the authors concatenated all user’s tweets

during a specified duration into one single representative document. The geolocation of the first tweet or the first available geo-tagged tweet in the collection was then the geolocation of the representative document. Using a concatenation provides circumstantial contents to develop a wide variety of techniques used in geo-locating such as content analysis with terms in a gazetteer [19], content analysis with probabilistic language models [11,16,1], metadata of various sorts such as follow-following relationships [17,22], behavior-based time zone [20]. Furthermore, the research conducted in [17] exploits the idea of geolocation prediction as label propagation by interpreting location labels spatially. Additionally, the work of [6] extends [17] by taking into account edge weights as a function reflecting user interactions.

Prerequisites to these directions are the representation of the earth’s surface. Geolocations can be captured as points, or clusters based on a pre-partitioning of regions into discrete sub-regions using city locations [5,18,26], named entities and location indicative words [14] as well as vernacular expressions with the aid of comprehensive gazetteers [15]. Another approach of partitioning the earth’s surface is to use a grid. While the simplest grid is a uniform rectangular one with cells of equal-sized degrees [30], more advanced grids are either an adaptive grid based on k -d trees [25], an equal-area quaternary triangular mesh [8] or a hierarchical structure [29].

However, these approaches have some drawbacks due to some reasons. First of all, as being classification methods, they heavily depend on pre-partitioning or a framing architecture that is used to split the regions into discrete sub-regions. Thus, they discard the natural properties of real physical coordinates. Moreover, concatenating tweets into one representative document requires a time-consuming collection as well as data abundance. In addition, concatenation of tweets during a particular duration, e.g. a month, leads to failure of capturing geolocation in near real-time situations. Effective geolocation of a user while posting a single short tweet based purely on its content is a direction worth-investigating and also constitutes a more difficult task.

In this paper, we address a novel geolocation prediction scenario via regression within indicative latent feature space. By working on the latent feature space, we have proved that regression models can be utilized to solve this prediction problem. We aim to predict the exact user geolocation at a given posting’s time, simply based on the textual content of tweets, ignoring their metadata.

2 Proposed Method

In this section, we present the general notation used in this paper as well as our approach. It is based on a matrix factorization of the individual tweets where we then learn a latent representation of tweets and words. This latent representation will then be used to predict the final geolocation. We also present a learning algorithm for our approach which is optimized by stochastic gradient descent.

2.1 Notation

Consider a dataset D containing a set of tweets where each tweet is described by n many features. The dataset will be split into a training D^{train} , a test D^{test} and a validation D^{valid} set, which will be used for hyperparameter optimization later. We have m , l and v tweets in the training D^{train} , test D^{test} and validation D^{valid} sets respectively. The tweet features are mapped from a dictionary that comprises all words/tokens/unigrams in the dataset. We denote the vocabulary size by $|V| = n$.

Each tweet is annotated with a ground-truth coordinate pair $\mathbf{y} \in \mathbb{R}^2$, $\mathbf{y} = (y^{lat}, y^{lon})$ where $y^{lat} \in \mathbb{R}$ is the latitude and $y^{lon} \in \mathbb{R}$ is the longitude of the associated tweet. By $\bar{\mathbf{y}}_{u_i} = (\bar{y}_{u_i}^{lat}, \bar{y}_{u_i}^{lon})$ we denote the average geolocation of a user in the training set, where $\bar{y}_{u_i}^{lat} \in \mathbb{R}$ is the average latitude and $\bar{y}_{u_i}^{lon} \in \mathbb{R}$ is the average longitude. Using $\bar{\mathbf{y}}_U = (\bar{y}_U^{lat}, \bar{y}_U^{lon})$, we denote the average geolocation of all users in the training set. Given some training data $X^{train} \in \mathbb{R}^{m \times n}$, and the respective labels $Y^{train} \in \mathbb{R}^{m \times 2}$, we seek to learn a machine learning model $f : \mathbb{R}^n \rightarrow \mathbb{R}^2$ which maps tweets to geolocations such that for some test data $X^{test} \in \mathbb{R}^{v \times n}$, the sum of distances

$$\sum_{i=1}^v d(f(X_i^{test}), Y_i^{test}) \quad (1)$$

is minimal. By $Y^{test} \in \mathbb{R}^{v \times 2}$ we denote the set of ground-truth labels for the test data. Note that, d is a distance metric where in our learning algorithm we use the Haversine distance.

2.2 The Geo Matrix Factorization Model

Over the last decade, Matrix Factorization (MF) models have gained much attention by the Netflix Prize competition where they have shown very good predictive performance as well as decent run-time complexity in terms of dealing with very sparse matrices. Based on the vanilla MF, we develop a more multi-relational-oriented factorization model for the geolocation regression task: the Geo Matrix Factorization (GMF) model. We approach the user geolocation problem as a text regression task where we aim to predict the exact latitude and longitude values using an individual tweet. However, instead of using the highly sparse word counts as features in a linear regression, we firstly factorize the input space by learning a matrix $T \in \mathbb{R}^{m \times k}$ for tweets and $W \in \mathbb{R}^{k \times n}$ for individual words of each tweet to reconstruct X as:

$$X \approx TW \quad (2)$$

As in the usual setting, the number of latent features k is usually much smaller than the number of words n , such that through this approach, tweets are projected into a lower dimensional latent feature space. This latent representation of a tweet is then used within a linear model to predict the geolocation of the user at the posting time of the tweet:

$$\begin{aligned}
\hat{y}_i^{lat} &= \bar{y}_{u_i}^{lat} + \phi_0 + \sum_{k=1}^K \phi_k T_{lk}^{lat} \\
\hat{y}_i^{lon} &= \bar{y}_{u_i}^{lon} + \theta_0 + \sum_{k=1}^K \theta_k T_{lk}^{lon}
\end{aligned} \tag{3}$$

where $\phi \in \mathbb{R}^{k+1}$ and $\theta \in \mathbb{R}^{k+1}$ are weight coefficients vectors for learning latitude and longitude respectively. Notice that we also actually perform two factorizations of X , one for latitude which yields T^{lat} , this is done for longitude as well. Our model then actually predicts the average training location of a user, plus a regression term on the latent feature space obtained by the factorization of X .

2.3 Model Fitting

Given the model, we have to learn parameters $T^{lat}, T^{lon}, W^{lat}, W^{lon}, \theta, \phi$, where the W matrices are only used for reconstructing X and not for predicting the actual geolocation. We optimize the prediction of the geolocation as well as the factorization of X for the least-squares error. In order to prevent the model from overfitting to the training data we apply a Tikhonov regularization on the regression parameters θ and ϕ , the latent feature matrices are regularized using the Frobenius norm. The overall loss term for learning the parameters associated to predicting latitude then looks like

$$\begin{aligned}
\mathcal{L}^{lat}(\hat{y}^{lat}, y^{lat}) &= \frac{1}{|X^{train}|} \|\hat{y}^{lat} - y^{lat}\|^2 + \lambda_\phi \|\phi\|^2 \\
&+ \|X^{train} - T^{lat}W^{lat}\|_F^2 + \lambda_T \|T^{lat}\|_F^2 + \lambda_W \|W^{lat}\|_F^2,
\end{aligned} \tag{4}$$

where the loss term associated to longitude $\mathcal{L}^{lon}(\hat{y}^{lon}, y^{lon})$ is similar. The only difference is that it involves θ, T^{lon} and W^{lon} . In Equation 4, the term $\|X^{train} - T^{lat}W^{lat}\|_F^2$ is the residual error of transforming X into T^{lat}, W^{lat} . The regularization terms $\lambda_\phi \|\phi\|^2$, $\lambda_T \|T^{lat}\|_F^2$, and $\lambda_W \|W^{lat}\|_F^2$ are multiplied by regularization parameters λ_ϕ, λ_T , and λ_W that control the amount of regularization.

These terms penalize parameters with high magnitudes, that typically lead to overly complex models with very small training errors but bad generalization performance. Certainly, these hyperparameters can not be learned from the data and will be optimized using a grid-search on the validation partition of the data. To solve the above optimization tasks, we apply Stochastic Gradient Descent (SGD) [2,13] where the learning rate is estimated using the Adaptive Subgradient Method (AdaGRAD) [10] which helps yielding a better run-time performance. The basic idea of SGD is that, instead of expensively calculating the gradient

of Equation 4 and its latitude counterpart, it randomly selects a tweet and calculates the corresponding gradient. Suppose we have chosen a tweet indexed by m , the partial derivatives of Equation 4 with the respect to T^{lat} can be computed as:

$$\begin{aligned} \frac{\partial \mathcal{L}^{lat}(\hat{y}_m^{lat}, y_m^{lat})}{\partial T_{ml}^{lat}} &= - \left(y_m^{lat} - \bar{y}_{u_m}^{lat} - \sum_{k=1}^K \phi_k T_{mk}^{lat} - \phi_0 \right) \phi_l \\ &\quad - \sum_{n=1}^N \left(\left(X_{mn} - \sum_{k=1}^K T_{mk}^{lat} W_{kn}^{lat} \right) W_{ln}^{lat} \right) + \lambda_T T_{ml}^{lat} \end{aligned} \quad (5)$$

The partial derivatives with respect to the latent feature matrix W^{lat} of the tokens is obtained by

$$\frac{\partial \mathcal{L}^{lat}(\hat{y}_m^{lat}, y_m^{lat})}{\partial W_{lj}^{lat}} = - \left(X_{mj} - \sum_{k=1}^K T_{mk}^{lat} W_{kj}^{lat} \right) T_{ml}^{lat} + \lambda_W W_{lj}^{lat} \quad (6)$$

Finally, the partial derivative of the regression parameters has the form:

$$\begin{aligned} \frac{\partial \mathcal{L}^{lat}(\hat{y}_m^{lat}, y_m^{lat})}{\partial \phi_j} &= - \left(y_m^{lat} - \bar{y}_{u_m}^{lat} - \sum_{k=1}^K \phi_k T_{mk}^{lat} - \phi_0 \right) T_{mj}^{lat} + \lambda_\phi \phi_j \\ \frac{\partial \mathcal{L}^{lat}(\hat{y}_m^{lat}, y_m^{lat})}{\partial \phi_0} &= - \left(y_m^{lat} - \bar{y}_{u_m}^{lat} - \sum_{k=1}^K \phi_k T_{mk}^{lat} - \phi_0 \right) \end{aligned} \quad (7)$$

The partial derivatives of the longitude loss with the respect to T^{lon} , W^{lon} and θ can be calculated in the exact same manner as Equations 5, 6 and 7.

2.4 Inference for Test Data

By optimizing the respective loss terms for the training data, we learn the latent representation T of all training tweets as well as the linear regression parameters θ and ϕ for predicting the final geolocation. However, as we want to predict geolocations of unseen test tweets, the latent representations T for the individual training tweets cannot be employed. Out of this reason, we perform a fold-in, where we factorize the feature matrix X^{test} of the test data, using the latent representation W of the word tokens that was learned on the training data. To avoid confusion, we denote the latent tweet representations for the test tweets by T^{lat} and T^{lon} and factorize X^{test} as

$$X^{test} \approx T^{lat} W^{lat} \quad (8)$$

as well as the respective term for longitude.

As we can see, W^{lat} and W^{lon} are reused from the learning phase. Subsequently, in the fold-in phase, we define the objective function that we need to minimize for T'^{lat} as follows:

$$\mathcal{L}^{lat}\left(X^{test}, T'^{lat}W^{lat}\right) = \frac{1}{|X^{test}|} \left\| X^{test} - T'^{lat}W^{lat} \right\|_F^2 + \lambda_{test} \left\| T'^{lat} \right\|_F^2 \quad (9)$$

The partial derivatives of Equation 9 with the respect to T'^{lat} can be computed by:

$$\frac{\partial \mathcal{L}^{lat}\left(X_{jn}^{test}, T'_{jk}{}^{lat}W_{kn}^{lat}\right)}{\partial T'_{jk}{}^{lat}} = - \left(X_{jn}^{test} - \sum_{k=1}^K T'_{jk}{}^{lat}W_{kn}^{lat} \right) W_{kn}^{lat} + \lambda_{test} T'_{jk}{}^{lat} \quad (10)$$

The partial derivatives with the respect to T'^{lon} can be also computed in the same manner as for Equation 10. Having learned the latent representation of the test tweets using the fold-in procedure, we can then perform predictions for the test users using Equation 3. However, not all users that appear in the test data necessarily have to appear in the training data, hence we cannot use their average geolocation for the final prediction. For those users, we then use the median geolocation of all users of the training data as:

$$\bar{\mathbf{y}}_{u_l} = \begin{cases} \bar{\mathbf{y}}_{u_l}, & \text{if } u_l \in D^{train} \\ \bar{\mathbf{y}}_U, & \text{otherwise} \end{cases} \quad (11)$$

Algorithm 1 illustrates how the overall GMF works.

3 Experiments

In this section, we first describe the datasets that we use as well as their pre-processing. Additionally, we describe how we optimized the hyperparameters of our model. Finally, we compare our approach to a set of competing methods.

3.1 Dataset

We have worked with three publicly available tweet datasets containing geolocation information and compiled them to fit the user geolocation prediction within the near real-time scenario. One dataset comprises the tweets posted within the United States, whereas the other dataset contains all tweets localized to north America and the world. Through this, we evaluate our model’s effectiveness and generality within different geographical scopes from a country to the whole world. A splitting protocol is then designed for these datasets. We randomly

Algorithm 1 GMF

Require: $X^{train} \in \mathbb{R}^{m \times n}$, $X^{test} \in \mathbb{R}^{l \times n}$, $Y \in \mathbb{R}^{m \times 2}$ **Ensure:** $T \in \mathbb{R}^{m \times k}$, $T' \in \mathbb{R}^{l \times k}$, $W \in \mathbb{R}^{k \times n}$, $\phi \in \mathbb{R}^{k+1}$, $\theta \in \mathbb{R}^{k+1}$

```
1: Initialize  $T^{lat} \leftarrow \mathcal{N}(0, 1)$ ,  $T^{lon} \leftarrow \mathcal{N}(0, 1)$ ,  $W^{lat} \leftarrow \mathcal{N}(0, 1)$ ,  
    $W^{lon} \leftarrow \mathcal{N}(0, 1)$ ,  $\phi \leftarrow \mathcal{N}(0, 1)$ ,  $\theta \leftarrow \mathcal{N}(0, 1)$ ,  $T'^{lat} \leftarrow \mathcal{N}(0, 1)$ ,  $T'^{lon} \leftarrow \mathcal{N}(0, 1)$   
2: // Learning phase  
3: for  $epoch \in 1, \dots, max\_epoch$  do  
4:   for  $iteration \in 1, \dots, M$  do  
5:     Pick  $m$  randomly  
6:     Pick  $X_{mn}^{train}$  randomly  
7:     for  $k \in 1, \dots, K$  do  
8:       Learning  $T_{mk}^{lat}$ ,  $T_{mk}^{lon}$ ,  $W_{kn}^{lat}$ ,  $W_{kn}^{lon}$ ,  $\phi_{T_{mk}}$ ,  $\theta_{T_{mk}}$   
9:     end for  
10:    Update  $\phi_0$ ,  $\theta_0$   
11:   end for  
12: end for  
13: // Fold-in phase  
14: for  $epoch \in 1, \dots, max\_epoch'$  do  
15:   for  $iteration \in 1, \dots, L$  do  
16:     Pick  $l$  randomly  
17:     if  $X_{ln}^{test}$  exists then  
18:       for  $k \in 1, \dots, K$  do  
19:         Learning  $T'_{lk}^{lat}$ ,  $T'_{lk}^{lon}$   
20:       end for  
21:     end if  
22:   end for  
23: end for  
24: // Prediction  
25: for  $l \in 1, \dots, L$  do  
26:    $\hat{y}_l^{lat} \leftarrow \bar{y}_{u_l}^{lat} + \phi_0 + \phi_{lk} T'^{lat}_{lk}$   
27:    $\hat{y}_l^{lon} \leftarrow \bar{y}_{u_l}^{lon} + \theta_0 + \theta_{lk} T'^{lon}_{lk}$   
28: end for  
29: return  $d_H(\mathbf{y}, \hat{\mathbf{y}})$ 
```

split *all tweets of each user* by a 60/20/20 scheme, denoted as LocalRandom (LR). Secondly, we also investigate how our model works with a user appearing in the test set might not exist in the training data by splitting *all tweets* using the 60/20/20 scheme, called GlobalRandom (GR).

US. This dataset is originally implemented by [12], and was later also used in [11,30,16]. The dataset comprises tweets gathered from the "Gardenhose" sample stream in the first week of March, 2010. In this dataset, the authors already provide geotagged tweets that we simply reuse. The implementing dataset contains 377,616 tweets posted by 9,475 users.

NA. The second dataset was collected by [25] and later implemented by [29,15]. This dataset contains tweets within north America, including the United

States, parts of Canada and Mexico from September 4th to November 29th, 2011. Because Twitter does not allow the distribution of complete tweets at that time, the NA dataset only contains user IDs and tweet IDs. Subsequently, we have to fetch the tweets from Twitter using its official API to check whether the tweets are available as well as their availability of embedded coordinates. Only 226,595 tweets out of 38 million posted by 10,950 users have geotags available and therefore are considered for the final dataset.

WORLD. The last dataset was compiled by [14] and later implemented by [29,15]. The dataset comprises tweets from all over the world. As being described in NA dataset, we also apply the same retrieving procedure. The implementing dataset then contains 121,327 tweets posted by 80,179 users. In the WORLD dataset, 70% of users has only one tweet. So that we only apply the GR 60/20/20 splitting scheme to it.

3.2 Data Preprocessing

In addition to length restriction, tweets are also characterized by the use of terms that are not found in natural language, including hashtags, abbreviations, emoticons and URLs. Through this, we propose a data preprocessing procedure as follows.

Tokenization. We apply a uni-gram tokenization procedure that preserves hashtags, @-replies, abbreviations, blocks of punctuation, emoticons and unicode glyphs and other symbols as tokens. We remove URL tokens to prevent the tweets where bots are posting information such as advertisement to enter our dataset.

Bag-of-words representation. After all tweets are tokenized, they are converted from sparse vectors of token counts into sparse vectors of bag-of-words representations using term frequency - inverse document frequency (TF.IDF) scores. By using the TF.IDF scores, we discard language and grammar structure, the token’s order, semantics and meaning as well as part-of-speech. The TF.IDF weights reflect how important a token is to an instance. The more common a token is to many instances, the more penalization it gets. The tokens with the highest TF.IDF weight are often the tokens that best characterize the instance.

3.3 Evaluation Metrics

Given the ellipsoidal shape of the earth’s surface, we apply the Haversine distance to calculate the distance of two points represented by their latitude in range of $\{-90, 90\}$ and longitude in range of $\{-180, 180\}$. The Haversine distance $d_H : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ is the great circle distance between two geographical coordinate pairs. We compute the distance between two points by the Haversine formula [24]. The formula of the central angle α between them is given by:

$$\alpha = \left(\sin^2 \left(\frac{|\hat{y}^{lat} - y^{lat}|}{2} \right) + \cos(y^{lat}) \cos(\hat{y}^{lat}) \sin^2 \left(\frac{|\hat{y}^{lon} - y^{lon}|}{2} \right) \right)^{\frac{1}{2}} \quad (12)$$

Then, the Haversine distance of the two points can be calculated by:

$$d_H(\mathbf{y}, \hat{\mathbf{y}}) = 2r \arcsin(\alpha) \tag{13}$$

where r is the radius of the earth. Because of the ellipsoidal shape of the earth, its radius varies from the equator to the poles. According to [7], we take the mean of the earth’s radius which amounts to $r = 6371$ km. Finally, the evaluation metrics are the mean and median Haversine distances d_H in kilometers between the ground-truth geolocation \mathbf{y} and the predicted geolocation $\hat{\mathbf{y}}$.

3.4 Hyperparameter Setup

In order to obtain good predictive performance, we also need to carefully tune the hyperparameters in our model. By $k \in \mathbb{N}^+$ we denote the number of latent features used within the factorization of X . By $\lambda_T, \lambda_W, \lambda_\phi, \lambda_\theta$ and $\lambda_{T'}$ we denote the regularization hyperparameters used when learning the latent feature matrices, latent vocabulary matrices, the linear regression parameters for predicting latitude and longitude and the latent features matrices for the test tweets respectively. With $\alpha_T, \alpha_W, \alpha_\phi, \alpha_\theta$ and $\alpha_{T'}$ we denote the respective learning rates. We tune the hyperparameters by assessing the validation performance of our model and choosing the hyperparameter configuration which performs best. The number of latent dimensions is selected among the range of $k \in \{2, 4, 8, 16\}$, while the value of all other hyperparameters are selected among the range of $\{0.1, 0.01, 0.001, 0.0001, 0.00001\}$. The preprocessed datasets used in the paper are publicly available unconditionally¹.

3.5 Results and Comparison

For the Support Vector Machine (SVM) and Factorization Machines (FM), we run them separately to predict latitude and longitude. To allow for a fair comparison, all these regression models also include the user bias in their estimation. Finally, we combine the predicted latitude and longitude to conduct a final distance calculation. For these models, we also apply a grid-search mechanism to find the best hyperparameter configurations for each prediction of latitude and longitude. On each dataset, we repeat running the models 10 times and take the average results. The final results can be observed in Table 1. We can see that all other regression models on average do not perform that well, mainly due to them using the extremely sparse 5,200 TF.IDF features. Our model, however, maps each tweet individually into an eight-dimensional latent feature space and uses those features for prediction. The number of k latent feature is found by grid-search mechanism. The results show that GMF outperforms all competitors with large margins.

We also report the state-of-the-art results by classification approaches (see Table 2). One might notice that there are significant differences in term of accuracy prediction in two geolocation prediction scenarios. By targeting user’s geolocation at a given posting’s time, the results show that our model significantly

¹ Available online at: <http://fs.ismll.de/publicspace/GMF/>

Table 1. The results by regression approaches targeting the user’s geolocation in a given posting’s time scenario using only textual information. The mean and median Haversine distance error are in *km*. The best distances are in **bold**.

Corpus	LR_US		LR_NA		GR_US		GR_NA		WORLD	
Model	mean	median	mean	median	mean	median	mean	median	mean	median
SVM (RBF kernel) [4]	34.63	7.81	157.81	8.42	32.29	8.22	171.72	10.23	3179.57	2654.17
FM [23]	29.67	0.68	164.51	7.27	27.09	0.66	177.53	7.26	3219.16	2650.48
Our model	29.15	0.66	157.22	6.95	26.44	0.65	170.08	7.19	2524.66	553.24

reduces the localization error on the US and NA datasets. For the WORLD dataset, the average individual tweet’s length is 5 tokens while being 49 tokens for the concatenation of tweets, our model still achieves reasonable results.

Table 2. The state-of-the-art results by classification approaches targeting the user’s geolocation in all-at-once scenario using only textual information. The mean and median Haversine distance error are in *km*. (“-” signifies no implemented results for the given dataset, and “?” signifies that no result was reported for the given metric).

Corpus	US		NA		WORLD	
Model	mean	median	mean	median	mean	median
Hierarchical clustering [29]	-	-	686.6	171.5	1669.6	490.0
Hierarchical topic model [1]	?	298	-	-	-	-

4 Conclusion and Future Work

We have investigated the geo matrix factorization model for the task of near real-time text-based geolocation in Twitter. In our work, we tackle the user geolocation prediction task in a regression perspective. We analyze a single tweet as the model’s input without any concatenation. Through this, we can further predict the user trajectory and achieve geolocation at a given posting’s time. This is a starting point for further investigation on the affection of tweet concatenation or the number of tweets needed to achieve an acceptable distance error. Furthermore, We also address the sparsity and imbalance of online conversational texts by a matrix factorization technique. Based on the experiment results, our model outperforms all the competitors including SVM and FM within the regression task using dedicated latent feature spaces. In comparison with current state-of-the-art results by classification approaches, our model still outperforms and/or achieve reasonable results. Our further improvement broadly falls into various directions: optimization or applying the model over different datasets. In the optimization direction, we will analyze direct optimization of the Haversine formula. We also expand our model to predict near real-time geolocation of another types of datasets such as Wikipedia articles and Flickr images.

Acknowledgments. Nghia Duong-Trung gratefully acknowledges the funding of his work by the Ministry of Education and Training of Vietnam under the national project no. 911.

References

1. Ahmed, A., Hong, L., Smola, A.J.: Hierarchical geographical modeling of user locations from social media posts. In: Proceedings of the 22nd international conference on World Wide Web. pp. 25–36. International World Wide Web Conferences Steering Committee (2013)
2. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT’2010, pp. 177–186. Springer (2010)
3. Burton, S.H., Tanner, K.W., Giraud-Carrier, C.G., West, J.H., Barnes, M.D.: ”right time, right place” health communication on twitter: value and accuracy of location information. *Journal of medical Internet research* 14(6) (2012)
4. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 27:1–27:27 (2011)
5. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 759–768. ACM (2010)
6. Compton, R., Jurgens, D., Allen, D.: Geotagging one hundred million twitter accounts with total variation minimization. In: *Big Data (Big Data), 2014 IEEE International Conference on*. pp. 393–401. IEEE (2014)
7. Decker, B.L.: World geodetic system 1984. Tech. rep., DTIC Document (1986)
8. Dias, D., Anastácio, I., Martins, B.: A language modeling approach for georeferencing textual documents. In: *Actas del Congreso Español de Recuperación de Información* (2012)
9. Dredze, M.: How social media will change public health. *Intelligent Systems, IEEE* 27(4), 81–84 (2012)
10. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research* 12, 2121–2159 (2011)
11. Eisenstein, J., Ahmed, A., Xing, E.P.: Sparse additive generative models of text. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 1041–1048 (2011)
12. Eisenstein, J., O’Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 1277–1287. Association for Computational Linguistics (2010)
13. Gemulla, R., Nijkamp, E., Haas, P.J., Sismanis, Y.: Large-scale matrix factorization with distributed stochastic gradient descent. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 69–77. ACM (2011)
14. Han, B., Cook, P., Baldwin, T.: Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012: Technical Papers* pp. 1045–1062 (2012)
15. Han, B., Cook, P., Baldwin, T.: Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research* pp. 451–500 (2014)

16. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J., Tsioutsoulouklis, K.: Discovering geographical topics in the twitter stream. In: Proceedings of the 21st international conference on World Wide Web. pp. 769–778. ACM (2012)
17. Jurgens, D.: That’s what friends are for: Inferring location in online social media platforms based on social relationships. In: ICWSM (2013)
18. Kinsella, S., Murdock, V., O’Hare, N.: I’m eating a sandwich in glasgow: modeling locations with tweets. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents. pp. 61–68. ACM (2011)
19. Li, R., Wang, S., Deng, H., Wang, R., Chang, K.C.C.: Towards social user profiling: unified and discriminative influence model for inferring home locations. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1023–1031. ACM (2012)
20. Mahmud, J., Nichols, J., Drews, C.: Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(3), 47 (2014)
21. McClendon, S., Robinson, A.C.: Leveraging geospatially-oriented social media communications in disaster response. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 5(1), 22–40 (2013)
22. McGee, J., Caverlee, J., Cheng, Z.: Location prediction in social media based on tie strength. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 459–468. ACM (2013)
23. Rendle, S.: Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3(3), 57 (2012)
24. Robusto, C.: The cosine-haversine formula. *American Mathematical Monthly* pp. 38–40 (1957)
25. Roller, S., Speriosu, M., Rallapalli, S., Wing, B., Baldrige, J.: Supervised text-based geolocation using language models on an adaptive grid. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1500–1510. Association for Computational Linguistics (2012)
26. Rout, D., Bontcheva, K., Preotjiuc-Pietro, D., Cohn, T.: Where’s@ wally?: a classification approach to geolocating users based on their social ties. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. pp. 11–20. ACM (2013)
27. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. pp. 851–860. ACM (2010)
28. Weng, J., Lee, B.S.: Event detection in twitter. *ICWSM* 11, 401–408 (2011)
29. Wing, B., Baldrige, J.: Hierarchical discriminative classification for text-based geolocation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 336–348 (2014)
30. Wing, B.P., Baldrige, J.: Simple supervised document geolocation with geodesic grids. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 955–964. Association for Computational Linguistics (2011)
31. Yin, J., Lampert, A., Cameron, M., Robinson, B., Power, R.: Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems* (6), 52–59 (2012)

Scalable Inference in Dynamic Admixture Models

Patrick Jähnichen, Florian Wenzel, and Marius Kloft
{jaehnicp, wenzelfl, kloft}@hu-berlin.de

Machine Learning Group
Humboldt-University of Berlin
Germany

Dynamic probabilistic models are standard in various time-series applications, including weather forecasting, stock market analysis, and robotics. Typically such models consist of a diffusion model that governs the state of the system and a model of measuring this state. As an example consider the simple non-mixture time series model

$$\begin{aligned}\mu_t &= \mu_{t-1} + v_t, v_t \sim \mathcal{N}(0, \nu^2) \\ x_t &= \mu_t + w_t, w_t \sim \mathcal{N}(0, \sigma^2).\end{aligned}$$

where μ_t is the state of the system at time t and x_t is a noisy measurement of that state. Note that this kind of model is akin to the well-known Kalman filter.

A drawback of such a simple model is that it does not capture data that is a mixture of several possible components in varying proportions. Such data emerges in e.g. corpora modeling where each document is comprised of words that are generated by different themes that are present in the corpus and underly a time dynamic. An example of a more complex model is the continuous time dynamic topic model (cDTM) [6]. In this model the time structure of the mixture components is modeled in terms of a Markov chain. We generalize this approach to general Gaussian processes (GPs). This allows for more flexible modeling of the diffusion process (time structure) by changing the GP covariance function, capturing a wider variety/combination of mixture component dynamics.

Inference in these models is a major challenge. The posterior we seek is generally intractable and we must appeal to an approximation. Up until recently, state-of-the-art approaches used variational inference as in [3, 6] and our own preliminary research [4]. As these approaches are limited to rather small datasets, [2] recently applied a stochastic gradient Langevin dynamics sampler [7] which allows for inference in these models using larger numbers of datapoints. However, [1, 5] have shown that this approach is amenable to considerable improvements.

We develop an inference method which is based on more evolved stochastic gradient based sampling techniques (as e.g. [2]) leading to a novel robust inference method which is applicable to millions of data points. Our preliminary empirical findings suggest that we can improve performance in terms of accuracy and speed over the state-of-the-art methods.

Bibliography

- [1] S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, 2012.
- [2] A. Bhadury, J. Chen, J. Zhu, and S. Liu. Scaling up Dynamic Topic Models. In *Proceedings of the 25th International ...*, 2016.
- [3] D. M. Blei and J. D. Lafferty. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 2006.
- [4] P. Jähnichen. *Time dynamic topic models*. PhD thesis, Leipzig University, Leipzig, Mar. 2016.
- [5] S. Mandt, M. D. Hoffman, and D. M. Blei. A Variational Analysis of Stochastic Gradient Algorithms. *arXiv.org*, 2016.
- [6] C. Wang, D. M. Blei, and D. Heckerman. Continuous time dynamic topic models. *Proc. of UAI*, 2008.
- [7] M. Welling and Y.-W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

Design of Knowledge Analytics Tools for Workplace Learning^{*}

Maria A Schett¹, Stefan Thalmann², and Ronald K Maier²

¹ University of Innsbruck, Faculty of Mathematics, Computer Science & Physics,
Department of Computer Science

² University of Innsbruck, School of Management, Department of Information Systems,
Production & Logistics Management, Information Systems I

`mail@maria-a-schett.net`
`{stefan.thalmann, ronald.maier}@uibk.ac.at`

Abstract. The amount of documented organizational knowledge steadily increases as well as the amount of knowledge available from external sources. At the same time the need for innovation at the workplace also increases and poses the challenge to support employee's workplace learning. Knowledge analytics seems to be a promising approach to help employees to sift through piles of documents and select knowledge suitable for their learning at the workplace. However, little is known about the requirements for knowledge analytics in general and in the context of workplace learning in particular. Therefore, we developed a knowledge analytics tool and applied it in a case study. We performed seven artifact-driven expert interviews within this case study to elicit the requirements for a knowledge analytics tool. Based on our investigation we developed three candidate design patterns: (1) provenance and traceability, (2) human factor and stakeholder rating, and (3) visualization of the proposed solution. Our design patterns can be used to inform the design of knowledge analytics tools, particularly in the context of workplace learning.

Keywords: knowledge analytics, workplace learning, design patterns

1 Introduction

The increasing amount of organizational knowledge and the increasing need for workplace learning poses a difficult challenge to organizations: How can an educational program manager provide suitable knowledge to the employees? How can a content developer select those contents in a large digital library, which are best suited for adaptation for workplace learning? In this paper we present the findings of a case study in order to answer these questions. The goal of this paper is to develop candidate design patterns for a knowledge analytics tool used

^{*} The research leading to the presented results was partially funded by the European Commission under the 7th Framework Programme (FP7), Integrating Project LEARNING LAYERS (Project no. 318209).

for workplace learning. Our case study is performed in the context of the EU FP7 Learning Layers³ (LL) project. We implemented a knowledge analytics tool to recommend contents for adaptive preparation in the context of workplace learning and discussed the recommendation proposed by the tool with seven experts using artifact-driven, semi-structured telephone interviews. Based on these interviews, we developed three candidate design patterns, which capture context, problem, and solution, intended to help the design of functionality for knowledge analytics tools.

2 Background

Informal learning is seen as the most important way to acquire and develop skills and competencies within the workplace [1]. Workplace learning is nested in everyday problem solving situations, where people learn through mistakes and interactions with colleagues as well as by learning from others' personal experiences [3]. Workplace learning is increasingly promoted because of the changes in work organizations and the appearance of new types of management [6]. The important interplay of both the informal and the social characteristics is further emerging in research on learning in the workplace [8]. Mobile devices enable access to documented knowledge (from inside and outside the organization) plus interactions with colleagues to foster workplace learning [19]. Workplace learning boosted by proceeding scalable learning solutions could take place with such devices useable from a variety of locations [23]. The quality of scalability particularly depends on user acceptance. Hence, the solution needs to be assimilated into the daily learning practices of a critical mass of users [18]. Compared to more traditional learning settings, the unstructured, creative, and expertise-driven informal learning cannot be designed with standardized management approaches and cannot be easily supported by IT [13]. Hence, much more contents prepared for more diverse learner needs are needed, which requires new ways of IT support. In this regard, knowledge analytics seems promising to cope with the increasing complexity.

Analytics is used in different settings, e.g., business analytics [2], learning analytics [10], or academic analytics [7] and we aim to take the analogy of analytics to documented organizational knowledge: knowledge analytics. The term analytics has many facets: It features data-driven decision making [22], by using mathematical techniques to analyze data [24] to drive fact-based planning, decisions, execution, management, measurement and learning [9]. The aim of analytics is to develop actionable insights, which give the potential for practical action [4]. Analytics can be used to model the past, recommend in the present, and predict, optimize, and simulate the future [5]. Van Barneveld et al. [22] identify three trends for combining analytics with a keyword: (1) topic of interest, e.g., learning analytics if we are interested in learning, (2) intent of the activity, e.g., predictive analytics, and (3) object of analysis, e.g., analytics based on Google (i.e., Google analytics).

³ See learning-layers.eu.

Following this categorization, we define knowledge analytics from perspective (3), object of analysis as analytics, based on knowledge, as opposed to data. First, we define knowledge after Zack [25] as organized accumulation of data enriched by context, and describe knowledge succinctly as “context and content”. Then, we define knowledge analytics as analytics which use knowledge as input to create value as output. In this paper we describe our knowledge analytics approach applied to a case study within the context of the LL project. Thereby we next apply our definition: content/data and context/metadata is used as input for analytics to create value in form of a proposed solution.

Content/Data. In the LL project case 58 knowledge elements were created. They are stored in text documents, presentations, spreadsheets, videos, and wiki pages and have topics clustered around theories of learning and knowledge. The goal was to use these knowledge elements to support workplace learning taking the preferences of the project members into account.

Context/Metadata. These knowledge elements were rated, or more specifically, factors which reflect benefits and efforts with respect to the knowledge elements were rated. Ratings of factors reflecting benefits were collected via an online survey of LL project members, the raters. Ratings of factors reflecting efforts were collected from a technical and a domain authority, and from employees, who are responsible for adapting the knowledge elements.

Analytics. Our analytics approach is the Knowledge Element Preparation model (KEP model) proposed by Thalmann [20]. The KEP model poses a linear, 0/1 combinatorial optimization problem. We employed a general purpose solver to implement the model in the KEP tool. We expressed the KEP model in a subset of the modeling language AMPL, namely GMPL, a declarative language with algebraic notation, which is close to the mathematical description of the KEP model, and utilized the free Gnu Linear Programming Kit. With the KEP tool, we computed a (*KEP*) *proposed solution* of knowledge elements best suited for preparation with respect to adaptation criteria based on the collected ratings of factors reflecting benefits and factors reflecting efforts. We selected five adaptation criteria from [21] to make the knowledge elements more accessible to learners in situations of workplace learning: device requirements, didactical approach, language, presentation preferences, and previous knowledge.

Value. The KEP proposed solution creates value in the form of decision support for a content developer or an educational program representative.

3 Procedure

After we had implemented the KEP model and applied the KEP tool to get the solution of knowledge elements to be prepared for workplace learning in the LL context, we conducted artifact-driven interviews, where the vehicles of

ID	Length	Gender	Role in Project	Work Exp.	Country of origin
Ex01	40:35	M	Senior Researcher	> 10 years	Germany
Ex02	23:40	F	Senior Researcher	> 10 years	United Kingdom
Ex03	34:45	M	Senior Researcher	> 10 years	Luxembourg
Ex04	32:03	M	Workpackage Leader	> 10 years	United Kingdom
Ex05	22:27	F	Senior Researcher	5-10 years	Spain
Ex06	26:58	F	Workpackage Leader	> 10 years	Finland
Ex07	37:26	M	Researcher	5-10 years	Germany

Table 1. Demographics of Interview Partners.

our interviews were our artifacts: the KEP model, the KEP tool, and the KEP proposed solution. The goal of the interviews was to identify design patterns. We interviewed seven experts from the LL project. Their demographics can be found in Table 1. The interviews were structured with an interview guideline and by open-ended questions. The interview guideline with the questions can be found in [12]. The interview was motivated by the evaluation of the KEP model and the KEP tool. It was structured in three blocks: (i) evaluation of the factors of the KEP model (based on [16]) and the reasoning behind rating them, (ii) investigation of the KEP proposed solution, and (iii) outlook with requirements on the graphical user interface (GUI). As we performed the interviews remotely, we used visual aids through screen sharing: slides presenting the KEP approach and a spreadsheet showing the KEP proposed solution. We presented this spreadsheet and collected the experts’ opinions. The interviews were transcribed verbatim and double-checked and took between 22 and 41 minutes.

We then analyzed the transcripts by a qualitative content analysis after Mayring [14] using deductive codes derived from the research design. The results of the interviews are given in Section 4. From the results we generated three *candidate design patterns*. Design patterns communicate high level and good solutions to recurring problems and can be seen as artifacts of design science research [17]. On an abstract level, design patterns follow the structure: “for problem P under circumstance C solution S has been known to work” [15]. Design patterns can be valuable for practitioners, as they describe practical and applicable solutions, and for researchers to synthesize and capture knowledge as well as to provide further research directions. In our work we developed candidate design patterns, which are discovered from experience and knowledge, and they are titled candidate, as they need to be validated [17]. We identified the three relevant candidate design patterns by discussing the reflections and suggestions given by the interviewees iteratively in three sessions within the group of co-authors (compare to the participatory pattern workshop methodology [15]). The patterns follow the structure of design patterns of Mor et al. [15] and the LL project [11].

4 Results

We present the results from the artifact-driven interviews structured into three parts: the KEP proposed solution, the rating procedure, and requirements on the graphical user interface.

KEP proposed solution. Four experts (Ex01, Ex02, Ex05, and Ex07) saw the proposed solution as a good way to provide a summary, an overview, and a starting point. However, three experts were not satisfied. Ex03 noted “everything fits to almost everything [...] there is no consistency”. Another expert stated that the solution proposed by the KEP tool misses “the core debate from [his] work package perspective, which is a bit of a shame (Ex04)”. Also Ex06 “noticed that there was nothing that was created [by her] work package”. To clarify: all accessible knowledge elements were included. However, the knowledge elements were perceived as missing because they were not included in the KEP proposed solution or not recognized. Overall, the interviewees were not fully satisfied with the recommendations of the KEP tool as they did not understand the logic behind. The analytics model was too complex and thus it was not clear how their individual input was considered.

Rating procedure. The experts employed a broad range of approaches to provide input for the KEP tool. Experts provided input based on their personal experience and they expected that their input was somehow reflected in the collective rating procedure. Particularly the collective way of gathering input data for the knowledge analytics approach was considered beneficial. One expert was “happy doing the collaborative rating [and finds] it is important [...] for the project to collect this kind of data (Ex05)”. However, two experts also identified challenges of the collective procedure: “you have got people like [A] defending [Topic A], [him] defending [Topic B], [C] defending [Topic C] and [D] defending [Topic D], and that’s [...] to mention a few (Ex04)”. Also Ex06 thinks that the tool could work in a context where “there is not so much of this, let’s say, social rules on the play”. Hence, the interviewees highlighted the need for a collection of ratings for knowledge analytics from different stakeholders, but also emphasized some challenges.

Graphical user interface. Two experts suggested to tag the knowledge elements with keywords or graphic icons. The interviewees also suggested to “have the link between the knowledge element and the corresponding result (Ex05)”. Two experts asked for additional information concerning key measures in the recommendation: “something like a benefit-effort ratio [...] some key measure (Ex07)”, and “of course [what would] be interesting here is the benefits [because] what’s the link now between the benefits and knowledge element and the adaptation criterion? (Ex03)”. Another expert suggested “more aggregate views on the results [and to] slice-and-dice results in a way (Ex03)”. He said, that the data “feels a bit raw” and suggested some “quality aggregators”, some kind of “red-green-orange traffic light thing”, which would “give you an idea of whether the outcome of it is clear cut”. Hence, our interviewees highlighted the importance of useful and meaningful representations of the results of a knowledge analytics tool. Particularly, they considered graphical associations, and aggregated numbers and figures as very important.

5 Candidate Design Patterns

The first candidate design pattern is generalized from the expert statements on the KEP proposed solution (cf. Section 4).

Candidate design pattern (1): “Provenance And Traceability”

Context. The knowledge analytics tool computes a recommendation on the basis of various input (i.e., the ratings). Thereby, the complexity of the proposed solution is very high.

Problem. The users do not accept the proposed solution, because they do not understand why this recommendation was proposed. They would intuitively select other recommendations or they do not understand how their own and others’ ratings were considered by the knowledge analytics tool.

Solution. The proposed solution and the major reasoning for the proposed solution is presented in a suitable way. By showing sub-factors and further details on demand, the analytics approach becomes more transparent. Now explanations for the proposed solution can be presented to the end users, which assumingly will increase their acceptance of the proposed solution.

The provision of provenance and traceability is important for knowledge analytics, because knowledge is a less straightforward object for analytics than data. As a consequence the tool’s suggestions are more difficult to understand for users of knowledge analytics tools. Therefore, following [25] we have to provide both in knowledge analytics approaches: content and context, where context is built through provenance and traceability.

The second candidate design pattern is based on the experts’ reflections on the rating procedure (cf. Section 4).

Candidate design pattern (2): “Human Factor and Stakeholder Rating”

Context. Several users provide their ratings, which reflect their individual knowledge and understanding. Therefore, the ratings stem from several different points of views and backgrounds.

Problem. The ratings are crucial for applying the knowledge analytics tool, concretely for computing the proposed solution. Thus, the ratings should not reflect only one individual perspective, rather it should reflect all relevant ratings.

Solution. The knowledge analytics tool can distinguish user groups and their perspectives on the ratings. It supports a collective approach to rating, the aggregation of ratings and it supports the building of consensus. If no consensus can be reached, then the tool offers the splitting of ratings into different user groups.

The stakeholder involvement and collective rating are considered as crucial for knowledge analytics approaches. As one expert puts it: everyone who has provided a rating has “their own insight knowledge (Ex06)”. The integration of this knowledge and the building of shared mental models should be fostered [23].

The final candidate design pattern is identified from the statements with respect to the graphical user interface (cf. Section 4).

Candidate design pattern (3): “Visualization of a Proposed Solution”

Context. The proposed solution is presented to the users in a spreadsheet which lists the recommended preparation tasks with the selected knowledge elements.

Problem. The visualization of the proposed solution in the spreadsheet was very data-oriented and simple, rather than designed according to the needs of the users. To them, the proposed solution seems raw and clunky.

Solution. The graphical user interface is directed towards the user. It enables the user to customize the interface to different views on the proposed solution. Moreover, the user interface enables to explore the proposed solution in detail.

Our interviewees demanded functionality that enables them to explore the proposed solution according to their own interests, preferences, and needs.

6 Conclusion

Our presented work is one step towards the goal of supporting educational content developers with selecting knowledge elements from a large digital library. To achieve this goal we applied our knowledge analytics approach for workplace learning in the case of the EU FP7 Learning Layers. We leveraged the KEP model and its prototype implementation the KEP tool to compute recommendations, which we used to frame seven artifact-driven expert interviews. We analyzed the interviews qualitatively and based on our findings we developed three candidate design patterns for knowledge analytics: (1) provenance and traceability, (2) human factors and stakeholder rating, and (3) visualization of the proposed solution. The next step is to ground the patterns with theories that explain the effects that the solutions are intended to create and to implement the functionality from our candidate design patterns in the KEP tool in order to validate the patterns.

References

1. Boud, D., Middleton, H.: Learning from others at work: communities of practice and informal learning. *Journal of Workplace Learning* 15(5), 194–202 (2003)
2. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: From big data to big impact. *MIS quarterly* 36(4), 1165–1188 (2012)
3. Collin, K.: Connecting work and learning: design engineers’ learning at work. *Journal of Workplace Learning* 18(7/8), 403–413 (2006)

4. Cooper, A.: What is analytics? Definition and essential characteristics. *CETIS Analytics Series* 1(5), 1–10 (2012)
5. Davenport, T., Harris, J., Morison, R.: *Analytics at Work: Smarter Decisions, Better Results*. Harvard Business Press (2010)
6. Garrick, J.: *Informal learning in the workplace: Unmasking human resource development*. Psychology Press (1998)
7. Goldstein, P.J., Katz, R.N.: *Academic analytics: The uses of management information and technology in higher education*. Tech. rep., EDUCAUSE Center for Analysis and Research (2005)
8. Hart, J.: *Social learning handbook*. Centre for Learning & Performance Technologies (2011)
9. Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., Haydock, M.: Analytics: The widening divide. *MIT Sloan Management Review* 53(2), 1 (2012)
10. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. ACM (2011)
11. Learning Layers D2.3: Tools for networked scaffolding in layers ecosystem. available at learning-layers.eu/deliverables/ (2015)
12. Learning Layers D3.2: Layers tools for creating and maturing instructional material. available at learning-layers.eu/deliverables/ (2014)
13. Maier, R., Thalmann, S.: Using personas for designing knowledge and learning services: results of an ethnographically informed study. *International Journal of Technology Enhanced Learning* 2(1–2), 58–74 (2010)
14. Mayring, P.: *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures And Software Solution*. Klagenfurt (2014)
15. Mor, Y., Mellar, H., Warburton, S., Winters, N.: Practical design patterns for teaching and learning with technology, chap. Introduction: Using Design Patterns to Develop and Share Effective Practice, pp. 1–11. Springer (2014)
16. Parboteeah, P., Jackson, T.W.: Expert evaluation study of an autopoietic model of knowledge. *Journal of Knowledge Management* 15(4), 688–699 (2011)
17. Petter, S., Khazanchi, D., Murphy, J.D.: A design science based evaluation framework for patterns. *ACM SIGMIS Database* 41(3), 9–26 (2010)
18. Pirkkalainen, H., Thalmann, S., Pawlowski, J., Bick, M., Holtkamp, P., Ha, K.H.: Internationalization processes for open educational resources. In: *Workshop on Competencies for the Globalization of Information Systems in Knowledge-Intensive Settings* (2010)
19. Schäper, S., Thalmann, S.: Addressing challenges for informal learning in networks of organizations. In: *23rd European Conference on Information Systems* (2015)
20. Thalmann, S.: Decision support framework for selecting techniques to prepare knowledge elements for adaptive use. Ph.D. thesis, University of Innsbruck (2012)
21. Thalmann, S.: Adaptation criteria for the personalized delivery of learning materials: A multi-stage empirical investigation. *Australasian Journal of Educational Technology* 30(1), 45–60 (2014)
22. Van Barneveld, A., Arnold, K.E., Campbell, J.P.: Analytics in higher education: Establishing a common language. *EDUCAUSE learning initiative* 1, 1–11 (2012)
23. Wang, M., Shen, R.: Message design for mobile learning: Learning theories, human cognition and design principles. *British Journal of Educational Technology* 43(4), 561–575 (2012)
24. Watson, H.J.: Business analytics insight: hype or here to stay. *Business Intelligence Journal* 16(1), 4–8 (2011)
25. Zack, M.H.: Managing codified knowledge. *MIT Sloan Management Review* 40(4), 45 (1999)

Building a Reference Model for Anti-Money Laundering in the Financial Sector

Felix Timm¹, Andrea Zasada¹, Felix Thiede¹

¹University of Rostock, Institute of Computer Science, Rostock, Germany
{felix.timm, andrea.zasada, felix.thiede}@uni-rostock.de

Abstract. Anti-Money Laundering (AML) can be seen as a central problem for financial institutions because of the need to detect compliance violations in various customer contexts. Changing regulations and the strict supervision of financial authorities create an even higher pressure to establish an effective working compliance program. To support financial institutions in building a simple but efficient compliance program we develop a reference model that describes the process and data view for one key process of AML based on literature analysis and expert interviews. Therefore, this paper describes the customer identification process (CIP) as a part of an AML program using reference modeling techniques. The contribution of this work is (i) the application of multi-perspective reference modeling resulting in (ii) a reference model for AML customer identification. Overall, the results help to understand the complexity of AML processes and to establish a sustainable compliance program.

Keywords: Reference Modeling, Compliance Management, Anti-Money Laundering, Financial Sector

1 Motivation and Introduction

The financial industry offers services for individuals and companies to realize money transactions and grant access to numerous financial products such as accounts, shares or credits. Typical financial institutes are banks, insurance and leasing companies. The economic impact of financial activities is enormous. For example, in 2013 the insurance, reinsurance and pension funding in Germany reaches an annual turnover of 251,140 million Euro achieved by only 158,308 employees of 848 companies [1].

The size and structure of the financial industry does not only leave a multitude of financial perspectives but also openings for criminal activities such as money laundering. Money laundering can be described as the process of transforming illegal into legal assets [2]. The *German Institute of Economic Research* (DIW) estimates that about 100 billion Euros are laundered in Germany per year. The observance of regulations that prevent illegal activities like money laundering is ensured by business process compliance management [3]. However, different asset classes and trading platforms make real-time risk and compliance monitoring a challenging and expensive task [4]. The monetary resources that companies need to invest in their compliance management include

implementation, remediation, and penalty associated costs [5]. A global survey with 200 hedge fund managers reveals that almost two thirds (64 percent) of respondents were spending over 5 percent in 2013 of their total operating costs on meeting compliance requirements [6].

The implementation of compliant business processes requires the collaboration of all involved stakeholders, such as compliance officers, IT and legal experts to build a reference model including necessary compliance requirements. The formal description of compliance requirements can be effectively supported by conceptual modeling techniques. These techniques are applied to improve the understanding and communication among stakeholders, which helps to prevent legal violations and reduces the operating costs of compliance management. An essential part of the compliance management of financial institutes is constituted by *anti-money laundering* (AML) regulations. In literature, the term reference model is often related to the *Enterprise Architecture Management* (EAM). EAM is used to reduce the complexity of business activities to create reference models abstracted from reality [7]. Current approaches tackle single information systems disciplines like e-government and miss so far a documented procedure to build the corresponding reference model [8]. With this paper, we drive attention to AML regulations and the necessity to develop a reference model that facilitates the application of compliance requirements in the financial industry. The research questions (RQs) are:

RQ1: *Which compliance regulations have to be adopted for AML prevention?*

RQ2: *How should a reference model for the AML CIP be constituted?*

The paper is structured as follows. In Section 2 we discuss the process of money laundering and give an overview on AML regulations and best practices. After presenting the research method in Section 3, we introduce the reference model for AML customer identification in Section 4. In the end, we discuss the evaluation approach in Section 5 before we conclude our work in Section 6.

2 Anti-Money Laundering Regulations

In general, the money laundering process consists of three stages: placement, layering and integration [9]. At the first stage illegal money is placed at a bank account. By using an account with a low risk, the money launderer avoids to be detected by authorities. At the layering stage the money is transferred from one to several other accounts, which lowers the chance that law enforcement detects and follows the money flow. At the last stage the money is actually laundered by investing in legal businesses like property or luxury articles [10]. Research indicates that effective AML is a resource-intensive quest and benefits from the collection, maintenance and dissemination of customer related information [11]. Another positive impact on AML can be observed regarding the employee work attitude and training [12].

Laws and guidelines are describing general principles and criteria to establish an AML process and to assign appropriate control activities. This paper covers the German *Money Laundry Law* (GwG) [13], the guidelines published by the *Federal Financial*

Supervisory Authority (BaFin) [16] and the two international financial supervision committees namely *Financial Action Task Force* (FATF) [14] and the *Wolfsberg Group* [15]. The GwG explains various levels of diligence that can be used by financial institutions to identify the customer or guarantee the *Know Your Customer* (KYC) principle. KYC means that financial institutions have to implement a suitable system of internal controls and policies to identify their customers and suspicious transactions [17]. It describes fines for financial institutions if their money laundering detection fails or AML mechanisms have not been implemented [18]. The BaFin publishes lists of non-cooperative countries and territories that can be used to identify single financial institutions which follow the law. Moreover, BaFin suggests guidelines that support the customer identification and the ascertainment of the beneficial owner of a company. The FATF recommends that financial institutions should establish compliance programs to prevent money laundering and counter terrorism [19]. The Wolfsberg Group, an association of eleven global financial institutions, built an industrial standard for compliance [20]. It is known as the Wolfsberg principles and motivates financial institutions to exchange information on AML cases [21]. Financial institutions that cannot establish these principles are disclosed [22]. To ensure that laws and guidelines are met, financial institutions have to establish an organizational framework to identify money laundering cases [23]. The steps of building an AML program are described in Table 1. After identifying and adapting financial regulations and guidelines, risk phenomena are measured [20]. Depending on the results, the AML process is defined usually supported by appropriate software [20]. Many guidelines also suggest to install organizational structures, which should at least encompass a compliance officer, whose task is to decide which countermeasures to take [12].

Table 1. AML program for financial institutions

Phase	Step	Name	Description	Example
Planning	1	Identify regulations	Compliance with legal requirements and official guidelines.	Wolfsberg principles
	2	Derive company guideline	Internal rules for handling money laundering cases.	Code of conduct
	3	Conduct risk analysis	Risk analysis for risk classes related to customer, product or location.	Money transaction
	4	Define process and control activities	Specification of the anti-money laundering process and control activities.	Customer identification
	5	Implement control system	Establishing of working routines and software for monitoring and reporting.	Business application
	6	Define control structure	Organizational function for money laundering reports to top management.	Report
Controlling	7	Define organ. function	Department for handling money laundering cases and conducting risk analyses.	Department
	8	Appoint representative	Head of the anti-money laundering department.	Agent
	9	Conduct employee training	Regular trainings and briefings on relevant regulations and the compliance program.	Seminar
	10	Conduct internal and external audits	Identification of deficiencies of the established compliance program.	Consultant

3 Methodological Approach

This approach integrates the process and data perspective in one reference model for a central AML process. The aim is to develop a reference model for AML exemplified for the CIP. We therefore conducted a literature analysis on common AML regulations as described in Section 2. To holistically capture the CIP and the data-centric nature of its related KYC paradigm a reference model should consider different perspectives. For developing such a multi-perspective reference model we adapted the procedure model by Rosemann and Schütte (1999) [24] because it explicitly defines different perspectives on the problem domain. The model consists of five phases: (1) Problem identification, (2) Design of the reference model frame, (3) Design of the reference model structure, (4) Finalization of the reference model and (5) Application of the reference model. The scope of this paper comprises phase (1) to (4) which are described in Section 4.

In the first phase a problem definition is given to determine modeling objectives, e.g. reducing the model complexity or improve the process efficiency. This requires a detailed process description including relevant regulations, stakeholders and modelling perspectives, e.g. process, data, application or technology [25]. As we are addressing customer identification and KYC in our approach, we will focus on the process and data perspective. The second phase is dedicated to the method applied for process modeling and a first sketch of the process, for which we propose the common *Business Process Model and Notation* (BPMN 2.0) standard. The third phase deals with the actual design of the process model, while phase four is used to enrich the process model with business data and evaluate the model constraints, for which we used the literature analysis. In phase four we conducted two expert interviews with senior IT consultants to complete the process information that has been gained from literature. The experts work for different IT vendors specialized for compliance software in the financial sector. Given their longtime experiences in supporting their customers (i.e. financial institutes) in implementing a successful AML program, we consider them as appropriate experts for our purpose.

4 Reference Model for Anti-Money Laundering

The processes of an institute's AML program can be seen as supporting processes related to the daily routines of the banking business, such as account opening, payments or account management. For instance, each transaction made by a customer will be monitored in terms of AML parameters like the transaction's amount. Further, the AML program can be divided into four different activities. The (i) AML hazard analysis is an upstream process, which analyzes all risks that are related to AML such as customer- or location-related risks. It results in a risk matrix used to assess a certain customer's likelihood to launder money. The (ii) CIP is triggered every time the institute enters a new business relationship with a customer [20]. This implies to follow the KYC principle discussed earlier. Every transaction is monitored during the (iii) *transaction monitoring process* and checked against threshold values depending on the customer's risk

assessment. Every suspicious activity triggers the (iv) *AML case handling* [20]. According to the experts, process (iii) is usually automated. Thus, we excluded it from our reference model. Process (i) is often performed with a global perspective on the institute, where AML risks are a subset of the holistic risk scheme. Although we consider process (i) vital for correct AML, we excluded it due to space limitations. In consequence, we focused on the processes (ii) and (iv) when performing reference modeling. In the following section, we will present the (ii) CIP of an AML program in BPMN and the KYC principles from a data perspective.

4.1 Reference Process Perspective on AML

The main source of information is a literature analysis we performed. On basis of the identified literature the first version of the reference models emerged. Then, two expert interviews were conducted. The resulting models are presented in the following section.

The AML CIP is triggered every time a new customer enters a relationship with the institute. Next to the usual customer data handling, on the one hand financial institutes face strict requirements by law in terms of data complexity, validation and screening. On the other hand, institutes have to assess the customer's risk regarding money laundering in order to adjust their AML monitoring systems and research activities. Three types of sources were used to model the reference process model. First, results from the literature analysis [20, 22, 26, 27] served as a process foundation. Second, laws and directives from different authorities were analyzed [13, 28, 29]. Third, known recommendations and best practices were incorporated into the reference model [14, 15, 30]. For the final reference model we use the BPMN 2.0 notation for the process perspective, which is visualized in Fig. 1.

There are five roles acting in the process, which are represented by the BPMN 2.0 swim lanes. While the customer and a service provider are modeled as a black box lane, the collaboration between three generic departments is modeled. First, the customer's account representative (AR) receives several sources of data from the customer during the *customer identification*. The amount of data depends on the customer's type (see Section 4.2). The institute needs defined internal guidelines for correct and complete customer data collection derived from national or international law. The guidelines also define how to *validate the customer's identity* by using official service providers like *Office of Foreign Assets Controls (OFAC)* or internal identity list. The next step *identify customer's purpose of usage* is important to predict future account movements and relate the customer to a risk cluster. Subsequently, the *AML employee (AE)* uses the validated customer data to assess her or his risk profile. Therefore, the *customer screening* compares the customer's identity with existing AML lists. For instance, the institute has to be aware whether the new customer is a *political exposed person (PEP)* or named in an official sanction list. Most of these lists can be accessed by service providers such as OFAC or *World Check by Thomson Reuters*. The institute should define against which lists the customer has to be checked. Afterwards, the AE *assesses the customer's risk* based on the risk matrix defined by the AML hazard analysis. The results of this risk assessment are then integrated into the monitoring system. The monitoring sys-

tem's threshold values are set depending on the customer's risk profile. The more precise and diligent the customer data is assessed, the more exact the monitoring system works. For instance, when a PEP, whose AML risk is set as high, receives a transaction from a country, which AML list providers rate as highly corrupt, the transaction can be identified as a possible AML case. This case will then be handled by the process (iv) *AML case handling*. The reference model in Fig. 1 also includes a BPMN 2.0 model of the AML hazard analysis with a low level of detail to highlight the dependencies with the customer identification. It is performed by an employee of the risk management department. In general, the institute has to decide which risk phenomena related to the institute contain AML risk and can be measured. For each of these phenomena values are defined, from which scenarios are derived instantiating the different values. These scenarios are assessed regarding their likelihood to represent an AML case. Usually, this is done by defining an AML risk of a scenario from low over medium and high to unacceptable. When a customer's risk is assessed, his profile is related to these scenarios.

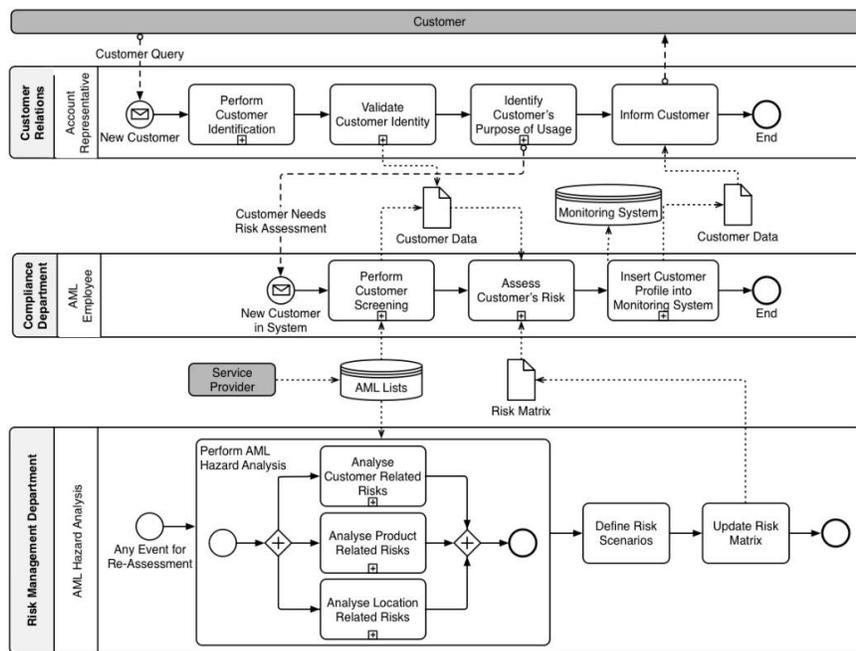


Fig. 1. The CIP reference process in an AML program

4.2 Reference Data Perspective on AML

In the CIP the processes data is customer data, risk matrix and AML lists. After structuring the identified data into these clusters, more detailed data structures were built. This was discussed within the expert interviews, which also served as an information source. Table 2 summarizes the data structure. The findings reveal that most of the analyzed data is related to the customer. Depending on the type of customer (natural or corporate) the required data fields differ. While the identification of private customers is primarily limited to personal information, the research activities of the AR and AE from Fig. 1 are very complex and cost-intense (e.g. to identify the beneficial owners). Furthermore, the type of relationship the customer enters with the institute needs to be distinguished in the stated data fields. The data about the customer and her or his business relationship with the institute are used to assess the customer's risk level. This is based on the prior developed risk matrix. Table 2 shows which risk phenomena should be captured in order to build AML risk categories, e.g. risks related to the customer, countries or even the institute's employees. The AML lists that are used for customer identification are usually provided by third parties (see Section 4.1). The structure in Table 2 serves as a general data view on the CIP. The elaboration of the concrete data objects exceeds the scope of the paper. In general, the complexity of data used for customer identification depends on its context, i.e. the type of customer and its environment. The authors derive a strong dependency between the process and data perspective in the CIP. For instance, the sub-tasks of *validate customer identity* changes with the type of customer. We identify the need to incorporate these dependencies within the reference model. The current reference model uses BPMN 2.0, which is restricted to model the control flow and lacks profound data modeling. Therefore, we suggest the *Enterprise Architectures* (EA) concept as a possible alternative to the current model structure. EAs capture the structure of an organization from different perspectives (e.g. business, data, application and technology layer) and reveal their interdependencies [25]. This would add value for institutes to identify the dependencies not only regarding AML but their whole compliance organization.

5 Evaluation of the Reference Model

The development of a reference model is an iterative process. This process is characterized by different versions of the considered model. The reference model should be evaluated using a validation method, which may lead to adjustments of the reference model [24]. In this work two iteration loops were traversed. Therefore, semi-structured telephone interviews with experts of two different vendors for financial compliance software were conducted [31]. While the first iteration loop concentrated on the process perspective, the second iteration loop focused on the data perspective of the AML program. The experts assessed the reference model as content wise correct, mentioning that the detailed sub-tasks may differ among different institutes. Furthermore, they pointed out that the usage of a complete data structure inside the institutes has a significant influence on the AML program's success. The expert interviews provided most input for the data perspective, which most literature did not discuss in detail.

Table 2. Data perspective of the CIP

Data Object	Contained Information	
<i>Customer Data</i>	A) Natural Person: <ul style="list-style-type: none"> • Personal data (e.g. name and nationality) • Occupation and industry • Sources of wealth • Relationships to other clients • Data of business relationship 	B) Corporate Identity: <ul style="list-style-type: none"> • Industry and legal form • Places of business (national vs. global) • Beneficial owners • Organizational structure • Data of business relationship
<i>Business Relationship</i>	<ul style="list-style-type: none"> • Purpose of account or product • Total assets 	<ul style="list-style-type: none"> • Type of account, currency and account opening • Predication of transactions
<i>Risk Matrix</i>	<ul style="list-style-type: none"> • Customer related • Product related • Country related • Business process related 	<ul style="list-style-type: none"> • Employee related • Transaction related • Information systems related • Derived risk categories
<i>AML Lists</i>	<ul style="list-style-type: none"> • PEP and related lists • Sanction lists • Black lists 	<ul style="list-style-type: none"> • Internal lists • Country risk lists

6 Conclusion

This work addresses the need of financial institutes to meet regulatory requirements defined on national and international level. Therefore, we present the results of applying multi-perspective reference modeling by Rosemann and Schütte for an AML program based on a literature analysis and expert interviews [24]. By analyzing related literature, legislative texts and recommendations from practitioners' working groups, requirements for an AML program have been derived (RQ1). On the basis of these results and two expert interviews, a reference model was developed capturing process and data perspectives of the CIP in an AML program (RQ2). From theoretical point of view this work contributes how to apply reference modeling. Further, practitioners can benefit from this approach in terms of evaluating their current practice of an AML program. Nevertheless, the authors want to point out that the used data base may not be complete in order to provide a sufficient level of detail of the reference model. Moreover, the interviewed experts may be biased since they represent the interests of their respective enterprise. In consequence, the authors see multiple areas for future research in this topic. First, the data base could be enriched by conducting interviews or workshops at the institutes' in order to gather their current state and identify practitioners' best practices, which would result in applying inductive reference modeling [32]. Second, the proposed reference model could be extended by concepts of EA. Finally, broadening the horizon to other domains of financial compliance like regulatory reporting might identify synergies among different data models, which then would be represented by a holistic reference model.

Acknowledgements. This work has been supported by the BITKOM funded project "IT gestützte Compliance im Finanzsektor".

References

1. Statistisches Bundesamt. (2016) Statistisches Jahrbuch 2015: Kapitel 27: Weitere Dienstleistungen
2. Friedrich Schneider, Ursula Windischbauer. (01/01/10) Money Laundering: Some Facts. Economics of Security Working Paper Series
3. Steffen Höhenberger, Dennis M Riehle, Patrick Delfmann. (2016) From Legislation to Potential Compliance Violations in Business Processes. Simplicity Matters. In: Proceedings of the European Conference on Information Systems (ECIS 2016), Istanbul, Turkey
4. Deloitte Center for Financial Services (2016) Banking reimaged: How disruptive forces will radically transform the industry in the decade ahead
5. Abdullah NS, Indulska M, Sadiq S (2016) Compliance management ontology – a shared conceptualization for research and practice in compliance management. *Inf Syst Front*. doi: 10.1007/s10796-016-9631-4
6. KPMG International The Cost of Compliance: 2013 KPMG/AIMA/MFA Global Hedge Fund Survey
7. ten Harmsen van der Beek, Wijke, Trienekens J, Grefen P (2012) The Application of Enterprise Reference Architecture in the Financial Industry. In: Aier S, Ekstedt M, Matthes F et al. (eds) Trends in Enterprise Architecture Research: 7th Workshop, TEAR 2012, Barcelona, Spain, October 23-24, 2012. Proceedings, vol 131. Springer, Berlin, Heidelberg, pp 93–110
8. Tambouris E, Kaliva E, Liaros M et al. (2014) A reference requirements set for public service provision enterprise architectures. *Softw Syst Model* 13(3): 991–1013. doi: 10.1007/s10270-012-0303-7
9. Reuter P, Truman EM (2004) Chasing dirty money: The fight against money laundering. Inst. for Internat. Economics, Washington, DC
10. Angela Samantha Maitland Irwin, Kim-Kwang Raymond Choo, Lin Liu (2012) Modelling of money laundering and terrorism financing typologies. *Journal of Money Laundering Control* 15(3): 316–335. doi: 10.1108/13685201211238061
11. Kemal MU (2014) Anti-money laundering regulations and its effectiveness. *Journal of Money Laundering Control* 17(4): 416–427. doi: 10.1108/JMLC-06-2013-0022
12. Hung Kwok TS Anti money laundering (“AML”) management and the importance of employees' work attitude. In: 2013 International Conference on Engineering, Management Science and Innovation (ICEMSI), pp 1–4
13. German Government (2008) Gesetz über das Aufspüren von Gewinnen aus schweren Straftaten (Geldwäschegesetz - GwG). https://www.gesetze-im-internet.de/bundesrecht/gwg_2008/gesamt.pdf
14. Financial Action Task Force (2013) National money laundering and terrorist financing risk assessment. http://www.fatf-gafi.org/media/fatf/content/images/National_ML_TF_Risk_Assessment.pdf
15. Wolfsberg Group (2009) Wolfsberg AML Guidance on Credit/Charge Card Issuing and Merchant Acquiring Activities. [http://www.wolfsberg-principles.com/pdf/standards/Wolfsberg_Credit_Cards_AML_Guidance_\(2009\).pdf](http://www.wolfsberg-principles.com/pdf/standards/Wolfsberg_Credit_Cards_AML_Guidance_(2009).pdf)
16. Federal Financial Supervisory Authority (since 2011) Circulars on Anti-Money Laundering.

https://www.bafin.de/EN/Aufsicht/Geldwaeschebekaempfung/geldwaeschebekaempfung_node_en.html

17. Olatunde Julius Otusanya, Solabomi Omobola Ajibolade, Eddy Olajide Omolehinwa (2011) The role of financial intermediaries in elite money laundering practices: Evidence from Nigeria. *Journal of Money Laundering Control* 15(1): 58–84. doi: 10.1108/13685201211194736
18. Bundesverband Deutscher Leasing-Unternehmen (2012) Anwendungsempfehlungen zur Geld- wäschebekämpfung bei Leasing-Unternehmen
19. Lagzdins A, Sloka B (2012) COMPLIANCE PROGRAM IN LATVIAS' BANKING SECTOR: THE RESULTS OF A SURVEY. *EurInsStud* 0(6). doi: 10.5755/j01.eis.0.6.1612
20. Sullivan K (ed) (2015) Anti-money laundering in a nutshell: Awareness and compliance for financial personnel and business. Apress, Berkley
21. Pieth M, Aiolfi G (2003) Anti-Money Laundering: Levelling the Playing Field
22. Verhage A (2009) Between the hammer and the anvil?: The anti-money laundering-complex and its interactions with the compliance industry. *Crime Law Soc Change* 52(1): 9–32. doi: 10.1007/s10611-008-9174-9
23. Basel Committee (2005) Compliance and the compliance function in banks.
24. Rosemann M, Schütte R (1999) Multiperspektivische Referenzmodellierung. In: Becker J, Rosemann M, Schütte R (eds) Referenzmodellierung: State-of-the-Art und Entwicklungsperspektiven. Physica-Verlag HD, Heidelberg, pp 22–44
25. Ahlemann F, Stettiner E, Messerschmidt M et al. (2012) Strategic Enterprise Architecture Management. Springer Berlin Heidelberg, Berlin, Heidelberg
26. Smet D de, Mention A (2011) Improving auditor effectiveness in assessing KYC/AML practices. *Managerial Auditing Journal* 26(2): 182–203. doi: 10.1108/02686901111095038
27. Tang J, Ai L (2013) The system integration of anti-money laundering data reporting and customer relationship management in commercial banks. *J of Money Laundering Control* 16(3): 231–237. doi: 10.1108/JMLC-04-2013-0010
28. European Union (2015) DIRECTIVE (EU) 2015/849 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015L0849&from=DE>
29. US Government (2001) USA Patriot Act. <https://www.gpo.gov/fdsys/pkg/PLAW-107publ56/pdf/PLAW-107publ56.pdf>
30. Financial Action Task Force (2012) International standards on the combating of money laundering and the financing of terrorism and proliferation. The FATF recommendations. http://www.fatf-gafi.org/media/fatf/documents/recommendations/pdfs/FATF_Recommendations.pdf
31. Runeson P, Höst M (2009) Guidelines for conducting and reporting case study research in software engineering. *Empir Software Eng* 14(2): 131–164. doi: 10.1007/s10664-008-9102-8
32. Loos P, Fettke P, Walter J et al. (2015) Identification of Business Process Models in a Digital World. In: Vom Brocke J, Schmiedel T (eds) BPM - driving Innovation in a digital world. Springer, Cham, pp 155–174

Is Web Content a Good Proxy for Real-Life Interaction?

A Case Study Considering Online and Offline Interactions of Computer Scientists

Mark Kibanov¹, Martin Atzmueller¹, Jens Illig¹,
Christoph Scholz², Alain Barrat³, Ciro Cattuto⁴, and Gerd Stumme¹

¹ University of Kassel, Knowledge and Data Engineering Group
{atzmueller, kibanov, illig, stumme}@cs.uni-kassel.de

² Fraunhofer Institute for Wind Energy and Energy System Technology
christoph.scholz@iwes.fraunhofer.de

³ Aix Marseille Université, Université de Toulon CNRS, CPT, UMR 7332
alain.barrat@cpt.univ-mrs.fr

⁴ Data Science Laboratory, ISI Foundation
ciro.cattuto@isi.it

Online social relationship can represent a significant share of an individual's *social profile*, in addition to the interactions that take place *offline*, e. g., when meeting friends, during a face-to-face conversation, etc. So far, the analysis of online social networks has received significant attention in the research community, while studies on offline interactions at large scale, e. g., focusing on networks of face-to-face proximity, has only been taken up recently.

We start filling this gap by analyzing *online* and *offline* social networks and published content of computer scientists who visited particular conferences. We investigate whether information from various online sources can be used as a proxy for the offline world.

Our results [1] indicate that in many cases online and offline datasets still have structural differences at large. However, strong ties seem to correlate better than weak ones and for “important persons” there are proxy relations between the online and the offline world. We also observed examples of a successful application of online data for offline scenarios, specifically relating to the analysis of content and characteristic subgroups. Our results show that the considered online data is not an ideal proxy for offline information overall, but still provides some important indications about offline relationships.

References

1. Kibanov, M., Atzmueller, M., Illig, J., Scholz, C., Barrat, A., Cattuto, C., Stumme, G.: Is web content a good proxy for real-life interaction? a case study considering online and offline interactions of computer scientists. In: Proceedings of the 2015 ACM/IEEE International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, Paris, France, August 25-28, 2015 (2015)

Predicting video game properties with deep convolutional neural networks using screenshots

Przemysław Buczkowski^{1,2*}, Antoni Sobkowicz¹

¹National Information Processing Institute, Warsaw, Poland

²Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland
pbuczkowski@opi.org.pl

The amount of visual data is enormous nowadays and it constantly increases. On the other hand information in images is still almost not readable for computer algorithms. In most of image classification problems humans still outperform computers. Sometimes it is not obvious who will perform better, because little is known about the classification problem. Good example of such case might be predicting various video game properties based only on in-game screen shots.

The authors have crawled the Steam platform storing information about genre, date of release, PEGI rating, etc. for thousands of video games. It is difficult to pinpoint a model that can infer those properties by looking only on images of gameplay. During research authors have focused on using standard deep convolutional neural networks in the first part of the network and fully connected layers at the end. As well as working on model we are preparing human annotator based baseline benchmark.

Interesting regularities have been found, e.g. approximation of release year is very difficult task on small images, because lot of details are lost. Downscaling an image is similar to anti-aliasing - in-game models look much less edgy. Both humans and CNNs struggle with this task. For now it is hard to tell who performs better because of small amount of human-annotated data. One of the notices made during the research was that people who don't describe themselves as gamers perform poorly on this task, and experienced players recognise the game from the image and then recall release date from memories. In the nearest future authors plan to use bigger images or even 1:1 crops for the task of release year approximation and extend our database of human-annotated images.



Fig. 1. Low resolution examples of screens: The Book of Unwritten Tales 2 (2015) and Counter Strike (2000).

Towards a case-based reasoning approach for cloud provisioning

Eric Kübler and Mirjam Minor

Wirtschaftsinformatik, Goethe University, Robert-Mayer-Str.10, Frankfurt am Main,
Germany,

{ekuebler, minor}@informatik.uni-frankfurt.de

Abstract. Resource provisioning is an important issue of cloud computing. Most of the recent cloud solutions implement a simple way with static thresholds to provide resources. Some more sophisticated approaches consider the cloud provisioning problem a multi-dimensional optimization approach. However, the calculation effort for solving optimization problems is significant. An intelligent resource provisioning with a reduced calculation effort requires smart cloud management methods. In this position paper, we propose a case-based reasoning approach for cloud management. A case records a problem situation in cloud management and its solution. We introduce a case model and a retrieval method for previously solved problem cases with the aim to reuse their re-configuration actions for a recent problem situation. The case model uses the container notion correlated with QoS problems. We present a novel, composite similarity function that allows to compare a recent problem situation with the cases from the past. During retrieval, the similarity function creates a ranking of the cases according to their relevance to the current problem situation. Further, we describe the prototypical implementation of the core elements of our case based-reasoning concept. The plausibility of the retrieval approach has been tested by means of sample cases with simulated data. The original version of this re-submission has been published at CLOSER 2016 [Kübler and Minor, 2016].

Keywords: cloud management, case-based reasoning, intelligent cloud provisioning

References

- [Kübler and Minor, 2016] Kübler, E. and Minor, M. (2016). Towards a case-based reasoning approach for cloud provisioning. In *Proceedings of the 6th International Conference on Cloud Computing and Services Science (Accepted for publication)*, volume 2, pages 290–295, Rome, Italy. SciTePress.

Modern Tools for Old Content – in Search of Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910

Kimmo Kettunen¹, Eetu Mäkelä², Juha Kuokkala³, Teemu Ruokolainen⁴, Jyrki Niemi⁵

¹ National Library of Finland, Centre for Preservation and Digitization, Mikkeli, Finland
kimmo.kettunen@helsinki.fi

² Aalto University, Semantic Computing Research Group, Espoo, Finland
eetu.makela@aalto.fi

³ University of Helsinki, Department of Modern Languages, Helsinki, Finland
juha.kuokkala@helsinki.fi

⁴ Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
teemu.ruokolainen@aalto.fi

⁵ University of Helsinki, Department of Modern Languages, Helsinki, Finland
jyrki.niemi@helsinki.fi

Abstract. Named entity recognition (NER), search, classification and tagging of names and name like frequent informational elements in texts, has become a standard information extraction procedure for textual data. NER has been applied to many types of texts and different types of entities: newspapers, fiction, historical records, persons, locations, chemical compounds, protein families, animals etc. In general a NER system's performance is genre and domain dependent and also used entity categories vary [1]. The most general set of named entities is usually some version of three partite categorization of locations, persons and organizations. In this paper we report first trials and evaluation of NER with data out of a digitized Finnish historical newspaper collection Digi. Digi collection contains 1,960,921 pages of newspaper material from years 1771–1910 both in Finnish and Swedish. We use only material of Finnish documents in our evaluation. The OCRed newspaper collection has lots of OCR errors; its estimated word level correctness is about 74–75 % [2]. Our principal NER tagger is a rule-based tagger of Finnish, FiNER, provided by the FIN-CLARIN consortium. We show also results of limited category semantic tagging with tools of the Semantic Computing Research Group (SeCo) of the Aalto University. FiNER is able to achieve up to 60.0 F-score with named entities in the evaluation data. SeCo's tools achieve 30.0–60.0 F-score with locations and persons. Performance of FiNER and SeCo's tools with the data shows that at best about half of named entities can be recognized even in a quite erroneous OCRed text.

Keywords: named entity recognition, historical newspaper collections, Finnish

1 Introduction

The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910 [3, 4]. This collection contains 1,960,921 million pages in Finnish and Swedish. Finnish part of the collection consists of about 2.39 billion words. The National Library's Digital Collections are offered via the *digi.kansalliskirjasto.fi* web service, also known as Digi. Part of the newspaper material (years 1771–1874) is freely downloadable in The Language Bank of Finland provided by the FIN-CLARIN consortium¹. The collection can also be accessed through the Korp² environment that has been developed by Språkbanken at the University of Gothenburg and extended by FIN-CLARIN team at the University of Helsinki to provide concordances of text resources. A Cranfield style information retrieval test collection has been produced out of a small part of the Digi newspaper material at the University of Tampere [5].

The web service *digi.kansalliskirjasto.fi* is used, for example, by genealogists, heritage societies, researchers, and history enthusiast laymen. There is also an increasing desire to offer the material more widely for educational use. In 2015 the service had about 14 million page loads. User statistics of 2014 showed that about 88.5 % of the usage of the Digi came from Finland, but an 11.5 % share of use was coming outside of Finland.

Named entity recognition has become one of the basic techniques for information extraction of texts. In its initial form NER was used to find and mark semantic entities like person, location and organization in texts to enable information extraction related to these kinds of entities. Later on other types of extractable entities, like time, artefact, event and measure/numerical, have been added to the repertoires of NER software [1], [6].

Our aim with usage of NER is to provide users of Digi better means for searching and browsing of the historical newspapers. Different types of names, especially person names and names of locations are used frequently as search terms in different newspaper collections [7]. They can provide also browsing assistance to collections, if the names are recognized and tagged in the newspaper data and put into the index [8]. A fine example of usage of name recognition with historical newspapers is La Stampa's historical newspaper collection³. After basic keyword search users can browse or filter the search results by using three basic NER categories of person (authors of articles or persons mentioned in the articles), location (countries and cities mentioned in the articles) and organization. Thus entity annotations of newspaper text allow a more semantically-oriented exploration of content of the large archive. A large scale (152 M articles) NER analysis and usage examples of the Australian historical newspaper collection Trove is described in Mac Kim and Cassidy [9].

¹ <https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiAineistotDigilibPub>

² <https://korp.csc.fi/>

³ <http://www.archiviolastampa.it/>

2 NER Software and Evaluation

For recognition and labelling of named entities we use principally FiNER software. SeCo's ARPA is of different type, it is mainly used for Semantic Web tagging and linking of entities [10]⁴, but it could be adapted for basic NER, too. Before choosing FiNER we also tried a commonly used trainable free tagger, Stanford NER⁵, but were not able to get reasonable performance out of it for our purposes.

FiNER is a rule-based named-entity tagger, which in addition to surface text forms utilizes grammatical and lexical information from a morphological analyzer (Omorfi⁶). FiNER pre-processes the input text with a morphological tagger derived from Omorfi. The tagger disambiguates Omorfi's output by selecting the statistically most probable morphological analysis for each word token, and for tokens not recognized by the analyzer, guesses an analysis by analogy of word-forms with similar ending in the morphological dictionary. The use of morphological pre-processing is crucial in performing NER with a morphologically rich language such as Finnish, where a single lexeme may theoretically have thousands of different inflectional forms.

The focus of FiNER is in recognizing different types of proper names. Additionally, it can identify the majority of Finnish expressions of time and e.g. sums of money. FiNER uses multiple strategies in its recognition task:

1) Pre-defined gazetteer information of known names of certain types. This information is mainly stored in the morphological lexicon as additional data tags of the lexemes in question. In the case of names consisting of multiple words, FiNER rules incorporate a list of known names not caught by the more general rules.

2) Several kinds of pattern rules are being used to recognize both single- and multiple-word names based on their internal structure. This typically involves (strings of) capitalized words ending with a characteristic suffix such as Inc, Corp, Institute etc. Morphological information is also utilized in avoiding erroneously long matches, since in most cases only the last part of a multi-word name is inflected, while the other words stay in the nominative (or genitive) case. Thus, preceding capitalized words in other case forms should be left out of a multi-word name match.

3) Context rules are based on lexical collocations, i.e. certain words which typically or exclusively appear next to certain types of names in text. For example, a string of capitalized words can be inferred to be a corporation/organization if it is followed by a verb such as *tuottaa* ('produce'), *työllistää* ('employ') or *lanseerata* ('launch' [a product]), or a personal name if it is followed by a comma- or parenthesis-separated numerical age or an abbreviation for a political party member.

The pattern-matching engine that FiNER uses, HFST Pmatch, marks leftmost longest non-overlapping matches satisfying the rule set (basically a large set of disjuncted patterns) [11, 12]. In the case of two or more rules matching the exact same passage in the text, the choice of the matching rule is undefined. Therefore, more

⁴ An older demo version of the tool is available at <http://demo.seco.tkk.fi/sarpa/#/>

⁵ <http://nlp.stanford.edu/software/CRF-NER.shtml>

⁶ <https://github.com/flammie/omorfi>

control is needed in some cases. Since HFST Pmatch did not contain a rule weighing mechanism at the time of designing the first release of FiNER, the problem was solved by applying two runs of distinct Pmatch rulesets in succession. This solves for instance the frequent case of Finnish place names used as family names: in the first phase, words tagged lexically as place names but matching a personal name context pattern are tagged as personal names, and the remaining place name candidates are tagged as places in the second phase. FiNER annotates 15 different entities that belong to five categories: location, person, organization, measure and time [12].

SeCo's ARPA [10] is not actually a NER tool, but instead a dynamic, configurable entity linker. In effect, ARPA is not interested in locating all entities of a particular type in a text, but instead locating all entities that can be linked to strong identifiers elsewhere. Through these, it is then for example possible to source coordinates for identified places, or associate different name variants and spellings to a single individual. For the pure entity recognition task presented in this paper, ARPA is thus at a disadvantage. However, we wanted to see how it would fare in comparison to FiNER.

The core benefits of the ARPA system lie in its dynamic, configurable nature. In processing, ARPA combines a separate lexical processing step with a configurable SPARQL-query -based lookup against an entity lexicon stored at a Linked Data endpoint. Lexical processing for Finnish is done with a modified version of Omorfi⁷, which supports historical morphological variants, as well as lemma guessing for out of vocabulary words. This separation of concerns allows the system to be speedily configured for both new reference vocabularies as well as the particular dataset to be processed.

2.1 Evaluation Data

As evaluation data for FiNER we used samples from the Digi collection. Kettunen and Pääkkönen [2] calculated among other things number of words in the data for different decades. It turned out that most of the newspaper data was published in 1870–1910, and beginning and mid of the 19th century had much less published material. About 95 % of the material was printed in 1870–1910, and most of it, 82.7 %, in two decades of 1890–1910.

We aimed at an evaluation collection of 150,000 words. To emphasize the importance of the 1870–1910 material we took 50 K of words from time period 1900–1910, 10 K from 1890–1899, 10 K from 1880–1889, and 10 K from 1870–1879. Rest 70 K of the material was picked from time period of 1820–1869. Thus the collection reflects most of the data from the century but is also weighed to the end of the 19th century and beginning of 20th century.

The final manually tagged evaluation data consists of 75,931 lines, each line having one word or other character data. The word accuracy of the evaluation sample is on the same level as the whole newspaper collection's word level quality: about 73 % of the words can be recognized by a modern Finnish morphological analyzer [2]. 71

⁷ <https://github.com/jiemakel/omorfi>

% of the tagger’s input snippets have five or more words, the rest have fewer than five words in the text snippet.

FiNER’s 15 tags for different types of entities is too fine a distinction for our purposes. Our first aim was to concentrate only on locations and person names, because they are mostly used in searches of the Digi collection, as was detected in an earlier log analysis [4]. After reviewing some of the FiNER tagged material, we included also three other tags, as they seemed important and were occurring frequently enough in the material. The final chosen eight tags are shown and explained below.

Entity/tag	Meaning
1. <EnameXPrsHum>	person
2. <EnameXLocXxx>	general location
3. <EnameXLocGpl>	geographical location
4. <EnameXLocPpl>	political location (state, city etc.)
5. <EnameXLocStr>	street, road, street address
6. <EnameXOrgEdu>	educational organization
7. <EnameXOrgCrp>	company, society, union etc.
8. <TimexTmeDat>	expression of time

The final entities show that our interest is mainly in the three most used semantic NER categories: persons, locations and organizations. With locations we use two sub-categories and with organizations one. Temporal expressions were included in the tag set due to their general interest in the newspaper material.

Manual tagging of the evaluation material was done by the fourth author, who had previous experience of tagging modern Finnish with tags of the FiNER tagger. Tagging took one month, and quality of the tagging and its principles were discussed before starting based on a sample of 2000 lines of evaluation data. It was agreed, for example, that words that are misspelled but are recognizable for the human tagger as named entities would be tagged (cf. 50 % character correctness rule in Packer et al. [15]). If orthography of the word was following 19th century spelling rules, but the word was identifiable as a named entity, it would be tagged, too.

2.2 Results of the Evaluation

We evaluated performance of FiNER and SeCo’s ARPA using the *conlleva1*⁸ script used in Conference on Computational Natural Language Learning (CONLL). Evaluation is based on “exact-match evaluation” [1], [16]. In this type of evaluation NER system is evaluated based on the micro-averaged F-measure (MAF) where precision is the percentage of correct named entities found by the NER software; recall is the percentage of correct named entities present in the tagged evaluation corpus that are found by the NER system. A named entity is considered correct only if it is an exact

⁸ <http://www.cnts.ua.ac.be/conll2002/ner/bin/conlleva1.txt>, author ErikTjong Kim Sang, version 2004-01-26

match of the corresponding entity in the tagged evaluation corpus: “a result is considered correct only if the boundaries and classification are exactly as annotated” [17]. Thus the evaluation criteria are strict, especially for multipart entities.

Detailed results of the evaluation of FiNER are shown in Table 1. Entities <ent/> consist of one word token, <ent> are part of a multiword entity and </ent> are last parts of multiword entities.

Label	P	R	F-score	Number of tags found	Number of tags in the evaluation data
<EnamexLocGpl/>	6.96	9.41	8.00	115	85
<EnamexLocPpl/>	89.50	8.46	15.46	181	1920
<EnamexLocStr/>	23.33	50.00	31.82	30	14
<EnamexLocStr>	100.00	13.83	24.30	13	94
</EnamexLocStr>	100.00	18.31	30.95	13	71
<EnamexOrgCrp/>	2.39	6.62	3.52	376	155
<EnamexOrgCrp>	44.74	25.99	32.88	190	338
</EnamexOrgCrp>	40.74	31.95	35.81	189	250
<EnamexOrgEdu>	48.28	40.00	43.75	29	35
</EnamexOrgEdu>	55.17	64.00	59.26	29	25
<EnamexPrsHum/>	16.38	52.93	25.02	1819	564
<EnamexPrsHum>	87.44	26.67	40.88	438	1436
</EnamexPrsHum>	82.88	31.62	45.78	438	1150
<TimexTmeDat/>	5.45	14.75	7.96	495	183
<TimexTmeDat>	68.54	2.14	4.14	89	2857
</TimexTmeDat>	20.22	2.00	3.65	89	898

Table 1. Evaluation results of FiNER with strict CONLL evaluation criteria. Data with zero P/R is not included in the table. These include categories <EnamexLocGpl/>, </EnamexLocGpl/>, <EnamexLocPpl/>, </EnamexLocPpl/>, <EnamexLocXxx/>, <EnamexLocXxx/>, </EnamexLocXxx/>, and <EnamexOrgEdu/>. Most of these have very few entities in the data, only <EnamexLocXxx/> is frequent with over 1200 occurrences

Results of the evaluation show that named entities are not recognized very well, which is not surprising, as the quality of the text data is quite low. Especially recognition of multipart entities is very low. Some part of the entities may be recognized, but rest is not. Out of multiword entities corporations and educational organizations are recognized best. Names of persons are the most frequent category. Recall of one part person names is best, but its precision is low. Multipart person names have a more balanced recall and precision, even if their overall recognition is not high.

In a looser evaluation the categories were treated so that any correct marking of an entity regardless its boundaries was considered a hit. Four different location categories were joined to two: general location <EnameLocXxx> and that of street names. End result was six different categories instead of eight. Table 2 shows evaluation results with loose evaluation. Recall and precision of the most frequent categories of person and location was now clearly higher, but still not very good.

Label	P	R	F-score	Number of tags
<EnamePrsHum>	63.30	53.69	58.10	2681
<EnameLocXxx>	69.05	49.21	57.47	1541
<EnameLocStr>	83.64	25.56	39.15	55
<EnameMemOrgEdu>	51.72	47.62	49.59	58
<EnameMemOrgCrp>	30.27	32.02	31.12	750
<TimeXTmeDat>	73.85	12.62	21.56	673

Table 2. Evaluation results of FiNER with loose criteria and six categories

Our third evaluation was performed for a limited tag set with tools of the SeCo's ARPA. First only places were identified so that one location, *EnameLocPpl*, was recognized. For this task, ARPA was first configured for the task of identifying place names in the data. As a first iteration, only the Finnish Place Name Registry⁹ was used. After examining raw results from the test run, three issues were identified for further improvement. First, PNR contains only modern Finnish place names. To improve recall, three registries containing historical place names were added: 1) the Finnish spatiotemporal ontology SAPO [18] containing names of historic municipalities, 2) a repository of old Finnish maps and associated places from the 19th and early 20th Century, and 3) a name registry of places inside historic Karelia, which does not appear in PNR due to being ceded by Finland to the Soviet Union at the end of the Second World War [19]. To account for international place names, the names were also queried against the Geonames database¹⁰ as well as Wikidata¹¹. The contributions of each of these resources to the number of places identified in the final runs are shown in Table 3. Note that a single place name can be, and often was found in multiple of these sources.

Source	Matches	Fuzzy matches
Karelian places	461	951
Old maps	685	789
Geonames	1036	1265
SAPO	1467	1610
Wikidata	1877	2186
PNR	2232	2978

⁹ <http://www.ldf.fi/dataset/pnr/>

¹⁰ <http://geonames.org/>

¹¹ <http://wikidata.org/>

Table 3. Number of distinct place names identified using each source

Table 4 describes the results of location recognition with ARPA. Without one exception (*New York*), only one word entities were discovered by the software

Label	P	R	F-score	Number of tags
<EnamexLocPpl/>	39.02	53.24	45.03	2673
</EnamexLocPpl>	100.00	5.26	10.00	1
<EnamexLocPpl>	100.00	4.76	9.09	1

Table 4. Basic evaluation results for ARPA

A second improvement to the ARPA process arose from the observation that while recall in the first test run was high, precision was low. Analysis revealed this to be due to many names being both person names as well as places. Thus, a filtering step was added, that removed 1) hits identified as person names by the morphological analyzer and 2) hits that matched regular expressions catching common person name patterns found in the data (I. Lastname and FirstName LastName). However, sometimes this was too aggressive, ending up for example in filtering out also big cities like Tampere and Helsinki. Thus, in the final configuration, this filtering was made conditional on the size of the identified place, as stated in the structured data sources matched against.

Finally, as the amount of OCR errors in the target dataset was identified to be a major hurdle in accurate recognition, experiments were made with sacrificing precision in favor of recall through enabling various levels of Levenshtein distance matching against the place name registries. In this test, the fuzzy matching was done in the query phase after lexical processing. This was easy to do, but doing the fuzzy matching during lexical processing would probably be more optimal, as currently lemma guessing (which is needed because OCR errors are out of the lemmatizer’s vocabulary) is extremely sensitive to OCR errors particularly in the suffix parts of words.

After the place recognition pipeline was finalized, a further test was done to test if the ARPA pipeline could be used for also person name recognition. Here, as a lexicon of names, the Virtual International Authority File was used, as it contains 33 million names for 20 million people. In the first run, the query simply matched all uppercase words against both first and last names in this database, while allowing for any number of initials to also precede such names matched. This way, the found names can’t actually be always any more linked to strong identifiers, but for a pure NER task, recall is improved.

Table 5 shows results of this evaluation without fuzzy matching of names and Table 6 with fuzzy matching.

Label	P	R	F-score	Number of tags
<EnamexLocPpl/>	58.90	55.59	57.20	1849
</EnamexLocPpl>	1.49	10.53	2.61	134
<EnamexLocPpl>	1.63	14.29	2.93	184
<EnamexPrsHum/>	30.42	27.03	28.63	2242
</EnamexPersHum>	83.08	47.39	60.35	656
<EnamexPersHum>	85.23	43.80	57.87	738

Table 5. Evaluation results for ARPA: no fuzzy matching

Label	P	R	F-score	Number of tags
<EnamexLocPpl/>	47.38	61.82	53.64	2556
</EnamexLocPpl>	1.63	15.79	2.96	184
<EnamexLocPpl>	1.55	14.29	2.80	193
<EnamexPrsHum/>	9.86	66.79	17.18	3815
</EnamexPersHum>	63.07	62.97	63.01	1148
<EnamexPersHum>	62.25	61.77	62.01	1425

Table 6. Evaluation results for ARPA: fuzzy matching

Recall of recognition increases markedly in fuzzy matching, but precision deteriorates. More multipart location names are also recognized with fuzzy matching.

Loose evaluation without fuzzy matching gave 44.02 % precision, 64.58 % recall and 52.35 F-score for locations with 2933 found tags. For persons it gave precision of 63.61%, recall of 45.27% and F-score of 52.90 with 3636 found tags.

Loose evaluation with fuzzy matching gave 44.02 % precision, 64.58 % recall and 52.35 F-score for locations. Number of found tags was 2933. For persons it gave precision of 34.49, recall of 78.09 and F-score of 51.57 with 6388 found tags.

3 Discussion

We have shown in this paper first evaluation results of NER for historical Finnish newspaper material from the 19th and early 20th century with two different tools, FiNER and SeCo's ARPA. Word level correctness of the digitized newspaper archive is approximately 70–75 %; the evaluation corpus had a word level correctness of about 73 %. Regarding this and the fact that FiNER and ARPA were developed for modern Finnish, the newspaper material makes a very difficult test for named entity recognition. It is obvious that the main obstacle of high class NER in this material is bad quality of the text. Also historical spelling variation has some effect, but it should not be that high.

Evaluation results in this phase were not very good, best basic F-scores were ranging from 30 to 60 in the basic evaluation, and slightly better in a looser evaluation. We have ongoing trials for improving word quality of our material, which may yield also better NER results. We made some unofficial tests with three versions of a 500,000 word text material that is different from our NER evaluation material but

derives from the 19th century newspapers as well. One version was manually corrected OCR, another old OCRred version and third a new OCRred version. Besides character level errors also word order errors have been corrected in the two new versions. For these texts we did not have a ground truth tagged version, so we could only count marking of NER tags. With FiNER total number of tags increased from 23,918 to 26,674 (+11.5 % units) in the manually corrected text version. Number of tags increased to 26,424 tags (+10.5 % units) in the new OCRred text version. Most notable increase in the number of tags was in categories *EnamexLocStr* and *EnamexOrgEdu*. With ARPA results were even slightly better. ARPA recognized 10 853 places in the old OCR, 11,847 in the new OCR (+ 9.2 % units) and 13,080 (+20.5 % units) in the ground truth version of the text. There is about a 10–20 % unit overall increase in the number of NER tags in both of the new better quality text versions in comparison to the old OCRred text with both taggers.

NER experiments with OCRred data in other languages show usually some improvement of NER when the quality of the OCRred data has been improved from very poor to somehow better [15, 16]. Results of Alex and Burns [18] imply that with lower level OCR quality (below 70 % correctness) name recognition is harmed clearly. Packer et al. [15] report partial correlation of Word Error Rate of the text and achieved NER result; their experiments imply that word order errors are more significant than character errors. On the other hand, results of Rodriguez et al. [17] show, that manual correction of OCRred material that has 88–92 % word accuracy does not increase performance of four different NER tools significantly. As the word accuracy of our material is low, it would be expectable, that somehow better recognition results would be achieved, if the word accuracy was round 80–90 % instead of 70–75 %. Our informal test with different quality texts suggests this, too. Our material has also quite a lot of word order errors which may affect results.

Another option for better recognition results is that we can use more historical language sensitive NER software. Such may become available, if the historically more sensitive version of morphological recognizer Omorfi can be merged with FiNER. A third possibility is to train a statistical name tagger described by Silfverberg [11] with labeled historical newspaper material.

Other causes for poor performance are probably due to 19th century Finnish spelling variation and perhaps also due to different writing conventions of the era. It is possible, for example, that the genre of 19th century newspaper writing differs from modern newspaper writing in some crucial aspects. Considering that both FiNER and ARPA are made for modern Finnish, our evaluation data is heavily out of their main scope [19], even if ARPA uses historical Finnish aware Omorfi.

In our case extraction of names is primarily a tool for improving access to the Digi collection. After getting the recognition rate of the NER tool to acceptable level, we need to decide, how we are going to use extracted names in Digi. Some exemplary suggestions are provided by archive of La Stampa and Trove Names [9]. La Stampa style usage of names provides informational filters after a basic search has been conducted. You can further look for persons, locations and organizations mentioned in the article results. This kind of approach enables browsing access to the collection and possibly also entity linking [20, 21, 22]. Trove Names’s name search takes the oppo-

site approach: you first search for names and then you get articles where the names occur. We believe that the La Stampa style of usage of names in the GUI of the newspaper collection is more informative and useful for users, as the Trove style can be already obtained with the normal search function in the GUI of the newspaper collection. If we consider possible uses of NER in Digi, FiNER does so far only basic identifying and classification of names. ARPA is basically not a NER software, but a semantic entity linking system, and thus of broader use. Our main emphasis with NER will be on the use of the names with the newspaper collection as a means to improve browsing and general informational usability of the collection. A good enough coverage of the names with NER needs to be achieved also for this use, of course. A good balance of P/R should be found for this purpose [15], but also other capabilities of the software need to be considered. These remain to be seen later, if we are able to connect some type of functional NER to our historical newspaper collection.

Acknowledgements

First author is funded by the EU Commission through its European Regional Development Fund, and the program Leverage from the EU 2014–2020.

References

1. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes* 30(1):3–26 (2007)
2. Kettunen, K., Pääkkönen, T.: Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. Accepted for LREC 2016. <http://lrec2016.lrec-conf.org/en/conference-programme/accepted-papers/> (2016).
3. Bremer-Laamanen, M-L.: In the Spotlight for Crowdsourcing. *Scandinavian Librarian Quarterly*, 1, 18–21 (2014)
4. Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., Kervinen, J.: Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. In: Proceedings of IFLA 2014, Lyon (2014) http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf
5. Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M. and Kettunen, K.: Information Retrieval from Historical Newspaper Collections in Highly Inflectional Languages: A Query Expansion Approach. *Journal of the Association for Information Science and Technology* doi: <http://onlinelibrary.wiley.com/doi/10.1002/asi.23379/epdf> (2015)
6. Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., Borin, L.: HFST-SweNER – a New NER Resource for Swedish. In: Proceedings of LREC 2014, http://www.lrec-conf.org/proceedings/lrec2014/pdf/391_Paper.pdf (2014)
7. Crane, G., Jones, A.: The Challenge of Virginia Banks: An Evaluation of Named Entity Analysis in a 19th-Century Newspaper Collection. In *Proceedings of JCDL '06*, June 11–15, 2006, Chapel Hill, North Carolina, USA. <http://repository01.lib.tufts.edu:8080/fedora/get/tufts:PB.001.001.00007/Archival.pdf> (2006)
8. Neudecker, C., Wilms, L., Faber, W. J., van Veen, T.: Large-scale Refinement of Digital Historic Newspapers with Named Entity Recognition.

- http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-neudecker_faber_wilms-en.pdf (2014)
9. Mac Kim, S., Cassidy, S.: Finding Names in Trove: Named Entity Recognition for Australian. In: Proceedings of Australasian Language Technology Association Workshop, pp. 57–65, <https://aclweb.org/anthology/U/U15/U15-1007.pdf> (2015)
 10. Mäkelä, E.: Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text. In Presutti, V. et al. (eds.), The Semantic Web: ESWC 2014 Satellite Events. Lecture Notes in Computer Science, vol. 8798, pp. 424–428 (2014)
 11. Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T.A., Silfverberg, M.: HFST—a System for Creating NLP Tools. In Mahlow, C., Piotrowski, M. (eds.) Systems and Frameworks for Computational Morphology. Third International Workshop, SFCM 2013, Berlin, Germany, September 6, 2013 Proceedings, pp. 53–71 (2013).
 12. Silfverberg, M.: Reverse Engineering a Rule-Based Finnish Named Entity Recognizer. https://kitwiki.csc.fi/twiki/pub/FinCLARIN/KielipankkiEventNERWorkshop2015/Silfverberg_presentation.pdf (2015)
 13. Packer, T., Lutes, J., Stewart, A., Embley, D., Ringger, E., Seppi, K., Jensen, L. S.: Extracting Person Names from Diverse and Noisy OCR Text. In: Proceedings of the fourth workshop on Analytics for noisy unstructured text data. Toronto, ON, Canada: ACM. (2010)
 14. Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J.M.: Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces* 35, 482–489 (2013)
 15. Rodrigues, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of Named Entity Recognition Tools for raw OCR text. In: Proceedings of KONVENS 2012 (LThist 2012 wordshop), Vienna September 21, pp. 410–414 (2012)
 16. Alex, B., Burns, J.: Estimating and Rating the Quality of Optically Character Recognised Text. In: DATECH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, pp. 97–102. <http://dl.acm.org/citation.cfm?id=2595214> (2014)
 17. Poibeau, T., Kosseim, L.: Proper Name Extraction from Non-Journalistic Texts. *Language and Computers*, 37, pp. 144-157 (2001)
 18. Hyvönen, E., Tuominen, J., Kauppinen T., Väättäinen, J: Representing and Utilizing Changing Historical Places as an Ontology Time Series. In Ashish, N. and Sheth, V. (eds.) Geospatial Semantics and Semantic Web: Foundations, Algorithms, and Applications, Springer-Verlag, (2011)
 19. Ikkala, E., Tuominen, J., Hyvönen, E.: Contextualizing Historical Places in a Gazetteer by Using Historical Maps and Linked Data. In: Proceedings of Digital Humanities 2016, short papers, Kraków, Poland (2016)
 20. Bates, M.: What is Browsing – really? A Model Drawing from Behavioural Science Research. *Information Research* 12. <http://www.informationr.net/ir/12-4/paper330.html> (2007)
 21. Toms, E.G.: Understanding and Facilitating the Browsing of Electronic Text. *International Journal of Human-Computer Studies*, 52(3), 423–452 (2000)
 22. McNamee, P., Mayfield, J.C., Piatko, C.D.: Processing Named Entities in Text. *Johns Hopkins APL Technical Digest*, 30(1), pp. 31–40. (2011)

Experience – the neglected success factor in enterprises?

Edith Maier¹, Werner Bruns², Sebastian Eschenbach³, and Ulrich Reimer¹

¹ Fachhochschule St. Gallen, Switzerland, {edith.maier,ulrich.reimer}@fhsg.ch

² Rheinische Fachhochschule Köln, Deutschland, werner.bruns@rfh-koeln.de

³ Fachhochschule Burgenland, Österreich, sebastian.eschenbach@fh-burgenland.at

Abstract. The effective use of experience as a valuable resource can give companies a competitive edge in a world characterised by an ageing workforce and globalisation. An online survey was conducted in Austria, Germany and Switzerland to find out managers' attitudes towards experience and if and how they capture, use and disseminate it. The results show that the majority consider experience an important asset, but do not actually support it in any systematic way. Company size and position rather than age or gender play a role when it comes to preferences, attitudes or practices. The survey shows great discrepancies between methods considered useful vs. those in regular use. Besides, there is a preference for classical people-oriented methods rather than modern IT-supported methods. Integrating experience management into project and process management practice may help overcome current barriers and reservations.

1 Introduction

More than ten years ago, studies by KPMG [2] and Fraunhofer [7] showed the high relevance of experience management for industry. At the time, the experience base of an enterprise was given top priority when IT support systems for knowledge management were installed and implemented. The rationale behind implementing experience management (EM) was to meet increasing demands in industry for process improvement approaches (e.g. Six Sigma etc.) to achieve higher and more repeatable product quality. This was to be enabled by better understanding, standardising and optimising processes and decisions by means of automation of production lines and business processes at a more fine-grained level [11].

In the meantime the rationale has shifted away from process improvement towards the demands and challenges posed by an ageing society and increasing globalisation. In the next few years many experts from the so-called 'baby boomer' generation are about to retire whose expertise and experience companies want and need to preserve. Besides, we are increasingly faced with incomplete knowledge in a world that is characterised by great uncertainties and imponderables as a result of disruptive innovations brought about mainly by digitalisation.

The experience we have accumulated over time may help us deal with these challenges, crises and conflicts. Companies and their managers are therefore called upon to make the best use of the experience and know-how of their employees. Instead, there appears to be a general trend as observed by [6] that organisations are failing to learn from their past experiences despite being surrounded by lessons learned models and guides on how to apply them.

To find out if and how companies actually document, exchange, manage and maintain this valuable resource, three universities of applied sciences from Austria (FH Burgenland), Switzerland (FHS St. Gallen) and Germany (Rheinische Fachhochschule Köln) carried out a survey in the autumn of 2015 together with German and Swiss national management associations (*Die Führungskräfte* with 8000 members and the *Schweizer Kader Organisation* with a membership of 8200, respectively), as well as the Austrian magazine *Die Presse*.

It is the first trans-border online survey of this kind and has been initiated by the newly formed European Institute for Experience-Based Knowledge (METIS), a think tank which brings together partners from research, industry, civil society and governmental institutions. The aim of METIS is to foster the dialogue between these players and contribute to and further develop successful methods for the transfer of experience-based or practical knowledge.

In the following sections we will briefly define the various concepts in relation to experience and EM, describe the methods we used for collecting and analysing the data and discuss the results of the survey. We shall pay particular attention to respondents' attitudes to methods and instruments that can be used for EM. We then explore the implications of the survey for both research and practice and ask how it could be made easier to exploit experience as a valuable asset.

2 Definitions, concepts and models

Webster's Dictionary defines experience as knowledge or practical wisdom gained from what one has observed, encountered, or undergone. The term experience or experience-based knowledge is closely related to terms such as good or best practice, lessons learned, tacit knowledge, knowledge-in-use etc. As early as 1958, Polanyi explored the distinction between tacit and explicit knowledge [15] and thus laid the foundation for Nonaka and Takeuchi [12] who made major contributions to knowledge management (KM) theory. They state that whereas explicit or codified knowledge is objective, easily communicated and transferred without in-depth experience, tacit knowledge is subjective, context-specific, personal, and difficult to communicate. It consists of cognitive elements such as cultural beliefs and viewpoints as well as technical elements, i.e. the existing know-how and skills.

The close link with KM is also made by [11] where EM is defined as a special form of KM and an Experience Management System (EMS) as a socio-technical system established for managing, reusing and recording experience or lessons learned. Research in EM therefore deals with methods and technologies suitable for collecting them from various sources, documenting, sharing, adapting and

distributing experience. It also includes the organisational and social measures required to assure that these are integrated into business processes (see also [4]).

According to [11] EM software should support a set of operations related to the reuse, adaptation and recording of experiences. But to make sure that EM activities are executed, they stress that its online components have to be directly linked to the business processes. IT solutions that can support and enable EM activities include incident management software, learning management systems, expertise location systems, enterprise content management systems, search technologies, e-discovery technology, and software for social exchange (e.g. instant messaging, blogging and micro-blogging), social networking and collaboration.

Different fields in artificial intelligence have also contributed to EM. Case-based reasoning, in particular, has played a role in the development, validation and maintenance of experience bases that may include case studies or lessons learned from projects. For dissemination and transfer of experience or lessons learned, various technology approaches are available such as formal reasoning as well as ontologies that can support the retrieval and adaptation of lessons learned.

Despite the continuous improvement of IT solutions, they have also been blamed for failure in the exchange and dissemination of knowledge and experience (e.g. [18]). As a result, there has been a move away from a reliance on IT to an approach that aligns and balances people, process and technology (see e.g. [3]).

3 Methodological considerations

For the survey, a questionnaire was developed by the academic partners and aimed at obtaining an overview of attitudes towards practices, instruments and methods with regard to the role of experience and its management and transfer in the corporate German-speaking world. Since the survey targeted senior and middle managers, the role of leadership in EM was another important issue raised in the questionnaire. Overall, we received 829 filled-in questionnaires out of which 359 from Germany, 147 from Switzerland and 51 from Austria.

The questionnaires were collected and analysed by the computing centre of the RHFH Cologne and interpreted by experts at the three universities of applied sciences. The statistics software SPSS was employed for univariate and bivariate statistical analysis to: (a) describe the attitudes of the total sample towards experience using a seven-part Likert scale and (b) to test for significant differences between subsamples, e.g. respondents from larger versus medium-size companies, using chi-squared and Mann-Whitney U tests which both allow the analysis of ordinal scaled non-normal data. At a significance level of less than or equal to 0.05 the null hypothesis, i.e. that the sub-samples (e.g. middle vs. senior managers) show the same distribution for a concrete variable, was rejected.

For comparing the three country subsamples we performed a Kruskal-Wallis H test in SPSS. It turned out that respondents from the three countries constitute three significantly different subsamples with regard to socio-economic at-

tributes (e.g. age, gender, education, position). This is largely due to differences in the membership of the German and Swiss associations and the readership of Austria's *Die Presse*. Although all respondents are managers (Führungskräfte), the socio-economic differences between the country subsamples make any meaningful comparison of national differences difficult.

Finally, we would like to point out certain constraints of our survey. For example, we had to adopt the management associations' preferred categories for company size rather than use the EU definitions. Random sampling was not possible because we do not know the total number of managers in Germany, Austria or Switzerland. Therefore we had to make do with a convenience sample and cannot make any representative statements about the total management population.

4 Results of survey

We have received just over 600 usable, i.e. completely filled-in replies, two thirds of which come from Germany (65%), 26% from Switzerland and 9% from Austria. Women account for almost a fifth of replies (18%). More than half of respondents (54%) are managers in large companies (>500 employees), 42% work in medium-size companies (10–500 employees). More than three quarters (77%) have graduated, about half have a technical, the other half a business or legal background.

The majority (85%) consider experience as an important resource for the success and productivity of their company, especially for making organisational routines and processes more efficient. However, only about a quarter of respondents claim that the exchange of experience enjoys the full support in their companies. Large companies, however, are more systematic and committed when it comes to promoting the transfer of experience.

Only about one fifth (21%) of managers show themselves satisfied with the exchange of experience across different levels of hierarchy or company divisions. The barriers are even higher when it comes to experience exchange beyond company boundaries, e.g. with customers or suppliers. This result shows that there is still a long way to go as far as the open exchange of experience and knowledge across organisational boundaries is concerned.

With regard to attitudes and preferred methods, the survey points to differences between senior and middle managers, whereas age appears to play a very minor role. Younger managers, however, are more likely to be aware of the "dark" side of experience, e.g. the danger of becoming professionally blinkered because one relies too much on established practices rather than open up to new possibilities. The professional background, e.g. whether someone has been to university or received vocational training, does not appear to have any influence on one's attitude to experience.

Generally, the motivation behind the implementation of EM is to learn from past experiences so as to avoid repeating mistakes. Figure 1 provides an overview of where and for which purposes experience-based knowledge is used.

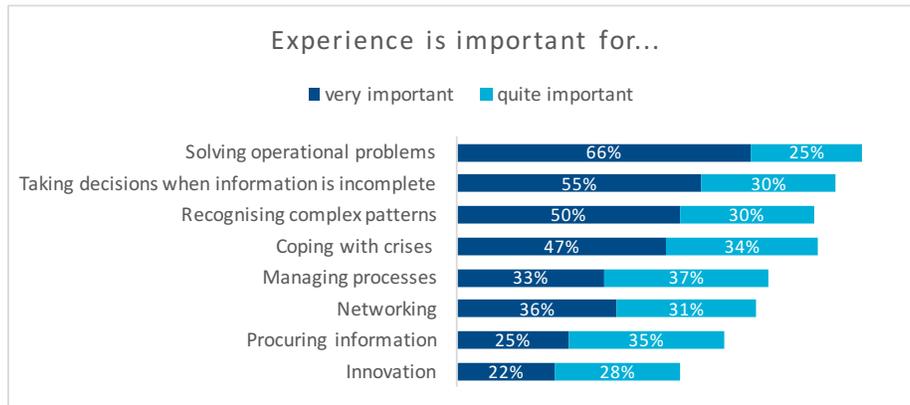


Fig. 1. Applications of experience-based knowledge

As shown in Figure 1, experience is considered as ‘very important’ and ‘quite important’ for solving operational problems, taking decisions when information is incomplete, for recognising complex patterns as well as coping with crises by a high percentage of respondents. It is deemed less important for process management, networking and information procurement. And only 50% believe that experience may foster innovation.

Figure 2 illustrates people’s attitudes towards different methods for the exchange and management of experience. What is striking is the considerable discrepancies between which methods are considered useful and their actual implementation. For example, the majority of respondents see the potential usefulness of both Succession planning and Induction programmes for new employees but few actually use them regularly in their organisations.

It also shows that many respondents have considerable reservations with regard to KM techniques such as world cafés, lessons learned workshops or storytelling (which can be subsumed under the term ‘Moderated experience exchange’), networking approaches such as communities of practice as well as social media platforms or intranets. They see them as ineffective and/or do not use them on a regular basis. Even younger managers are sceptical with regard to such tools and tend to prefer the classical management and communication tools such as informal talks and meetings. What is interesting is that women on the whole appear to be more open with regard to the possibilities offered by online platforms or social networks.

Whilst people-oriented methods such as induction programmes for new employees or mentoring are considered useful, in the ‘real world’ it is the more formal methods such as written reports, meetings or professional or further training courses that tend to dominate.

These results have been confirmed by several informal interviews conducted by W. Bruns, one of the founders of METIS, as well as in a series of in-depth interviews currently conducted in a follow-up study based on the METIS questionnaire. Preliminary findings from interviews with CEOs from companies of

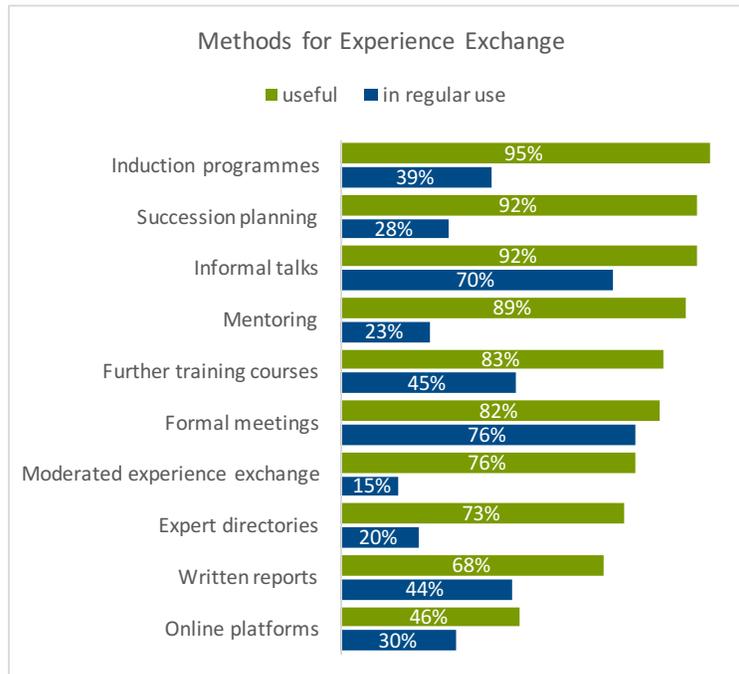


Fig. 2. Methods used for the exchange of experience

varying size and different industries show a wide-spread disenchantment with modern KM methods on the part of those (esp. large companies) who have actually experimented and/or implemented them, as well as wide-spread scepticism and reservations on the part of those who have not.

We can conclude that whilst experience is held in high esteem, little is done to actually manage and cultivate it. For example, rarely do companies offer incentives or rewards for EM. When asked for the reasons in the interviews, lack of time and resources are cited most frequently. It appears that the pressures from daily business and competitors do not allow the management to dedicate more resources to EM, even though they may consider it valuable.

5 Implications for research and practice

The findings from the survey are not particularly encouraging for those engaged in promoting modern IT-supported methods of KM, in general, and EM, in particular. At the same time, they may be a call for action because it is clear that experience and EM are attributed great importance but that there is a lack of know-how about how best to exploit this valuable resource and thus, a need for support. This in line with the findings of [6] that whilst processes for identifying lessons do exist, organisations fail to disseminate and apply them.

This raises the issue how such support can be delivered more effectively and efficiently? Do we have to change the “packaging” of our methods, e.g. avoid terminology (CoPs, world cafés etc.) that may (still) sound outlandish to the down-to-earth manager of a small enterprise? Or do we have to shift the focus of EM towards management communication or project management, e.g. by focussing on how companies can distribute successful project know-how across an organisation to ensure that lessons are learned and mistakes of the past are not repeated?

However, when managers try to turn inherently tacit knowledge into explicit knowledge they often encounter pitfalls. Xerox is an example that is often quoted in the literature (e.g. [8]). They attempted to embed the know-how of its service and repair technicians into an expert system that was installed in the copiers and expected that technicians responding to a call could be guided by the system and complete repairs from a distance. That’s not what happened. Rather the copier designers discovered that technicians learned from one another by sharing stories about how they had fixed the machines. The expert system could not replicate the nuance and detail that were exchanged in face-to-face conversations.

Nowadays, many such conversations happen in discussion fora or blogs on the Internet. Software engineers, in particular, consult them when they encounter a tricky problem. Many technology firms also offer Q&A sections where users can find answers to problems. Usually such platforms are not as well structured as expert systems, which is why text mining and intelligent search algorithms may be used to help people find what they are looking for and serve to “pre-codify” relevant knowledge.

It is generally recognised that social media as well as the dramatic advance and widespread use of mobile devices, social software and online social networking are having a positive impact on KM [13]. These trends have rekindled the debate about how technology may contribute to an effective sharing of knowledge and experience across units and organisations [14]. Since the open innovation concept actually emphasises the idea of alliances and cooperation between partners across organisational boundaries to create new products and services, it is surprising that the respondents in the survey do not appear to recognise the role that EM might play in innovation.

In a recent article in the Harvard Business Review [5], the authors discuss a key obstacle to innovation, i.e. the absence of any systematic review of lessons a company might learn from mistakes or failed projects. They suggest to rigorously extract value from failure by means of a three-step process:

1. Learn from every failure, e.g. the insights one has gained (e.g. about customers or markets, one’s team, personal growth) as well as the liabilities (e.g. costs in time and money, reputation)
2. Share the lessons across the organisation (e.g. by means of regular reviews for sharing lessons incl. informal approaches such as capturing critical lessons with stories)
3. Review one’s pattern of failure from a bird’s-eye view (e.g. is our organisation learning from unsuccessful endeavours?)

The aim is to nudge people toward greater openness to failure, which will be less painful according to the authors when one manages to extract the maximum return from it. According to the authors this can only be achieved when we learn from mistakes, share those lessons and periodically check that processes such as lessons learned workshops or debriefings help one's organisation move more efficiently in the right direction.

Finally, both the survey and the preliminary results from the follow-up study show that people will not engage in EM if it implies additional effort. Therefore, EM activities have to be integrated into workflow and process management approaches to provide the context in which experience is reused on the one hand, and to provide best practices, i.e. proven procedures for performing certain tasks, on the other hand. This demand is not really new and has been voiced by other researchers who wrote about how best to support knowledge-intensive work (see e.g. [16, 1], but these kinds of approaches have received much less attention in the last ten years.

Similarly, EM activities – especially those related to lessons learned and best practices – should be integrated with project management. As has been pointed out in [10], there is actually a gap in project management practice and suggested that there is a need for more research in understanding the role KM plays in project management methodologies.

In this respect it may be worth mentioning the so-called “Syllk” model, which stands for Systemic Lessons Learned Knowledge model. According to its proponents [6] it could assist in identifying the KM barriers that need to be overcome for an effective transfer of lessons learned. Others such as [9] have demonstrated how the Syllk model can support knowledge sharing and integration between an organisation and its suppliers, customers and partners. As is the case with experience and its transfer, the human factor plays a major role in the studies on as well as applications of the Syllk model because it recognises that for organisations to learn, people and systems (processes and technology) have to be working together closely [17].

6 Conclusions

As we have seen, experience and its management may well be one of the most neglected success factors in companies in the German-speaking corporate world. Although the majority of managers consider experience an important asset, few actually support it in any systematic way. This finding has been corroborated by a series of interviews conducted as a follow-up to the survey.

Company size and position rather than age or gender play a role when it comes to preferences, attitudes or practices with regard to experience and its management. The preference for classical and people-oriented methods rather than more modern IT-supported methods, however, appears to be shared by all. To overcome the current reservations with regard to potentially effective methods for experience exchange, we suggest looking further into how to integrate experience and its management into project and process management practice

as an automatic part that does not require any additional effort. Only then will it be possible to exploit the full potential of EM.

References

1. Abecker, A., Bernardi, A., Hinkelmann, K.: Context-aware, proactive delivery of task-specific knowledge: The KnowMore Project. *International Journal on Information Systems Frontiers* 2(3/4), 139–162 (2000)
2. Abele, J., Pfaff, M.: Bedeutung und Entwicklung des multimediebasierten Wissensmanagements in der mittelständischen Wirtschaft. Studie im Auftrag des Bundesministeriums für Wirtschaft und Technologie. KPMG (Hrsg.) (2001)
3. Barnes, S.: *Aligning people, process and technology in knowledge management*. Ark Group (2011)
4. Bergmann, R.: *Experience management: Foundations, development methodology, and internet-based applications*. Springer-Verlag (2002)
5. Birkinshaw, J., Haas, M.: Increase your return on failure. *Harvard Business Review* 94(5), 88–93 (2016)
6. Duffield, S., Whitty, S.J.: Developing a systemic lessons learned knowledge model for organisational learning through projects. *International Journal of Project Management* 33(2), 311–324 (2015)
7. Fraunhofer Wissensmanagement-Community: *Wissen und Information 2005*. Fraunhofer IRB-Verlag (2005)
8. Hansen, M.T., Nohria, N., Tierney, T.: What’s your strategy for managing knowledge? *Harvard Business Review* 77(2), 106–116 (1999)
9. Leal-Rodríguez, A.L., Roldán, J.L., Ariza-Montes, J.A., Leal-Millán, A.: From potential absorptive capacity to innovation outcomes in project teams: The conditional mediating role of the realized absorptive capacity in a relational learning context. *International Journal of Project Management* 32(6), 894–907 (2014)
10. Lindner, F., Wald, A.: Success factors of knowledge management in temporary organizations. *International Journal of Project Management* 29(7), 877–888 (2011)
11. Nick, M., Althoff, K.D., Bergmann, R.: Experience management. *Künstliche Intelligenz* 21(2), 50–52 (2007)
12. Nonaka, I., Takeuchi, H.: *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press (1995)
13. O’Dell, C., Hubert, C.: *The new edge in knowledge: How knowledge management is changing the way we do business*. John Wiley & Sons (2011)
14. Orlikowski, W.J.: Sociomaterial practices: Exploring technology at work. *Organization Studies* 28(9), 1435–1448 (2007)
15. Polanyi, M.: The logic of tacit inference. *Philosophy* 41(155), 1–18 (1966)
16. Reimer, U., Novotny, B., Staudt, M.: Micro-modeling of business processes for just-in-time knowledge delivery. In: Roy, R. (ed.) *Industrial Knowledge Management*, pp. 283–297. Springer (2001)
17. Virolainen, T.: *Learning from projects: a qualitative metasummary*. Master thesis, Lappeenranta University of Technology, School of Industrial Engineering and Management (2014)
18. Williams, T.: *Post-project reviews to gain effective lessons learned*. Project Management Institute (2007)

Linked Data City - Visualization of Linked Enterprise Data

Joachim Baumeister^{1,2}, Sebastian Furth¹, Lea Roth¹ and Volker Belli¹

¹ denkbares GmbH, Friedrich-Bergius-Ring 15, 97076 Würzburg

² University of Würzburg, Am Hubland, 97074 Würzburg

Abstract. A generic technique for the visualization of hierarchical structures is introduced. The actual visualization is not only defined by the underlying data but also the application of domain-driven metrics. The paper shows two use cases for the analysis of linked enterprise data in the domain of technical service information systems.

1 Introduction

In the age of digitalization and automation of industries, many companies are consolidating their business information systems and product meta-data, such as ERP, CRM, file directories, and extranet data. In many cases, not all elements of these information resources are accessible to all relevant users. The intransparent access hinders effective work processes and often threatens business success. Therefore, the primary goal of many ICT projects is the linkage of the existing information silos into an integrated information infrastructure. Here, semantic technologies, and especially linked data models, are a successful enabler for building such *knowledge warehouses* mediating the information silos. Linked Enterprise Data [1] transfers the ideas and technologies of linked data [2] into the much more restricted world of business and enterprises. Standard semantic languages, such as RDF and SPARQL, are used to represent the core entities of the enterprise. Useful de-facto standard vocabularies for the enterprise usage already exist, see for instance SKOS [14] and GoodRelations [8]. Within a semantic infrastructure, all information resources are uniformly and semantically accessible by the user and novel services. In consequence, a number of advanced applications with business added value become possible [13]:

- Semantic enterprise search
- Semantic B2B portal with standardized data exchange
- Semantic assistants
- Automated data quality and curation processes

During the migration from the existing information structure to linked enterprise data, existing information sources need to be linked with semantic concepts. Here, a toolbox of core technologies ranging from Natural Language Processing/Information Extraction to Information Retrieval methods [4] is employed.

In Figure 1 the semantification process of enterprise data is depicted [5]. Each step of the process includes a detailed analysis:

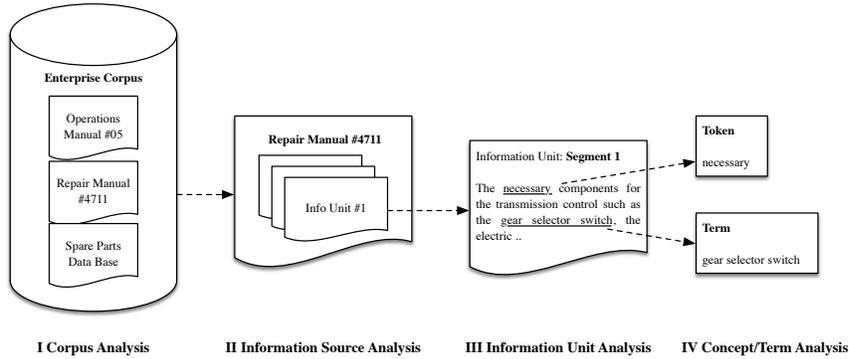


Fig. 1. Simplified process for the semantification of enterprise information sources.

I Corpus Analysis The existing data is collected and described. All relevant information systems are analyzed with respect to the included information sources.

II Information Source Analysis The included information sources are analyzed in more detail, for instance, with respect to the size and number of elements in the corpus.

III Information Unit Analysis Information sources are composed of information units, i.e., segments of a source that are not dividable anymore. Number and distribution of segments for information sources are relevant factors.

IV Concept/Term Analysis The use of ontological concepts in the information units is analyzed and improved by automated methods.

During the process, the exploration and visualization of the data is of core importance [11]. Since much of the data is generated during the process, the manual inspection and evaluation of the results need to be supported. Visualization methods help to understand existing deficiencies and motivate further process steps. In this paper, we introduce the generic visualization method *Linked Data City* that effectively supports the exploration and analysis of linked enterprise data during the semantification phase. The visualization of linked data differs from general ontology visualization methods [3, 6, 9], since usually linked data models exploit less relational structure but tend to be larger by orders-of-magnitude. We emphasize that the method is general usable for hierarchic structures in a way that it can be deployed very easily in different scenarios.

The rest of the paper is organized as follows: First we introduce the core components of a linked data city, namely *buildings* and (nested) *districts*. Then, we describe the current implementation of the approach and demonstrate the usefulness of the visualization method by examples. In the conclusions we show promising steps for further work.

2 Linked Data Cities

The metaphor for visualizing linked data as a city is inspired by the work of Wetzel [15]. In the original approach, the code of software applications is visualized as structures of a city. Figure 2 shows an example of a city visualization. Classes of software code are represented as *buildings* and code packages are defining the *districts* of a city. Special properties of classes and packages are communicated via color and size of the artifacts. Later this idea was adapted for the visualization of the test coverage after the evaluation of knowledge bases [7], where knowledge base elements are represented as buildings.

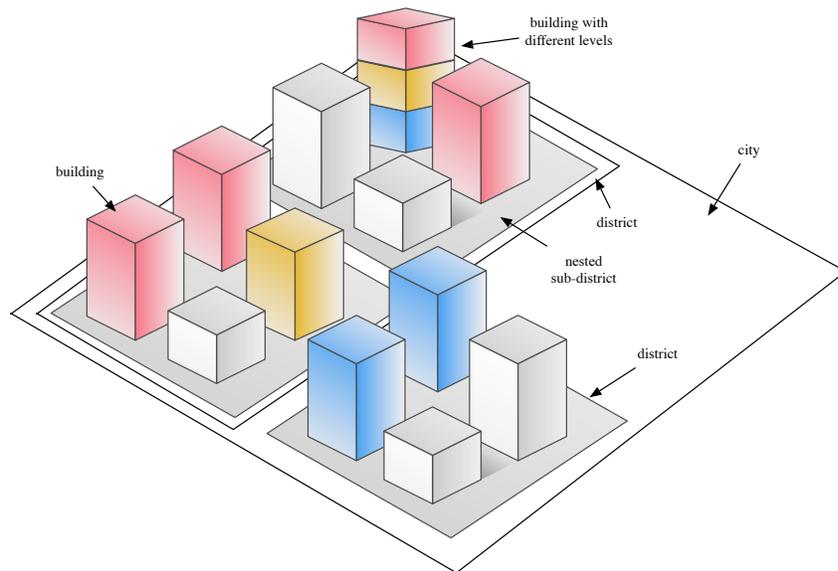


Fig. 2. Building, districts, and sub-districts of a data city.

Complex artifacts with part-of or hierarchical relations can be naturally visualized as a city: As we see in Figure 2, core elements of the artifacts are usually depicted as *buildings* of the city, that are grouped within *districts*. For deeper part-of or hierarchy relations the districts can be nested in *sub-districts*. The size of districts correlates with the number and sizes of the included buildings, whereas the size and color of the buildings represent specific performance indicators that are defined by the current analysis query. The specific configuration of buildings, districts, sizes, and colors is called *data city metric*. At the top of the figure we also see that buildings can have different levels. Different levels of a building are commonly used to include different metric attributes in the visualization.

For humans it is easy to understand the city metaphor. Like in the real world, local areas of the city are represented in districts and often the size of houses corresponds to their weight. For large cities (real and artificial) the user is familiar in incrementally explore the particular districts and buildings. For this reason, an advanced visualization application need to allow for the interactive exploration by drill-down and roll-up operations within the city.

3 Implementation

The presented visualization technique is implemented as a JavaScript library. That way, it can be easily integrated into (web-based) knowledge engineering tools but also runs as a stand-alone tool in a web browser. The definition of the city itself is represented as a JSON document. The following example shows books of a technical documentation that are represented as districts. Sections of a book are represented as buildings contained in the district.

```
{
  "Label": "Linked Data City 0815",
  "Districts": [
    {
      "Label": "Technical Documentation",
      "color": "#297B48"
      "Districts": [
        {
          "Label": "Repair Manual #4711",
          "color": "#848484#"
          "Buildings": [
            {
              "Label": "Section #4711.1",
              "color": "#0F2A65"
              "depth": 2,
              "height": 1,
              "width": 2
            },
            {
              "Label": ["Section #4711.2a",
                "Section #4711.2b"],
              "color": ["#0F2A65", "#AF5C0B"],
              "height": [1, 3],
              "depth": 2,
              "width": 2
            }
          ],
          ...
        }
      ]
    }
  ]
}
```

The definition of the city is straight-forward, as we can see that districts can be nested in other districts. A leaf district contains a collection of buildings in

a corresponding element “Buildings”. For the district “Repair Manual #4711” two buildings are shown. The first building “Section #4711.1” is an example for a simple building having only one level included. The second building shows two labels, colors and heights representing the two levels of the building.

4 Case Study: Metrics for Linked Enterprise Data in Technical Service Information Systems

The architecture of a city is defined by the applied *data city metric*, i.e., the definition of colors, sizes, and levels of the buildings and districts. In this section, we demonstrate the approach by two use cases.

The presented metrics were used in the context of an industrial semantification project, where information sources of technical service information were analyzed. Here, buildings of a city represent a special kind of elements of the linked enterprise data. In enterprise systems the data refers to a domain-specific ontology. For instance, machinery builders typically align their data (documentation, parts, 3D models, etc.) to an ontology of products, components and functions, cf. [10, 12]. Enterprise information resources—document sections, parts, and wiring diagrams—are usually annotated by one or more instances of this hierarchy, for example a repair paragraph is annotated by the involved components and influenced functions.

In the following we present two basic metrics that investigate (a) the use of the product structure within the available information resources and (b) the availability of annotations in the information resources.

Use Case: Usage of Product Structure (UPS)

The *product structure* of an enterprise defines how products are organized in different levels. This organization includes multiple hierarchies for representing the relation of components and parts, but also for the functions of the product.

The primary subject of the UPS analysis is the actual use of the elements defined in the product structure. The use of the elements corresponds to annotations done with these elements included in the data of the investigated information systems. The metric is applied to find out how well the product structure is used in current enterprise information.

In the visualization, leaf elements of the product structure are represented by buildings and upper elements of the product structure are represented as wrapping districts. The height of the buildings correlates with the number of uses within all considered information sources. Higher buildings are thus used more often.

At the left side of Figure 3, a zoomed building representing the component “Engine block” is shown. The building itself is located in the district with the name “Engine”. We see that the building consists of multiple levels specializing the location of occurrences. Here, the element was used most often in the resource “doc#1” and the resource “3d#3”.

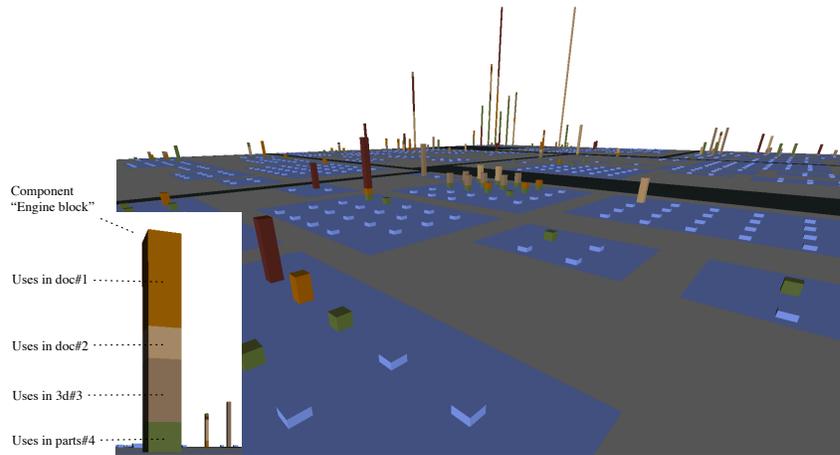


Fig. 3. Example for the usage of a product structure for a selection of technical documentation.

The visualization gives a very quick overview of actual application of a product structure. Unused areas can be easily spotted as well as elements with heavy use. Applied to a representative corpus of information resources, the visualization method points to areas in the structure that need refinement; both for lazy and frequent elements.

An interactive visualization is appropriate for very deep hierarchical structures. Then, buildings do not necessarily represent leaf-elements of the hierarchy but aggregated elements. Entering a building (e.g., by clicking on it) the building will drill-down the product structure and build a city visualization for all sub-elements contained in the aggregated element.

Use Case: Corpus Annotation Frequency (CAF)

Besides the actual use of the product structure the annotation frequency of the information resources is of prime interest. Usually, meta-data is attached to the information units to formally describe the contents. This meta-data mainly corresponds to elements of the product structure.

For the metric CAF, the city visualization is created as follows: The information sources in the corpus are represented as districts of the city, e.g., technical documentation, spare parts catalog, or FAQ data base. Sub-elements of these districts are further represented as nested sub-districts, e.g., a particular repair manual contained in the technical documentation or the spare parts catalog of a specific machine. Core information units are represented as buildings, for instance, a specific chapter of a repair manual in the technical documentation. The height of a building corresponds to its number of meta-data annotations; a

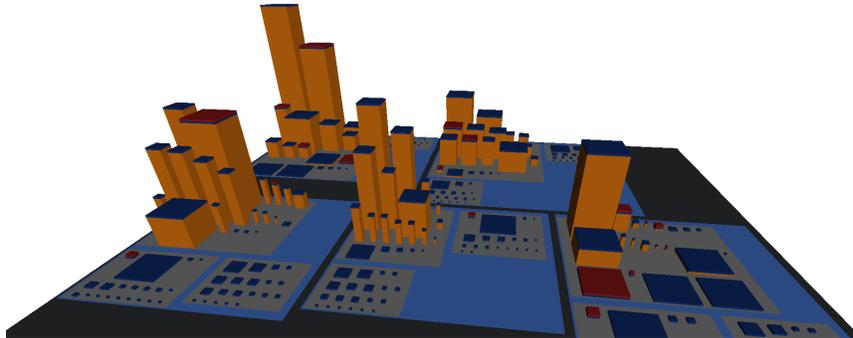


Fig. 4. Visualization of the annotations existing particular information units of a technical documentation for a specific machine.

buildings can have more than one level when different types of meta-data are included corresponding information unit. For instance, a chapter may include annotations of a component hierarchy but also of the functional hierarchy.

This visualization gives an overview of the corpus size and the existing annotations. Less annotated areas can be easily spotted but also districts (information sources, books types, etc.) with a high annotation quality. The results can help to motivate which areas of the structures need to be used much more frequently in information resources. With this knowledge, annotation initiatives (automated or manual) can be motivated and precisely planned.

5 Conclusions

Recently, more and more business information systems are transformed to linked enterprise data models. Appropriate visualization and exploration techniques support the semantification process of enterprise data. In this paper we presented Linked Data Cities, an interactive and generic method for the visualization of hierarchical structures. The actual visualization is defined by the application of a domain-specific metric. We introduced a number of metrics that showed its usefulness in an industrial semantification project.

In the future we are planning to improve the simplicity of the visualization by drill-down techniques, where similar buildings are clustered in aggregated building or districts. Then, even very large system structures can be (interactively) explored. Furthermore, we are working on the automated linkage of the city structure to existing linked data. Currently, scripts are used to transfer the information into the city data notation (the shown JSON). In the future, the automated transformation by SPARQL queries could be a possible simplification of this process.

References

1. Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Lehmann, J., Martin, M., Mendes, P.N., van Nuffelen, B., Stadler, C., Tramp, S., Williams, H.: Managing the life-cycle of linked data with the LOD2 stack. In: Proceedings of International Semantic Web Conference (ISWC 2012) (2012)
2. Berners-Lee, T.: Linked data (2009), <http://www.w3.org/DesignIssues/LinkedData.html>
3. Fluit, C., Sabou, M., van Harmelen, F.: Ontology-based information visualization. In: Visualizing the Semantic Web, pp. 36–48. Springer (2006)
4. Furth, S., Baumeister, J.: On the semantification of 5-star technical documentation. In: Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. pp. 264–271 (2015)
5. Furth, S., Baumeister, J.: Semantification of large corpora of technical documentation. In: Atzmueller, M., Oussena, S., Roth-Berghofer, T. (eds.) Enterprise Big Data Engineering, Analytics, and Management. IGI Global (2016), <http://www.igi-global.com/book/enterprise-big-data-engineering-analytics/145468>
6. Geroimenko, V., Chen, C. (eds.): Visualizing the Semantic Web. Springer, 2 edn. (2006)
7. Hatko, R., Baumeister, J., Puppe, F.: Coveragecity: Test coverage for clinical guidelines. In: The 8th Workshop on Knowledge Engineering and Software Engineering (KESE2012) (2012), http://ceur-ws.org/Vol-949/kese8-01_02.pdf
8. Hepp, M.: GoodRelations: An ontology for describing products and services offers on the web. In: Gangemi, A., Euzenat, J. (eds.) EKAW. Lecture Notes in Computer Science, vol. 5268, pp. 329–346. Springer (2008), <http://dblp.uni-trier.de/db/conf/ekaw/ekaw2008.html#Hepp08>
9. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E.: Ontology visualization methods - a survey. ACM Comput. Surv. 39(4) (Nov 2007), <http://doi.acm.org/10.1145/1287620.1287621>
10. Lin, J., Fox, M.S., Bilgic, T.: A product ontology. Enterprise Integration (1997)
11. Mader, C., Martin, M., Stadler, C.: Facilitating the exploration and visualization of linked data. In: Auer, S., Bryl, V., Tramp, S. (eds.) Linked Open Data—Creating Knowledge Out of Interlinked Data, pp. 90–107. Lecture Notes in Computer Science, Springer International Publishing (2014), http://dx.doi.org/10.1007/978-3-319-09846-3_5
12. Mohammad, N.N., Amin, I.M., Othman, R.M., Asmuni, H., Hassan, R., Kasim, S.: Design and implementation of product structure ontology. Ontology-Based Applications for Enterprise Systems and Knowledge Management p. 246 (2012)
13. Oberle, D.: How ontologies benefit enterprise applications. Semantic Web 5(6), 473–491 (2014)
14. W3C: SKOS Simple Knowledge Organization System reference: <http://www.w3.org/TR/skos-reference> (August 2009)
15. Wettel, R., Lanza, M.: Visualizing software systems as cities. In: Visualizing Software for Understanding and Analysis, 2007. VISSOFT 2007. pp. 92–99 (2007)

Case Representation and Similarity Assessment in the selfBACK Decision Support System

Kerstin Bach¹, Tomasz Szczepanski¹, Agnar Aamodt¹, Odd Erik Gundersen¹
and Paul Jarle Mork²

¹ Department of Computer and Information Science

² Department of Public Health and General Practice

Norwegian University of Science and Technology, Trondheim, Norway

<http://www.idi.ntnu.no>, <http://www.ntnu.no/ism>

Abstract. In this paper³ we will introduce SELFBACK , a decision support system that facilitates, improves and reinforces self-management of non-specific low back pain.

Keywords: Case-Based Reasoning, Case Representations, Data Streams, Similarity Assessment

1 Introduction

Low back pain is one of the most common reasons for activity limitation, sick leave, and disability. It is the fourth most common diagnosis (after upper respiratory infection, hypertension, and coughing) seen in primary care [2].

Self-management in the form of physical activity and strength/stretching exercises constitute the core component in the management of non-specific low back pain; however, adherence to self-management programs is poor because it is difficult to make lifestyle modifications with little or no additional support. In the SELFBACK project we will develop and document an easy-to-use decision support system to be used by the patient him/herself in order to facilitate, improve and reinforce self-management of non-specific low back pain. The decision support system will be conveyed to the patient via a smart-phone app in the form of advice for self-management.

The SELFBACK system will constitute a data-driven, predictive decision support system that uses the Case-Based Reasoning (CBR) methodology to capture and reuse patient cases in order to suggest the most suitable activity goals and plans for an individual patient. This will be based on data from two sources. One is a questionnaire, presented to the patient at suitable intervals, in order to capture general information (e.g. age) and subjective symptoms (e.g. the current degree of pain). Initially, patient information from the patient's clinician or general practitioner will also be added. The other is a stream of activity data

³ This paper is a resubmission from the 24th International Conference on Case Based Reasoning. Full paper: <http://www.idi.ntnu.no/kerstinb/paper/2016-ICCBR-Bach-et-al.pdf>

collected using a wristband. The incoming data will be analyzed to classify the patients current state and recent activities, and matched against past cases in order to derive follow-up advices to the patient. Two main challenges are to detect the activity pattern represented at a suitable level of abstraction, and to match that structure against existing patient descriptions in the case base. Combined with patient profile data from the questionnaire, and the current goal setting, this should enable the system to suggest the best next activity goal and plan for the patient.

Stratified care for patients with low back pain, based on initial pain intensity, disability related to low back pain, and fear-avoidance beliefs have been shown to improve patient outcomes as well as being cost-effective [1]. The SELFBACK system aims at further improving the stratified care approach by including data on the patients health and coping behaviour (i.e., the adherence to basic self-management principles) in order to support and prompt appropriate actions thereby empowering the patient to improve the self-management of their own low back pain. The SELFBACK system targets the self-management of non-specific low back pain by incorporating existing knowledge in the SELFBACK system to recommend advice that is personalised to the information input by the patient.

The overall SELFBACK hypothesis is that CBR can be applied to the general condition and activity pattern streams of patients with non-specific low back pain in order to effectively improve their rehabilitation processes. Based on this hypothesis, we are currently studying two core research issues: The case representation, i.e. what exactly should be in a case and how should this be expressed, and the corresponding similarity assessment method that operate on that structure. The primary focus of this paper is on case representation, with similarity assessment discussed in relation to the representation.

In the presentation we describe the case representation and case content as well as we introduce the applied similarity assessment. For both, case representation and similarity assessment, we conducted experiments using already existing data set from the domain and discuss these in the course of this work as well.

Acknowledgement The SELFBACK project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 689043.

References

1. Hill, J.C., Whitehurst, D.G.T., Lewis, M., Bryan, S., Dunn, K.M., Foster, N.E., Konstantinou, K., Main, C.J., Mason, E., Somerville, S., Sowden, G., Vohora, K., Hay, E.M.: Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *The Lancet* 378(9802), 1560–1571 (Oct 2011)
2. Wändell, P., Carlsson, A.C., Wettermark, B., Lord, G., Cars, T., Ljunggren, G.: Most common diseases diagnosed in primary care in stockholm, sweden, in 2011. *Family Practice* 30(5), 506–513 (2013)

Understanding Mathematical Expressions: An Eye-Tracking Study

Andrea Kohlhasse

Michael Fürsich

Neu-Ulm University of Applied Sciences

Extended Abstract

Intuitive knowledge management and user interfaces need a good understanding of the respective users and practices. We are especially interested in the target group of users of mathematical knowledge. One important mathematical practice consists in using mathematical expressions, which evolved as concise and precise representations of mathematical knowledge and as such guide and inform mathematical thinking.

We present an exploratory eye-tracking study in which we investigate how humans perceive and understand mathematical expressions. The study confirms that math-oriented and not math-oriented users approach them differently and reveals implicit mathematical practices in the decoding and understanding processes. Mathematical expressions are widely used in technical documents, therefore fitting mathematical user interfaces have to be created that empower their users to better utilise those tools of the trade. With our eye-tracking experiment we have identified a set of mathematical practices when decoding a math expression.

We observed that math-oriented and non-math-oriented participants have different visual approaches for math detection. In particular, non-math-oriented subjects read even complex mathematical expression from left to right as if it were text. In contrast, math-oriented probands decomposed mathematical expressions with clearly organized procedures. They chunked the expression into sub-expressions. They decoded the mathematical expression from left to right until the first meaningful sub-expression was grasped and then turned their attention towards unciphering this subexpressions. The relations among the distinct items within the subexpressions were also attended too, even among cross-relations between distinct subexpressions. Literals and variables were easily distinguished by most readers, whereas variable names were not the pivotal point essential for problem solving. This might have been different with mathematical symbols, that did not necessarily include all parameters of their information. The suggested six hypotheses enable a deeper understanding of the way mathematical expressions are perceived.

Interestingly, we could verify that the math-oriented subjects decoded a complex mathematical expression according to the MathML content tree of the expression in question. The existence of a content-oriented markup format at the same time as a presentation-oriented one is therefore justified for the aimed-for target-group of the MathML web format.

Note, that the set of mathematical practices is still hypothetical but it gives rise to interesting new perspectives for the design of mathematical user interfaces, particularly mathematical search engines.

Acknowledgements

The original paper was published in the Joint Proceedings of the Conference on Intelligent Computer Mathematics 2016 <http://www.cicm-conference.org/2016/eur-ws/CICM2016-WIP.pdf> at the Mathematical User Interfaces Workshop 2016 (see <http://www.cicm-conference.org/2016/cicm.php?event=mathui&menu=general>).

Synthesizing Invariants via Iterative Learning of Decision Trees

Pranav Garg, Daniel Neider, P. Madhusudan, and Dan Roth
University of Illinois at Urbana-Champaign, USA

Automatically synthesizing inductive invariants (i.e., statements about the configurations of a program that remain true during its execution) lies at the heart of automated program verification. Once an inductive invariant is found or given by the user, the problem of checking whether the program satisfies a specification can be reduced to logical validity of so-called verification conditions, which is increasingly tractable with the advances in automated logic solvers.

In recent years, the *black-box* or *learning* approach to finding invariants has gained popularity. In this data-driven approach, the synthesizer of invariants is split into two components. One component is a *teacher*, which is essentially a program verifier that can verify the program using a conjectured invariant; in case that the conjecture is inadequate to verify the program against the specification, the teacher can return a counterexample (i.e., a concrete program configuration) as a witness to why this is the case. The other component is a *learner*, which learns a candidate invariant from counterexamples returned by the teacher. Both components are run in alternation until the learner proposes an invariant that is adequate to verify the program.

One of the biggest advantages of the black-box learning paradigm is the possible usage of *machine learning techniques* to synthesize invariants. The learner, being completely agnostic to the program (its programming language, semantics, etc.), can be seen as a machine learning algorithm that learns a Boolean classifier of program configurations. However, Garg et al. [2] argue that learning from positively or negatively classified program configurations alone does not form a robust paradigm for learning invariants. Instead, the authors propose to allow the teacher to return implications $x \Rightarrow y$, where both x and y are program configurations, with the constraint that the learner must propose a classifier such that if x is classified as true, so is y (intuitively, implications capture the semantics of a program). The resulting learning setting is called *ICE learning* (short for implication-counterexamples) and entails that the learner now additionally needs to handle implications.

The goal of our paper [1] is to adapt Quinlan’s decision tree learning algorithms to the ICE learning model in order to synthesize invariants. Our key contributions are: (i) a generic top-down decision tree learning algorithm that learns from training data with implications; (ii) several novel “information gain” measures that are used to determine the best attribute to split on the current collection of examples and implications; (iii) a convergence mechanism that guarantees learning an invariant if one exists; and (iv) an extensive experimental evaluation.

1. Pranav Garg, P. Madhusudan, Daniel Neider, and Dan Roth. *Learning Invariants using Decision Trees and Implication Counterexamples*. In POPL 2016, pages 499–512. ACM, 2016.
2. Pranav Garg, Christof Löding, P. Madhusudan, and Daniel Neider. *ICE: A Robust Framework for Learning Invariants*. In CAV 2014, volume 8559 of LNCS, pages 69–87. Springer, 2014.

BISHOP – Big Data Driven Self-Learning Support for High-performance Ontology Population

Daniel Knoell¹, Martin Atzmueller², Constantin Rieder¹, and Klaus Peter Scherer¹

¹ Karlsruhe Institute of Technology
D-76344, Eggenstein-Leopoldshafen, Germany
firstname.lastname@kit.edu

² University of Kassel, Research Center for Information System Design
Wilhelmshöher Allee 73, 34121 Kassel, Germany
atzmueller@cs.uni-kassel.de

Abstract. Self-learning support systems are already being successfully used to support sophisticated processes. For the widespread industrial use, there are still challenges in terms of accessibility with respect to the process and the scalability in the context of large amounts of data.

This paper provides an example-driven view on the Bishop project for Big Data driven self-learning support for high-performance ontology population. We outline workflows, components and use cases in the context of the project and discuss methodological as well as implementation issues.

1 Introduction

Linked Enterprise Data requires the effective and efficient learning of ontologies. Typically, only large data sources provide the means for obtaining results with sufficient quality. Therefore, methods that work at large-scale are necessary, e. g., using high performance methods, resulting in increasing efforts concerning Big Data processing and management. In addition, typically specialized infrastructure needs to be set-up and configured, which is usually complicated and costly. Therefore, both accessibility and scalability of the applied methods and techniques need to be increased.

This paper presents the Bishop project that addresses these issues in order to provide a systematic approach towards large-scale self-learning support systems. We present an example-driven approach on the project and discuss specific workflows, components and use cases supported by appropriate tools. Hence, the remainder of the paper is structured as follows: We first provide an overview on the Bishop project, putting it into the context of related work, and discuss exemplary workflows and components in Section 2, before we present a set of use cases that are elaborated in a requirements engineering step in order to identify first measures and process forces. Overall, These use cases are used as a reference for the different architectural variants, e. g., in the context of natural language processing methods for self-learning from texts. Furthermore, we discuss suitable tool support in that context. Finally, we conclude with a summary and outlook on further steps in Section 3.

2 BISHOP by Example: Workflow and Components

BISHOP is part of the APOSTLE project, which is the acronym for “Accessible Performant Ontology Supported Text Learning“. While learning ontologies from text is not a novel approach and is e. g., used to learn the concept hierarchy out of web data [10], the Bishop project tackles the efficient and effective self-learning of ontologies for large data. The TELESUP Project [9], for example also deals with the automatic ontology population by using textual data, however, it does not consider Big Data.

In a first step, a conceptual framework is derived from the requirements that captures the decisions for integrating self learning methods into high performance environments. For that, different Big Data frameworks like Map/Reduce (Hadoop), Spark and Flink need to be investigated, in order to estimate the performance in the scope of the targeted data. Then a test scenario for the comparison of the results will be defined. After the set-up of the big data infrastructure, it will be evaluated with different persistence strategies. In parallel, it is necessary to find an easy way for the set-up of the big data environment and the deploying of existing Java applications. An additional parallel task is to find and efficient way for storing and querying huge amounts of semantic structures. Here, also intelligent mechanisms for persistence, distribution and parallelization will be devised.

By optimizing the accessibility and scalability, significant efficiency improvements in technical services for the creation of self-learning systems, such as expert systems and knowledge-based support systems are enabled.

2.1 Exemplary Workflows

The project consists of different parts which lead to different workflows. These workflows are processed in parallel and are described in the following subsections.

Calculating a Thesaurus The automatic generation of a thesaurus requires the steps, described in Figure 1.

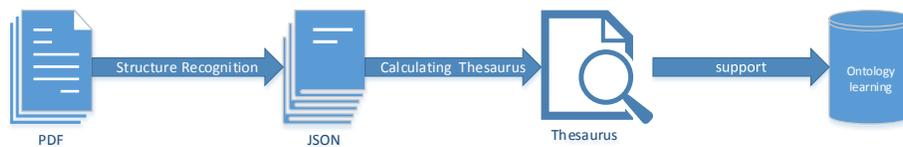


Fig. 1: Workflow: Calculating a Thesaurus

The data format in industry is often PDF. So in the first step the PDF files need to be converted to a more structured format like JSON. This is an important step, which is also necessary for other areas of the processing and is detailed in section 2.3. The JSON files serve as input for the application which calculates the thesaurus. A possible application for this task could be “JoBimText” [15]. A further description is given in section 2.3. Finally, as last step, the thesaurus can be used for ontology learning.

Easy Deployment The set up of an **infrastructure** to fulfill necessary tasks can be very difficult and time consuming. Furthermore it should be noted that each user has a different work equipment. Therefore, an easy deployment of such an infrastructure is needed to start the data processing quickly and platform independent keeping the frustration level low by avoiding installing, configuring and setting up activities.

A modern technology facing these restrictions could be a **container based approach** of deployment. One possible solution considering these limitations could be the open-source project Docker that provides suitable features by deploying applications inside software containers. The so called docker images are providing the applications which are running in the docker containers and accelerating the distribution and deploying efforts. By deploying a ready to start configuration with a preset environment and set of applications the expectation is a more user friendly set up process that allows a quick start.

In addition to the prepared configurations and on the basis of the above a further important step is to design a set of conventions to reduce the complexity of mandatory configurations. One possible solution could be the design of an appropriate configuration and **set up wizard** that guides the user through the complex processes. This kind of support could be a helpful extension because it has been in use for decades (e. g. classic installer wizards) and has proven its worth.

A second point of the easy deployment is how to get an existing **Java application** running on the big data environment, see Figure 2. Here appropriate conversion methodologies need to be developed.

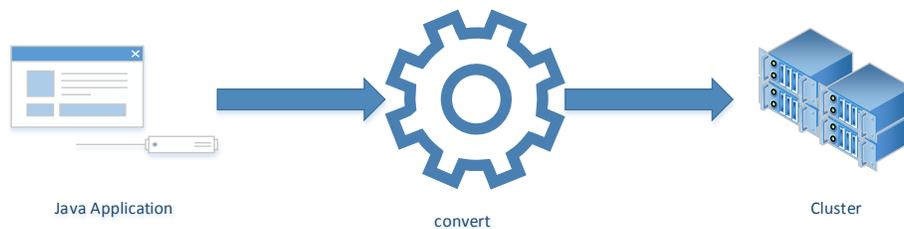


Fig. 2: Workflow: deploying a single computer application on a cluster

Storing and Querying Semantic Structures According to the current state of the art, the management of huge amounts of semantic data (ontologies) is still inefficient. For larger amounts of data the current solution is to merely use larger amounts of main memory. However, if the limit of the currently used memory exceeds the volume of the semantic data, there is at the moment no effective solution. Therefore, a problem-solving approach, which can handle the requirements of huge ontologies is necessary. An approach is intended, which retains the advantages of current solutions as well as possible and fixes the weaknesses in dealing with large amounts of data. For this, innovative methods needs to be developed and integrated into the overall approach.

2.2 Components

In Figure 3 the components of the whole project are illustrated. Via *Business Services* all components of the system are loosely connected. This also includes the *Self-learning Methods* which are evaluated under the aspects of parallelizability and scalability. These methods can either be in the field of text learning [6], for example NLP, and in the field of learning with structured data. After the evaluation follows a review and then based on the results an adaptation.

Various forms of parallelization are implemented and evaluated. The component *Persistence* allows the permanent storage of the documents. The aim is to develop a library which offers different options how to store the documents. It detects depending on the required scaling and the used system which persistence method is to be used. In the first step the documents are stored on the local file system. After that the implementation of additional storage capabilities, such as the Hadoop Distributed File System (HDFS) or MapR-DB, is done. These can be automatically selected when storing large amounts of data.

Therefore the scalability of the infrastructure has to be evaluated. The results of the evaluation allow the construction (or potentially refinement) of rules for decision-making for the storage strategy used. Parameters such as size and type the data are also considered. The last component is the *Ontology Proxy* which enables the storing and accessing of the semantic representation of the documents. Therefore various existing solutions are evaluated and adjusted substantially or completely redeveloped. Furthermore, it is examined whether techniques from the database environment can be applied and whether these yield performance improvements.

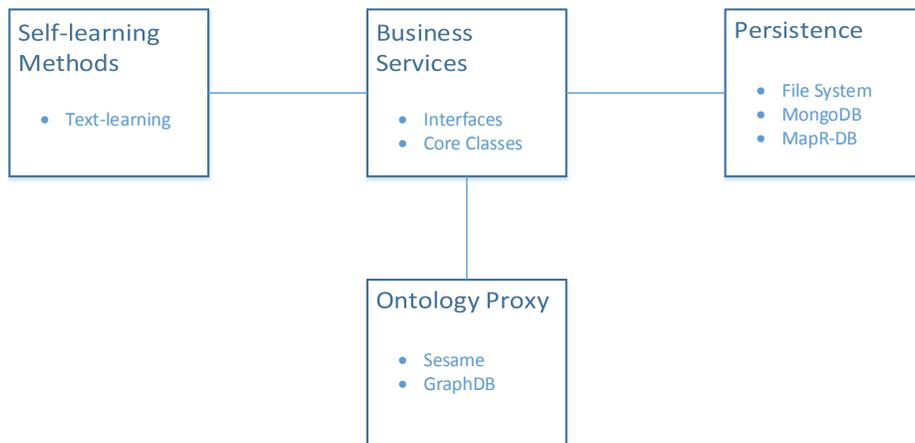


Fig. 3: Components

2.3 Use Cases

This section outlines two use cases in the context of the Bishop project concerning basic techniques for learning from texts, i. e., structure recognition and calculating a Thesaurus. After that, we discuss options for tool support in that scope, considering potential Big Data processing and management methods in the context of processing unstructured, i. e., textual information.

Use Case 1: Support Structure Recognition in PDF Files A common problem in the enterprise environment is, that important data is only available as PDF-Files. It is easy to get content of the PDF-Files as plain text, but that is usually not that helpful, because the structure of the documents gets lost. Without the structure, there is no way to find out if a specific phrase in the document was a heading, a headline, a footnote or even a caption. In the most cases this information is extremely important for the further processing. The recognition of the structure of a PDF-File is a difficult task which is very time consuming and only a few applications are good at it. In combination with a huge amount of PDF-Files, like it occurs for example in the field of the technical documentation, there is a long time waiting for results.

In order to decrease the overall processing time, it is useful to distribute the application for example on a High Performance Cluster. There are at least two ways for the distribution. The first way is to process every PDF-File on a different node in the cluster. This is expected to be a good solution if there are a lot of files, which are not that big. If there are only a few, but huge PDF-Files it can be helpful to split the files into many parts and distribute the parts of the files. This would be the second way. It has to be evaluated, if the two ways perform as expected, to know which way fits to the correlated case. The optimal behaviour of the resulting application would be, that it picks the right way of distribution, depending on the dataset.

Use Case 2: Calculating a Thesaurus A domain-specific thesaurus can have huge advantages on the task of ontology learning. Especially on the lower layers of the ontology learning layer cake [7], like terms and synonyms layer, it can be useful and improve the results drastically. The problem is, that for the most domains, there is no suitable domain-specific thesaurus available. Furthermore, building up a thesaurus is a time consuming task which needs the involvement and knowledge of experts. However, the time of the experts is typically rare and expensive. These two facts make the manual construction of an domain-specific thesaurus difficult. An automation of this task is also difficult and needs a lot of domain specific documents.

There are approaches like “JoBimText” [15] which can calculate a thesaurus out of huge corpora, but do not take advantage of the structure of the documents. This could have an enormous impact on the quality of the results. The calculation of the thesaurus should also be distributed, because of the huge amount of documents, which need to be processed. Otherwise it would take too much time for the industrial usage.

2.4 Tool Support

According to the *four V* criteria [11] (i. e., velocity, volume, variety, and veracity), big data requires efficient methods to handle the rapidly incoming data with appropriate response time (velocity), the large number of data points (volume), many different heterogeneously structured data sources (variety), and data sources with different quality and provenance standards (veracity). In the context of unstructured and semi-structured data, several challenges have to be addressed, such as information extraction [1] for textual data, as well as integration techniques for the comprehensive set of data sources. For semi-structured data, e. g., rule-based methods [3] or case-based reasoning systems [5] can often be successfully applied. According modeling and indexing techniques can be implemented, e. g., using the Map/Reduce framework [8], see below.

Before starting with a data processing framework, different questions and requirements need to be clarified, e. g., according to the types, structure and accuracy of data that is to be implemented, cf. [12]. We focus on according tools for Big Data processing, analytics, and management in the following.

Lambda Architecture According to Marz and Warren [14], system properties of a Big data system typically exhibit the following system properties: They should provide a *general* data framework that is *extensible*, enables *ad-hoc queries* with *minimal maintenance*, and *debugging capabilities*. For data storage, this implies mechanisms for handling the *complexity* of data, e. g., for preventing corruption issues and maintenance issues. Further, *robustness and fault-tolerance* should be enforced, as well as *low latency reads and updates*. This also points to *scalability* issues concerning horizontal and vertical scalability, and the option of obtaining *intermediate results* and views, according to some concept of reproducibility.

The lambda architecture incorporates these system principles and especially tackles the concept of reproducibility of results and views for dynamic processing. Essentially, it allows to compute arbitrary functions on arbitrary datasets in real-time [14]. The lambda architecture is structured into several layers briefly, summarized as follows:

- Batch layer: continuously (re-)computes batch views using immutable data records.
- Serving layer: indexes query view, performs updates, and provides access to the dataset. Only batch updates and random reads are supported, no (distributed) writes.
- Speed layer: high-latency updates; fix batch layer lag; needs fast algorithms for incremental updates.
- Complexity isolation: random writes only need to be supported in speed layer. Results are then merged with the precomputed data from the batch layer.

Map/Reduce Map/Reduce[8] is a paradigm for scalable distributed processing of big data, that can be utilized for implementing, e. g., the batch layer. Its core ideas are based on the functional programming primitives *map* and *reduce*. Whereas *map* iterates on a certain input sequence of key-value pairs, the *reduce* function collects and processes all values for a certain key. The Map/Reduce paradigm is applicable for a computational task, if it can be divided into *independent* subtasks, such that there is no required communication between these. Then, large tasks can be split up into subtasks according to a typical divide-and-conquer strategy, e. g., for local exceptionality detection [4].

Map/Reduce is a powerful paradigm for processing big data – with a prominent implementation given by the *Hadoop* framework³ supported by the HDFS filesystem, and big data databases such as Hive⁴ and HBase⁵. Map/Reduce tasks can also be utilized for batch processing in the Lambda architecture discussed above, such that continuous views are (re-)computed by the respective Map/Reduce jobs. These batch tasks can then be complemented by tools for distributed realtime computation like the Storm framework⁶, or the Flink⁷ platform. This allows a comprehensive data processing pipeline for big data in the Lambda architecture, combining realtime together with Map/Reduce techniques. Alternatives to Map/Reduce, especially considering *in-memory computation* with large datasets include, for example, the Spark⁸ [17] and Flink platforms.

Big Data Management NoSQL Databases (Not Only SQL) offer high performance and high availability [16], if no ACID (Atomic, Consistent, Isolated, Durable) transactions are needed. These databases perfectly fit in our Big Data environment. In our case, we use JSON files, which should be stored in the database. A lot of document based NoSQL databases use this format to store the data on the filesystem. So it is quite simple to use a document based database like MongoDB⁹, Apache CouchDB¹⁰ or MapR-DB¹¹. MongoDB and Apache CouchDB have own solutions for the distribution of the database. MapR-

DB is a In-Hadoop NoSQL database that supports JSON document models and wide column data models and can be run in the same cluster as Apache Hadoop and Apache Spark. This has the benefit of an easy integration in the big data environment, which will contain Hadoop and/or Spark. The architecture is shown in Figure 4. The *Databases* and the *File Systems* are connected to the *Interface Layer*, which enables the access of the *Frontend* and the *Big Data Framework*.

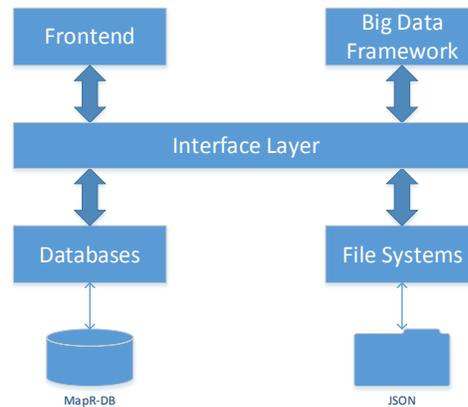


Fig. 4: Big Data Management Architecture based on MapR-DB.

³ <http://hadoop.apache.org/>

⁴ <http://hive.apache.org/>

⁵ <http://hbase.apache.org/>

⁶ <http://storm.apache.org/>

⁷ <http://flink.apache.org/>

⁸ <http://spark.apache.org/>

⁹ <https://www.mongodb.com/>

¹⁰ <http://couchdb.apache.org/>

¹¹ <https://www.mapr.com/products/mapr-db-in-hadoop-nosql>

3 Conclusions

This paper presented the Bishop project that investigates methods for Big Data driven large-scale self-learning support for high-performance ontology population. In an example-driven approach we discussed workflows, components, use cases, and tools.

For future work, we will investigate the proposed Big Data frameworks and develop corresponding data processing and analytics methods, also aiming at a methodology for easy cluster set up. Other interesting future directions are given by efficient (distributed) information extraction, e. g., [13] and refinement methods, e. g., [2], for advancing high-performance approaches for ontology population using self-learning support systems.

Acknowledgements. The work described in this paper is funded by grant ZIM-KOOP ZF4170601BZ5 by the German Federal Ministry of Economics and Technology (BMWi).

References

1. Adrian, B.: Information Extraction on the Semantic Web. Ph.D. thesis, DFKI (2012)
2. Atzmueller, M., Baumeister, J., Puppe, F.: Introspective Subgroup Analysis for Interactive Knowledge Refinement. In: Proc. FLAIRS. pp. 402–407. AAAI, Palo Alto, CA, USA (2006)
3. Atzmueller, M., Kluegl, P., Puppe, F.: Rule-Based Information Extraction for Structured Data Acquisition using TextMarker. In: Proc. LWA. University of Würzburg, Germany (2008)
4. Atzmueller, M., Mollenhauer, D., Schmidt, A.: Big Data Analytics Using Local Exceptionality Detection. In: Enterprise Big Data Engineering, Analytics, and Management. IGI Global, Hershey, PA, USA (2016)
5. Bach, K.: Knowledge Acquisition for Case-Based Reasoning Systems. Ph.D. thesis, Dr. Hut Verlag München (2012)
6. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press (2005)
7. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer, New York, N.Y. and London (2006)
8. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM 51(1), 107–113 (Jan 2008)
9. Furth, S., Baumeister, J.: TELESUP - Textual Self-Learning Support Systems. In: Proc. LWA 2014 (FGWM Workshop). RWTH Aachen University, Aachen, Germany (2014)
10. Karthikeyan, K., Karthikeyani, V.: Ontology Based Concept Hierarchy Extraction of Web Data. Indian Journal of Science and Technology 8(6), 536 (2015)
11. Klein, D., Tran-Gia, P., Hartmann, M.: Big Data. Inform. Spektrum 36(3), 319–323 (2013)
12. Klöpfer, B., Dix, M., Schorer, L., Ampofo, A., Atzmueller, M., Arnu, D., Klinkenberg, R.: Defining Software Architectures for Big Data Enabled Operator Support Systems. In: Proc. IEEE International Conference on Industrial Informatics. IEEE, Boston, MA, USA (2016)
13. Kluegl, P., Atzmueller, M., Puppe, F.: Meta-level information extraction. In: Proc. KI. LNCS, vol. 5803, pp. 233–240. Springer, Berlin / Heidelberg, Germany (2009)
14. Marz, N., Warren, J.: Big Data: Principles and Best Practices of Scalable Realtime Data Systems. Manning Publishers, Shelter Island, NY, USA, 1. edn. (2013)
15. Riedl, M., Biemann, C.: Scaling to Large Data: An Efficient and Effective Method to Compute Distributional Thesauri. In: EMNLP. pp. 884–890 (2013)
16. Tudorica, B.G., Bucur, C.: A Comparison between Several NoSQL Databases with Comments and Notes. In: International RoEduNet Conference. pp. 1–5. IEEE (2011)
17. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: Cluster Computing with Working Sets. In: Proc. USENIX. HotCloud, Berkeley, CA, USA (2010)

The Review of Subject Analysis: A Knowledge-based Approach facilitating Semantic Search

Sebastian Furth¹, Volker Belli¹, and Joachim Baumeister^{1,2}

¹denkbare GmbH, Friedrich-Bergius-Ring 15, 97076 Würzburg, Germany

²University of Würzburg, Institute of Computer Science,
Am Hubland, 97074 Würzburg, Germany

{sebastian.furth,volker.belli,joachim.baumeister}@denkbare.com

Abstract. Semantic Search emerged as the new system paradigm in enterprise information systems. However, usually only small amounts of textual enterprise data is semantically prepared for such systems. The manual semantification of these resources typically is a time-consuming process. The automatic semantification requires deep knowledge in Natural Language Processing. Therefore, in this paper we present a novel approach that makes the underlying Subject Indexing task rather a Knowledge Engineering than a Natural Language Processing task. The approach is based on a simple but powerful and intuitive probabilistic model that allows for the easy integration of expert knowledge.

Keywords: Subject Indexing, Document Classification, Semantic Search

1 Introduction

Historically, Subject Analysis and *Subject Indexing* [10,1] had been a rather manual task, where librarians or catalogers tried to index large corpora of documents according to a given set of controlled subjects. A more technical but prominent example for large scale Subject Indexing is the web catalog from the early Yahoo times, where websites had been indexed with certain topics. Regardless of which medium was used, catalogers typically tried to determine the overall content of a work in order to identify key terms/concepts that summarize the primary subject(s) of the work. An indexing step enabled in-depth access to parts of the work (chapters, articles, etc.). Therefore, the item was conceptually analyzed (what is it about?) and subsequently tagged and cataloged with subjects from a controlled vocabulary.

Nowadays, *Semantic Search* [8] applications belong to the state of the art in Information Retrieval. In contrast to traditional search engines ontologies are used to connect multi-modal content with semantic concepts, which can then be exploited during the retrieval to improve search results. Therefore, users of Semantic Search applications typically formulate their search queries as semantic concepts. Then a retrieval algorithm might expand the query considering

ontological information. Finally, a look-up method maps the concepts to actual search results using an index from concepts to information resources.

With the growing amount of information manually maintaining catalogs or indices became almost impossible. However, catalogs and indices are typically built for a specific problem domain. For many domains formal knowledge in form of ontologies exists and comprises decent amounts of terminology and relational information. Thus, we describe a novel approach for automatic Subject Analysis that allows for the easy integration of formal domain knowledge. We have built an intuitive probabilistic model that makes Subject Analysis not a Natural Language Processing but rather a Knowledge Engineering task. Therefore, the approach allows that domain experts can control the analysis by expressing their knowledge about relations between concept classes and the importance of certain document structures.

The remainder of the paper is structured as follows: Section 2 formally defines the Subject Analysis problem and discusses related work. In Section 3 we present our Knowledge-based Subject Analysis approach. Section 4 describes experiences made with our approach in industrial scenarios. We conclude with a discussion of our approach in Section 5.

2 Problem Description

2.1 Controlled Vocabularies

The fundamental requirement for Subject Analysis and the subsequent Subject Indexing is the existence of a controlled vocabulary. Historically, a controlled vocabulary defined the way how concepts were expressed, provided access to preferred terms and contained a term's relationships to broader, narrower and related terms. Nowadays, such information is typically modeled by standardized ontologies [9,13,12], where terms are embedded in complex networks of concepts covering broad fields of the underlying problem domain. Typical examples are ontologies powering semantic enterprise information systems. In such systems users interact using concepts that are company-wide known and valid. An increasing amount of companies maintain corresponding ontologies as they are the key element for the interconnection of enterprise systems and data [18]. If such ontologies do not exist, the construction is usually very reasonable under cost-benefit considerations, as they support not only semantic information systems but are also a vehicle for the introduction of more elaborate services like Semantic Autocompletions [11] or Semantic Assistants. In this paper, we formally define a controlled vocabulary as follows:

Definition 1 (Controlled Vocabulary). *A controlled vocabulary is an ontology $O = (T, C, P)$ that contains a set of terms T that are connected to a set of concepts C . Concepts $c \in C$ are connected to other concepts using properties $p \in P$.*

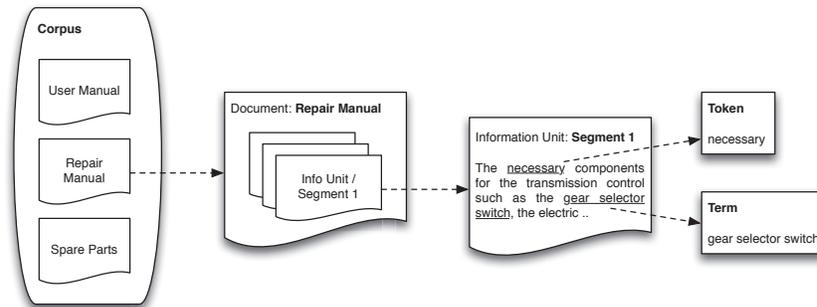


Fig. 1. Partitioning a corpus to information units.

2.2 Subject Indexing

Assuming that such an ontology/controlled vocabulary exists the task is to examine the subject-rich portions of the item being cataloged to identify key words and concepts. Therefore existing textual resources must be partitioned to sets of reasonable *Information Units* [6,3,17,4] (see Figure 1). Then the task can be defined as follows:

Definition 2 (Subject Indexing from information units). *For each Information Unit $i \in I$ find a set of concepts $C_i \subseteq C$ from an ontology O that describe the topic of the corresponding text best.*

An information unit i has an associated bag of term matches M_i , i.e. a list of terms from a domain ontology/controlled vocabulary that occur in a particular information unit. Given the bag of term matches the task can be specialized as follows:

Definition 3 (Subject Indexing from bags of words). *Given a bag of term matches M_i determine the underlying topics in the form of a set of concepts $C_i \subseteq C$ from an Ontology O .*

The availability of formalized domain knowledge is usually a valuable support factor for tasks that cover certain aspects of a problem domain [14,15,16]. We claim that this is also true for Subject Indexing where the selection of topics can profit from formalized background knowledge. Thus, the integration of domain knowledge in the annotation mechanism becomes a critical success factor and the task can be further refined as follows:

Definition 4 (Subject Indexing with background knowledge). *Given a bag of term matches M_i determine the underlying topic in the form of a set of concepts $C_i \subseteq C$ from an Ontology O considering the domain knowledge contained in Ontology O .*

2.3 Related Work

Topic Analysis is a relatively wide field of research and is strongly influenced by Document Classification and Document Clustering approaches. Notable approaches exist in particular among latent methods, i.e. topics are not expressed in form of explicit concepts but as a set of key terms. Prominent examples are Latent Dirichlet Allocation (LDA)[2] and Latent Semantic Analysis (LSA)[5]. Regarding the deduction of explicit topics Explicit Semantic Analysis [7] is a well-known approach.

3 Probabilistic Subject Analysis

In the following, we will first present a basic probabilistic model that is based mainly on weighted semantic relations between terms and concepts. The model can be tailored to integrate expert knowledge for a certain domain specific controlled vocabulary. The basic model will be extended in order to also consider document characteristics, like important document structures (e.g. headlines) or formatting information (e.g. bold text).

3.1 Basic Probabilistic Model

The basic probabilistic model is founded on observable text/term matches, relations between these terms and potential topics (concepts) and a strong independence assumption between all features. The model connects the features as follows:

1. Starting from a text match $match \in M_i$ in an information unit $i \in I$ the model derives potentially corresponding terms $t \in T$.
2. The model *optionally* weights the term $t \in T$ with respect to the covering document structure of the corresponding text match $match \in M_i$, e.g. term occurrences in headlines might be more important.
3. Given a term $t \in T$ the model looks for concepts $c \in C$ that can be described with this term, i.e. which concepts have this term as label and how specific is this label.
4. The concepts $c \in C$ derived from the model on basis of the text/term match $m \in M_i$ might have relations to topic concepts $topic \in C_i$ with $C_i \subseteq C$. The model exploits ontological information for the derivation of topic concepts $topic \in C_i$ from observed (term) concepts $c \in C$ resulting in a topic probability for a text/term match.
5. The derived topic probabilities for each text match $match \in M_i$ get aggregated in order to compute the overall topic probabilities for an information unit $i \in I$.

Given a bag of term matches M_i for an information unit $i \in I$, we realized steps (1) to (3) by computing the topic probabilities for each text/term match $match \in M_i$:

$$\mathbf{P}(\mathbf{topic} \mid \mathbf{match}) = \alpha * P(\mathbf{topic} \mid c) * P(c \mid t) * P(t \mid \mathbf{match}). \quad (1)$$

Therefore, we consider the confidence of a term match $P(t | match)$, i.e. the probability of a certain term t given a textual match $match$. Additionally we take the specificity $P(c | t)$ of a term t for a certain concept c into account. Unique labels have the maximum specificity of 1.0. The relevance of the concept in focus c for a topic concept $topic$ is $P(topic | c)$. This relevance gets computed on basis of ontological information between both concepts. The relevance is maximum if both considered concepts are equal (identity). Finally, we use the constant prior α to express the linguistic uncertainty that a certain topic is not meant given a certain term match. This avoids that one perfect term match pretends other topics to get more important, i.e. it regulates how many related term matches are necessary in order to outperform one perfectly matching term. We then compute the topic probabilities for an information unit $i \in I$ (step 4) on basis of the topic probabilities for each term match $match \in M_i$:

$$\mathbf{P}(\mathbf{topic}) = 1 - \prod_{\substack{M_i \\ match}} (1 - P(topic | match)). \quad (2)$$

The result is a set of topics $topic \in C_i$ with associated probabilities that express how well a certain topic fits to the terminology observed in the information unit. This computation assumes independence between the term matches in M_i according to Bayes' Theorem. The independence assumption might not perfectly reflect reality but is a sufficient approximation in this application scenario.

3.2 Extended Probabilistic Model

The basic probabilistic model can be extended, such that it also considers distinctive document characteristics as valuable background knowledge. In many specialized publications like technical documents or textbooks document structures indicate the underlying topic or support at least the discrimination of multiple topic candidates. Typical examples are headlines or formatted text (italics, bold, underlined).

The basic probabilistic model uses a constant prior α that expresses the linguistic uncertainty that a topic is not meant by a certain term match. We extend the basic model, such that the prior is not constant but depends on the document structure where the term match was observed. Therefore, document structures get weighted according to their importance for the deduction of a topic for an information unit. Assuming that for each document structure a weight w exists (default 1.0) the value for the prior α is computed as follows:

$$\alpha_{\mathbf{adaptive}} = 1 - (1 - \alpha_{\mathbf{constant}})^w. \quad (3)$$

This procedure also allows to discriminate document structures that are inappropriate for the topic deduction, e.g. references/links to other documents.

3.3 Knowledge Representations and Derivation of Probabilities

The preceding sections introduced a simple but powerful and intuitive probabilistic model for Subject Analysis. However, the primary target remains that the Subject

Analysis of large document corpora becomes rather a Knowledge Engineering than a Natural Language Processing task. Therefore, the proposed probabilistic model allows for the easy adaptation to characteristics of a domain specific controlled vocabulary and the corresponding corpus of documents that shall be subject indexed. The following section describe the knowledge-based adaptation, i.e. the definition of basic conditions for the derivation of probabilities.

Term Confidence $P(t | match)$ The term confidence $P(t | match)$ expresses how certain a text match is actually a term occurrence. The computed confidence depends on the quality of the text match. A perfect match, i.e. the text match $match$ is equal to the term t results in the maximum confidence of 1.0. The usage of fuzzy string matching techniques like order independent matching, stemming etc. might lower the confidence of term matches. Therefore, implementations of the presented probabilistic approach should allow for the configuration of different fuzziness levels and adjust the confidence accordingly.

Term Specificity $P(c | t)$ Given a term the model must derive all concepts $c \in O$ that can be described by this term. The model must also express how specific a term is for a concept $P(c | t)$, i.e. handle ambiguous terms like “apple” which can be the name of a company or a fruit. In the context of technical documents, we might encounter terms like “nut”, “engine” or “screw” that are very ambiguous and thus unspecific. Therefore, the specificity of a term must be distributed over all potential concepts. In the simplest case the specificity can be distributed equally over all concepts. Unambiguous terms always have a specificity of 1.0. However, experts’ knowledge might be used to prefer certain concepts. This might be useful if some concepts of an ontology are not applicable, e.g. because components they represent are not included in certain machines.

Concept Relevance $P(topic | c)$ Then, given a concept the model must be able to determine how relevant it is for certain topics $P(topic | c)$. The procedure is always the same and is explained by the example of technical documents. In technical documents the occurrence/observation of a concept describing a component might be relevant for a couple of concept topics: (1) machine functions relying on this component, (2) parent components or (3) the component itself.

In general, we assume that the relevance of a concept for a topic decreases the larger the distance between both concepts is in the underlying ontology. However, experts’ might know that in certain situations (documents) the occurrence of a concept is much more indicative for specific topics than for others. For example in operator manuals component terms might also indicate functions while they typically do not in repair manuals because usually an operator wants to "operate a function", whereas a technician usually wants to "repair a component".

For the calculation of the concept relevance distances between concepts and topic concepts are extracted/queried from the ontology. Expert knowledge can be used to weight these distances according to the properties $p \in P$ involved. This

way background knowledge regarding the relevance of certain concepts under certain circumstances can be included in the model. Finally the weighted distances between the concept in focus c and the topic concept $topic$ get transformed to a probability. We propose the usage of a normalized sigmoid function to avoid overestimation of the distance. The parameters β and γ can be used to control the sigmoid function and thus the overall importance of the concept relevance:

$$\mathbf{P}(\mathbf{topic} \mid \mathbf{c}) = \frac{1 + e^{(-\beta)*\gamma}}{1 + e^{(distance-\beta)*\gamma}} \quad (4)$$

Linguistic Uncertainty α In the basic probabilistic model the parameter α is constant. In the extended model the parameter α can be adjusted, such that it can prefer or discriminate term occurrences in certain document structures. Therefore, domain experts can define weights w for certain document structures (default 1.0). Values for w greater than 1.0 prefer, values smaller than 1.0 discriminate terms in certain structures respectively. During the computation of the value for the adaptive linguistic uncertainty $\alpha_{adaptive}$ an implementation has to consider the value accordingly.

4 Extended Example

An exhaustive and thorough evaluation of the presented approach is subject to future work. However, we have already applied the probabilistic model in an ongoing industrial semantification project with promising results. In the following we briefly describe the key aspects of the case study.

4.1 The data set

In the case study the task is to semantify a given corpus of technical documents provided in PDF format. The corpus comprises several thousand pages of technical information, spreaded over different documents like operator manuals, functional descriptions or repair and maintenance instructions. The semantification partitions the PDF files to reasonable segments (information units). Then, each information unit is subject indexed with respect to an existing ontology.

The ontology contains information about the hierarchical structure of components in the corresponding machine as well as functional connections between components (see Figure 2 for a simplified visualization). Labels are attached to all concepts.

4.2 Parametrization of the Probabilistic Model

The probabilistic model has been parametrized to incorporate existing domain knowledge. Therefore, we used the tailoring possibilities described in Section 3.3 as follows:

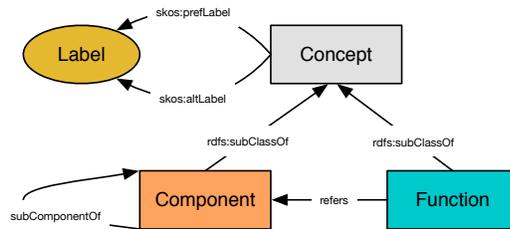


Fig. 2. Simplified visualization of the domain ontology.

- **Term Confidence** $P(t | match)$: We allowed order independent lookups without decreasing term confidences. Matches that had only been possible due to stemming have been discriminated.
- **Term Specificity** $P(c | t)$: We have distributed the specificity equally over all concepts, i.e. if a term is attached to two concepts, the specificity of the term is 0.5 for both concepts.
- **Concept Relevance** $P(topic | c)$: For operating manuals we slightly preferred the **refer** property, in descriptive manuals the **subComponentOf** property respectively.
- **Linguistic Uncertainty** $\alpha_{adaptive}$: We defined weights w greater than 1.0 for headlines and captions, i.e. preferred term matches occurring in the heading of sections and the descriptions of images.

A formal evaluation has not yet been performed. However, experts reviewed the derived topics and confirmed a noticeable improvement over a previous implementation based on Explicit Semantic Analysis.

5 Conclusion

In this paper we presented a novel approach for automatic Subject Indexing, i.e. the indexing of information units with respect to a controlled vocabulary/ontology. The presented approach is based on a simple but powerful and intuitive probabilistic model. We claim that this approach does not require training data but facilitates the easy incorporation of experts' domain knowledge and thus is highly adaptive. The adaptiveness through experts' knowledge makes automatic Subject Indexing rather a Knowledge Engineering than a Natural Language Processing task. Thus, large scale semantification of enterprise corpora becomes possible. The approach has not yet been evaluated thoroughly. However, its application in industrial case studies yielded promising results.

Besides an exhaustive evaluation future directions include the addition of learning methods. Therefore, we consider incorporating latent approaches as preprocessors to adjust concept relevances based on term frequencies in the underlying corpus. Additionally, we plan to investigate whether simulated annealing

can be used to learn weights w for document structures. We also plan to consider background knowledge about (hierarchical) connections between document structures in the model, e.g. the consideration of neighbour or parent segments' topics.

Acknowledgments

The work described in this paper is supported by the Bundesministerium für Wirtschaft und Energie (BMWi) under the grant ZIM ZF4172701 "APOSTL - Accessible Performant Ontology Supported Text Learning".

References

1. Albrechtsen, H.: Subject analysis and indexing: from automated indexing to domain analysis. *Indexer* 18, 219–219 (1993)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 993–1022 (2003)
3. Borkar, V.R., Deshmukh, K., Sarawagi, S.: Automatic segmentation of text into structured records. In: Mehrotra, S., Sellis, T.K. (eds.) *SIGMOD Conference*. pp. 175–186. ACM (2001), <http://dblp.uni-trier.de/db/conf/sigmod/sigmod2001.html#BorkarDS01>
4. Choi, F.Y.Y.: Advances in domain independent linear text segmentation. In: *ANLP*. pp. 26–33 (2000), <http://dblp.uni-trier.de/db/conf/anlp/anlp2000.html#Choi00>
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6), 391–407 (1990)
6. Furth, S., Baumeister, J.: Semantification of Large Corpora of Technical Documentation. IGI Global (2016), <http://www.igi-global.com/book/enterprise-big-data-engineering-analytics/145468>
7. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: *Proceedings of the 20th international joint conference on artificial intelligence*. vol. 6, p. 12 (2007)
8. Guha, R., McCool, R., Miller, E.: Semantic search. In: *Proceedings of the 12th international conference on World Wide Web*. pp. 700–709. ACM (2003)
9. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S. (eds.): *OWL 2 Web Ontology Language: Primer*. W3C Recommendation (27 October 2009), available at <http://www.w3.org/TR/owl2-primer/>
10. Hutchins, W.J.: The concept of 'aboutness' in subject indexing. In: *Aslib Proceedings*. vol. 30, pp. 172–181. MCB UP Ltd (1978)
11. Hyvönen, E., Mäkelä, E.: Semantic autocompletion. In: *The Semantic Web—ASWC 2006*, pp. 739–751. Springer (2006)
12. Klyne, G., Carroll, J.J.: *Resource Description Framework (RDF): Concepts and Abstract Syntax* (Feb 2004), <http://www.w3.org/TR/rdf-concepts/>
13. Miles, A., Bechhofer, S.: *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation 18 August 2009. (2009), <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>
14. Milne, R., Nicol, C., Trave-Massuyès, L., Quevedo, J.: TIGER: Knowledge based gas turbine condition monitoring. *AI Communications* 9(3), 92–108 (1996)

15. Padma, T., Balasubramanie, P.: Knowledge based decision support system to assist work-related risk analysis in musculoskeletal disorder. *Knowledge-Based Systems* 22(1), 72–78 (2009)
16. Puppe, F., Buscher, G., Atzmueller, M., Huettig, M., Buscher, H.P.: Clinical experiences with a knowledge-based system in sonography (sonoconsult). In: Workshop on Current Aspects of Knowledge Management in Medicine (KMM05), Proceedings 3rd Conference Professional Knowledge Management - Experiences and Visions, Kaiserslautern, Germany (2005)
17. Reynar, J.C.: Statistical models for topic segmentation. In: Dale, R., Church, K.W. (eds.) *ACL. Association of Computer Linguistics* (1999), <http://dblp.uni-trier.de/db/conf/acl/acl1999.html#Reynar99>
18. Stephens, S.: The enterprise semantic web. In: *The Semantic Web*, pp. 17–37. Springer (2007)

Using key phrases as new queries in building relevance judgments automatically

Mireille Makary¹, Michael Oakes¹, and Fadi Yamout²

¹Research Group in Computational Linguistics, University of Wolverhampton, UK
{m.makary, Michael.oakes}@wlv.ac.uk

²Computer Science Department, Lebanese International University, Lebanon
fadi.yamout@liu.edu.lb

Abstract. We describe a new technique for building a relevance judgment list (qrels) for TREC test collections with no human intervention. For each TREC topic, a set of new queries is automatically generated from key phrases extracted from the top k documents retrieved from 12 different Terrier weighting models when the initial TREC topic is submitted. We assign a score to each key phrase based on its similarity to the original TREC topic. The key phrases with the highest scores become the new queries for a second search, this time using the Terrier BM25 weighting model. The union of the documents retrieved forms the automatically-build set of qrels.

Keywords: Evaluation, automatic qrels, key phrases, relevance judgments.

1 Introduction

We propose a new technique based on Efron's [1] work which used query aspects to automatically build a set of qrels. The qrels did not involve any human intervention but the query aspects created for each TREC topic were mostly created manually. To explain what an aspect is, consider TREC topic 402 that has "behavioral genetics" as its title. The same information need might be represented by different aspects such as "behavioral disorders" or "genetics addictions". Each manually derived aspect was considered as a query and the union of the top 100 documents retrieved for each topic was considered to be the set of "pseudo-qrels" or "aspect qrels". We generate these new query aspects automatically from key phrases extracted from documents and use them to generate a relevance judgment list.

2 Experiments

Following Efron, we use the TREC-8 and TREC-7 test collections. We start initially by submitting each TREC topic to 12 weighting models found in Terrier (BM25, DFR_BM25, LGD, In_expC2, In_expB2, IFB2, TFIDF, LemurTF_IDF, PL2, BB2, DLH13 and DLH) [3] as surrogates for different information retrieval systems. The top K (K=10) documents retrieved by all 12 weighting models are collected in a set

(S) because they have a high probability of being relevant to the topic. Next, we extract 25 keyphrases using KEA [2] from each document in (S) where each key phrase consists of 3-5 terms for TREC-8 and 2-3 terms for TREC-7. Values were determined empirically. We assign a score to each keyphrase depending on its similarity to the initial topic. We then select the key phrases with the highest scores for each topic ($s \geq 0.4$ for TREC-8 and $s \geq 0.33$ for TREC-7) and put them in a set Q. The key phrases in Q are submitted as queries to the BM25 weighting model and since we are using another query for the same topic, this leads to new relevant documents that were not retrieved in the initial topic submission. We combine the union of the documents retrieved by the key phrases in Q. These documents are considered to be the newly generated qrels for the initial topic. To compare with Efron, we used a subset of the TREC systems, the “automatic” runs. In TREC-8 there were 116 automatic runs and in TREC-7 there were 86. We computed the MAP values using the original qrels for the test collection and then the MAP values using the newly generated qrels. We ranked the systems and computed the correlation with the TREC rankings. As shown in table 1, for TREC-7 the newly generated qrels provide a better correlation than those generated from Efron’s aspects, while for TREC-8 they are similar. This is acceptable considering that there is no human intervention in our method.

Test Collection	Efron’s aspects qrels		Keyphrases generated qrels	
	Kendall’s tau	Spearman	Kendall’s tau	Spearman
TREC-7	0.867	0.974	0.914	0.986
TREC-8	0.77	0.92	0.762	0.912

Table 1: Kendall’s tau for TREC-7 and TREC-8 automatic runs for different techniques

3 Conclusion

In this paper, we automatically generated a set of qrels based on keyphrases extracted from documents retrieved from 12 Terrier models for a particular topic and we used them as new queries instead of formulating new ones manually. The union of the documents obtained after this process was proven to be better than the aspect qrels generated by Efron. Future work can include testing this method on non-English and non-TREC test collections to evaluate its performance for any test collection.

4 References

1. Efron M.: Using multiple query aspects to build test collections without human relevance judgements, ECIR 2009
2. Witten I.H., Paynter G.W., Frank E., Gutwin C. and Nevill-Manning C.G. (2000) "KEA: Practical automatic keyphrase extraction." Working Paper 00/5, Department of Computer Science, the University of Waikato.
3. Ounis I., Amati G., Plachouras V., He B., Macdonald C. and Johnson D. Terrier Information Retrieval Platform. In Proceedings of the 27th European Conference on Information Retrieval (ECIR 05).

CAPLAN: An Accessible, Flexible and Scalable Semantification Architecture (Project Description)

Sebastian Furth¹, Volker Belli¹, Alexander Legler¹,
Albrecht Striffler¹, and Joachim Baumeister^{1,2}

¹denkbare GmbH, Friedrich-Bergius-Ring 15, 97076 Würzburg, Germany

²University of Würzburg, Institute of Computer Science,
Am Hubland, 97074 Würzburg, Germany

{sebastian.furth,volker.belli,joachim.baumeister}@denkbare.com

Abstract. The popularity of semantic information systems requires more data to be semantically prepared. However, the subsequent semantification process is still reserved for experts in Natural Language Processing. In this paper we define requirements for a state-of-the-art semantification architecture. Additionally we present a concept for a new semantification architecture meeting these requirements. Key strengths of the presented concepts are accessibility for non-experts, scalability and flexibility.

Keywords: Semantification, Information Management Architecture, Knowledge Management

1 Introduction

Semantic Search [7] emerged as the new system paradigm for enterprise information systems. In contrast to traditional information systems Semantic Search exploits ontologies during the retrieval process. The search performance usually outperforms traditional text based retrieval engines. However the underlying semantic search engines require resources to be semantically prepared. The semantic preparation [5] / semantification process of such resources typically comprises the partition of resources to reasonable segments, so called information units, and the subsequent semantic annotation with concepts from an ontology. The process is typically realized as a sequence of process steps.

The popularity of Semantic Information Systems leads to an increased need for migrating existing resources to semantic representations. However, existing implementations of the semantification process typically require a decent amount of knowledge in Text Analytics / Natural Language Processing and are thus hardly accessible for non-expert users. Additionally, implementations usually lack scalability and are thus not well prepared for processing large amounts of data. In most cases they are also inflexible with respect to the underlying data model and are thus hardly customizable to specific project needs.

In this paper we present a concept for a novel semantification architecture that is part of the ongoing research project APOSTL. The architecture is powered by a flexible state-of-the-art data model that is well prepared for the usage in scalable high performance environments. The easy management of project resources, import of existing data as well as assessment and review components open the semantification process for non-experts.

The remainder of the paper is structured as follows: In Section 2 we first describe requirements for a state-of-the-art semantification architecture. In Section 3 we explain required components and give some remarks to future implementations. Related work is briefly considered in Section 4. We conclude in Section 5.

2 Requirements

The overall requirements to the architecture are accessibility for non-expert users, scalability to large-scale data sets and flexibility for new project requirements. In the following we break down these requirements.

2.1 Accessibility

The increasing amount of semantification projects requires that semantification processes are accessible for non-experts (wrt. to Text Analytics/Natural Language Processing). This requires that the architecture is able to **hide the complexity** of underlying NLP processes. Users without expert knowledge in Natural Language Processing should be able to configure the semantification process on an abstract level, without having to know specific details of underlying approaches.

The opening of the semantification process to non-expert users requires that the architecture provides **documentation** for each of the underlying process steps. The documentation for each process step has to state clearly what data in which format is required as input and which results can then be derived from this data as output.

The generated data should be provided with **provenance and versioning information** that states clearly how (which method and parametrization) and when the data has been produced. The availability of such information facilitates the reproducibility of results and the comparison of parameter configurations.

The architecture should also provide ways to **examine generated results** on a high level. Therefore, the data visualization techniques should be a vital element in the architecture to open the assessment of results to a wide user range. Additionally, interactive review tools should allow the users to easily correct generated results.

Another aspect of accessibility affects the representation of the underlying data. Due to their subsequent usage in semantic applications all (intermediate) results should have a **semantic representation**, i.e. all data elements should at least be identifiable using a URI and provide type information.

2.2 Scalability

Scalability has a two-fold meaning in the context of semantification architectures. It is primarily concerned with the support of **large scale data processing** (Big Data), i.e. the architecture should be prepared to be employed in high performance environments for high throughputs. This requires that underlying algorithms are available for Big Data processing frameworks like Apache Spark [10] and the underlying data model supports distributed data storages like Hadoop's HDFS [11].

However, scalability in this context is also concerned with the aspect that a wide range of users should be able to use the semantification architecture. Therefore, the architecture should be realized as **Business Process as a Service**. A business process as a service is typically realized as a cloud service. In the context of a semantification architecture this means that the whole semantification process is available as web application or API.

2.3 Flexibility

A semantification process typically comprises a series of complex operations that successively prepare a resource for the usage in a semantic information system. However, in some cases some of the operations are not necessary, because data is already prepared to a certain extend (cf. Figure 1). Therefore, users should be able to enter the semantification process at an **arbitrary process step** if they can provide data in the necessary format.



Fig. 1. Maturity schema for documents in the semantification process.

Sometimes the semantification process must not necessarily be completed, e.g. because intermediate results are sufficient for specific application scenarios. Typical examples include specialized Information Extraction tasks that operate on semantically represented document structures. Hence, the architecture should allow to **query and export intermediate results**.

Although the process steps of semantification processes are usually similar in various application scenarios it might be necessary to parametrize, extend or adapt the process to new process requirements. Typical scenarios include the

existence of a previously unknown source format or new approaches/parameter configurations for specific process steps like segmentation, term matching or subject indexing. Thus, the architecture shall be **extensible**, such that new process steps or variants of existing process steps can easily be integrated. The extensibility should also be reflected in the data model.

3 Architecture

In the following we present an architecture facilitating the semantification of resources under the requirements stated in Section 2. Therefore, we first introduce the key components of the architecture and then close the section with some remarks regarding future implementations.

3.1 Components

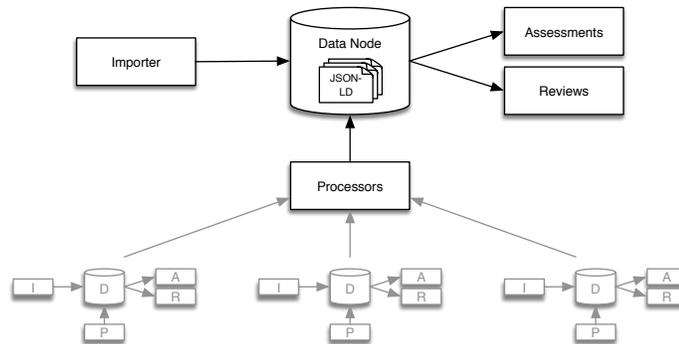


Fig. 2. Key components of semantification architecture.

Referring to the requirements in Section 2, a semantification architecture should be accessible, scalable and highly flexible. The flexibility mainly demands for a high extensibility and standardized import, export and processing functionality in all process steps while accessibility is concerned with hiding complexity from non-experts, providing easy-to-use assessment and reviewing functionalities and standardized data representations. Thus, we propose an architecture (see Figure 2) that is composed of interweaved modules, that are represented as quintuples $Q = \{D, I, P, A, R\}$, with:

- **Data Nodes D :** Contain the data and a data description for the process step, e.g. a description of document structures and instance data for concrete documents.

- **Importers** *I*: Provide and document import functionalities for data nodes, i.e. describe possible import formats and handle the import of data nodes from raw/source data. Also creates provenance information for the imported data.
- **Processors** *P*: Process data nodes in order to produce new or update existing data nodes respectively. Also creates provenance information for the generated/updated data.
- **Assessments** *A*: Provide possibilities/metrics/visualizations to assess a set of data nodes.
- **Reviews** *R*: Allow manually changing/reviewing existing data nodes.

All elements of the quintuple except the data nodes are optional. A semantification system can be built by combining multiple modules to a complete process, where each module encapsulates specialized functionality for a certain process step.

The interconnection of the encapsulated functionalities is realized through the data nodes. All data nodes are stored in a common schema-less data base (NoSQL) and are from there accessible from all modules. This way, the output of one module can be used as data source from another module which itself can produce new data nodes and so on. Additionally the usage of a schema-less NoSQL data base ensures the extensibility of a system, as new data can be stored without constraints.

The interconnection of modules in a semantification system is explained by the example of segmentation, term matching and subject indexing. Therefore, we assume that we have three modules encapsulating the aforementioned functionalities. Then the procedure is as follows:

1. **Segmentation**: A importer imports raw documents and stores them as data nodes (when appropriate using references to original sources).
2. **Segmentation**: A processor partitions the raw documents to segments and stores them as data nodes.
3. **Term Matching**: A term matching processor configured with a list of relevant terms scans the stored segment data nodes for term occurrences. Discovered occurrences are stored as new/complementary data nodes.
4. **Subject Indexing**: A subject indexing processor accesses the segment data nodes and the corresponding term match data nodes. Based on the information it determines topics for the segments and stores them as new/complementary data.
5. **Subject Indexing**: An assessment component visualizes the subject indexing result, e.g. highlights segments with many or few subject annotations.
6. **Subject Indexing**: Based on the assessment, the parametrization of step 4 may be revised and step 4 repeated. With stored provenance information multiple outcomes can be compared and the most appropriate one selected.
7. **Subject Indexing**: A review component allows to edit subject annotations, e.g. remove unnecessary or add missing subjects respectively.

The (intermediate) results, namely segments, term matches and annotated subjects can then be exported for subsequent usage in other systems.

3.2 Implementation Remarks

The implementation of the proposed architecture or rather the corresponding framework has not yet started. However, we have already defined some parameters specifying the subsequent implementation. These parameters affect the data model, the graphical user interface and the module mechanism.

Data Model The complete architecture builds upon a very flexible schema-less data model. The data model will be implemented as document-oriented NoSQL data base, where documents are the basic storage entity. We require JSON-LD [12] as storage format, which is standardized, light-weight, well-supported in common data base systems and allows to use explicit semantics. The availability of JSON-LD also allows to export (intermediate) results as standardized ontologies [8,14]. Furthermore, JSON(-LD) is compatible with common high performance data bases that work upon Apache Hadoop, e.g. MapR-DB. Importers I and processors P has to enhance the JSON-LD documents with provenance information from the PROV-O [13] ontology.

Module Mechanism The architecture is based upon the idea that a semantification system can be composed of modules that encapsulate specialized functionality. Besides a description of the data nodes (if appropriate as JSON-LD context), a module can define importers I , processors P , assessments A and reviews R . For the integration in the framework each of these components must provide specific information. Additionally, each component might define additional parameters that are necessary for configuration. Therefore, we plan to use a standardized plugin framework like OSGi [2].

Considering the scalability requirements modules should also report whether they are capable of running in high performance environments. Therefore, modules should express there high performance capability in their plugin definitions. If they claim to be high performance capable, we require them realize their functionality using a high performance computing framework like Apache Spark [10] or Apache Flink [1].

Graphical User Interface (GUI) and API As one requirement is a high accessibility for non-experts the framework will have a standardized graphical user interface. The graphical user interface shall guide users through existing semantification processes and allow for the creation of new/customized processes. Therefore, some components of the modules like importers or processors will be presented in a standardized way to allow the configuration by the user. Other components like assessments or reviews require a specialized user interface. Hence, these components must also provide user interface definitions as part of a module. The functionality that is accessible through the graphical user interface shall also be available as API to facilitate the process or module integration in other applications.

3.3 Requirement Tracing

In the following we give a brief requirement tracing, i.e. which requirement is realized by which component.

Accessibility

- **Hide Complexity:** Importers I and Processors P allow for the import and processing of data in a documented format.
- **Documentation:** Importers I provide documentation of importable data formats.
- **Provenance and Versioning:** Provenance and Versioning information are stored along with the data nodes in the common data base.
- **Examine Results:** Assessments A and Reviews R allow for the easy evaluation and review of results.
- **Semantic Representation:** All (intermediate) results are stored as JSON-LD documents with an explicit semantic.

Scalability

- **Large Scale Data Processing:** Module functionality can be implemented using high performance computing frameworks.
- **Business Process as a Service:** The framework will provide a standardized graphical user interface and an API.

Flexibility

- **Enter process at arbitrary steps:** Each module can have importers that allow the direct import of the required data.
- **Export (intermediate) results:** The results of each processing step can be exported as standardized ontology.
- **Extensibility:** The architecture allows for the easy extension through a module mechanism that will be realized using a plugin framework.

4 Related Work

To the best of our knowledge we are not aware of a framework that meets the requirements stated in Section 2 for a accessible, flexible and scalable semantification architecture. However, there are extensible frameworks for Natural Language Processing/Text Analytics tasks. Prominent examples are Apache UIMA [9] or GATE [4]. However, they usually need expert knowledge to be employed and come with a couple of shortcomings, cf. Bank et al. [3] for details. The idea of building specialized applications from standardized modules is not new, cf. for example Gu et al. [6].

5 Conclusion

In this paper we described early work from the ongoing research project CAPLAN. We presented requirements for a state-of-the-art semantification architecture. The requirements can be summarized with accessibility, scalability and flexibility. We then presented a novel semantification architecture that is composed of specialized modules that are interconnected through a very flexible and standardized data model based on JSON-LD. We showed that our architecture meets all the requirements and briefly named existing alternatives and their shortcomings.

Future directions include a further refinement of the presented architecture. Subsequently the concept will be realized in a prototypical implementation. The implementation will comprise the framework as well as sample modules for specialized semantification use cases.

Acknowledgments

The work described in this paper is supported by the Bundesministerium für Wirtschaft und Energie (BMWi) under the grant ZIM ZF4172701 "APOSTL - Accessible Performant Ontology Supported Text Learning".

References

1. Alexandrov, A., Bergmann, R., Ewen, S., Freytag, J.C., Hueske, F., Heise, A., Kao, O., Leich, M., Leser, U., Markl, V., et al.: The stratosphere platform for big data analytics. *The VLDB Journal* 23(6), 939–964 (2014)
2. Alliance, O.: Osgi Service Platform, Release 3. IOS Press, Inc. (2003)
3. Bank, M., Schierle, M.: A survey of text mining architectures and the uima standard. In: *LREC*. pp. 3479–3486 (2012)
4. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: an Architecture for Development of Robust HLT Applications. In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)* (2002)
5. Furth, S., Baumeister, J.: Semantification of Large Corpora of Technical Documentation. IGI Global (2016), <http://www.igi-global.com/book/enterprise-big-data-engineering-analytics/145468>
6. Gu, T., Pung, H.K., Zhang, D.Q.: Toward an osgi-based infrastructure for context-aware applications. *IEEE Pervasive Computing* 3(4), 66–74 (2004)
7. Guha, R., McCool, R., Miller, E.: Semantic search. In: *Proceedings of the 12th international conference on World Wide Web*. pp. 700–709. ACM (2003)
8. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S. (eds.): *OWL 2 Web Ontology Language: Primer*. W3C Recommendation (27 October 2009), available at <http://www.w3.org/TR/owl2-primer/>
9. Lally, A., Verspoor, K., Nyberg, E.: *Unstructured Information Management Architecture (UIMA) Version 1.0* (March 2009), <http://docs.oasis-open.org/uima/v1.0/uima-v1.0.html>
10. Meng, X., Bradley, J., Yuvaz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., et al.: Mllib: Machine learning in apache spark. *JMLR* 17(34), 1–7 (2016)

11. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The hadoop distributed file system. In: 2010 IEEE 26th symposium on mass storage systems and technologies (MSST). pp. 1–10. IEEE (2010)
12. Sporny, M., Kellogg, G., Lanthaler, M., Group, W.R.W., et al.: Json-ld 1.0: a json-based serialization for linked data. W3C Recommendation 16 (2014)
13. W3C: PROV-O: The PROV Ontology: <http://www.w3.org/TR/prov-o> (April 2013)
14. W3C: RDF Schema 1.1 – W3C Recommendation. <http://www.w3.org/TR/rdf-schema> (February 2014)

Topical Video-On-Demand Recommendations based on Event Detection

Tobias Dörsch, Andreas Lommatzsch, and Christian Rakow

DAI-Labor, TU Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany
{Tobias.Doersch,Andreas.Lommatzsch,Christian.Rakow}@dai-labor.de

Abstract. Recommender systems help users to discover relevant items. Traditionally, recommender systems rely on both detailed knowledge of the domain and an extensive user profile. However, small numbers of users, privacy concerns, or a very specific domain limit access or availability to this information. In this work, we present an approach for recommending items based on events relevant to the target group of our system. We exemplify the approach with the aid of a Video-On-Demand platform specialized in independent and art-house movies. Our recommender analyzes domain-specific blogs and news. It extracts current events that can be used for triggering topical recommendations. We show that our approach successfully identifies relevant events and provides highly relevant results without requiring detailed user profiles.

Keywords: recommender, event detection, privacy preserving recommender, Linked Open Data, video on demand

1 Introduction

The rapidly growing amount of items in online shops and entertainment services make it very hard for users to find relevant items. Recommender systems have been developed for supporting users to discover items potentially unknown and matching the user preferences. A widely used approach is user-based collaborative filtering computing the similarity between users and suggesting items that users with similar interests liked. A weakness of collaborative filtering is that current trends and temporal aspects are not taken into account. In many scenarios the context and seasonality have a high influence on the user preferences. Traditionally, experts (e.g. “curators”) compile topical recommendations in video shops or libraries taking into account new releases, trends as well as current events. This motivates us to develop a recommender that scans several different news streams, detects relevant events, and uses this information for computing recommendations.

Video-on-Demand (VoD) systems allow users to watch almost any movie at any time. The challenge for a VoD recommender system is not only identifying items

matching the user preferences but also computing when to recommend an item. In the past, curators knowing the typical seasonal user preferences and the relevant events (awards ceremonies, holidays, etc.) created a schedule when to broadcast a movie. We bring this principle to the VoD recommender. The recommender determines events and trends relevant to a specific target group. Based on these events we compute topical recommendations, which can be weighted by individual preferences. The event-based recommendations are often helpful for escaping the filter bubble and for suggesting items related to current trends.

We develop a recommender system for a VoD service focused on independent and art-house movies. In contrast to main stream VoD services, our portal does not offer blockbuster movies but a carefully selected catalog of films tailored to the needs of a niche market. A remarkable fraction of the offered films are documentaries and films related to current political topics. The requirements in the scenario are providing new relevant recommendations every day without relying on user profiles. We build our system on the idea that recognizing events relevant to our target group is a valuable basis for recommending relevant movies.

The identification of events suitable for recommending items leads to several challenges. To extract events, suitable sources must be identified and appropriate ways of processing and storing the contained information must be developed. This task requires learning algorithms able to identify events in streams of news data suitable for recommending items (films). Dependent from the different events types, adequate methods are needed for the event recognition and for linking events and films. In addition, explanations for the suggested items should be provided for improving the trust in the suggestions since recommendations based on news events are still unfamiliar for most users.

The remaining paper is structured as follows. Sec. 2 summarizes related work and discusses the connection to relevant research domains. Our approach is presented in Sec. 3. In Sec. 4 we evaluate our approach and discuss the strengths and weaknesses of our approach. Finally, a conclusion and an outlook to future work are given in Sec. 5.

2 Related Work

The task of recommending films on a daily basis is related to different domains.

CF-based Recommender Most movie recommender system focus on collaborative filtering (CF) [4]. CF-based approaches analyze the ratings users assign to items. The predictions are calculated by computing the similarity between either users (“user-based CF”) or the similarity between items (“item-based CF”). A requirement for getting high-quality recommendations is that a sufficient number of ratings for every user and every item are available. Well-known problems of CF-based approaches are the popularity bias [6] and the cold start problem [1]. CF-based algorithms tend to suggest popular, often already known items.

Context and Event Detection Beside the individual user preferences several different aspects influence the perceived relevance of movies, e.g. seasonality or the relation

to events. Studies analyzing the messages in social networks show that holidays and recent events have a high impact on the discussed topics [2]. The detection of events and the aggregation of messages related to the events are research topics in the analysis of social networks and news streams. Hennig et al. [3] applied clustering algorithms to news streams for identifying events in the news. The focus of the work lies on extracting and tracking topics but not on recommending items. Macedo et al. [5] developed a system that recommends social events. Based on the analysis of the user's past behavior, the proposed system recommends events based on social distance, and both location and time preferences.

Discussion Contexts and events have a high impact on the interest of users. Hence, building recommender systems computing recommendations based on relevant events is a promising approach helping users to escape the filter bubble and to find items related to the current topics of interest.

3 Approach

We develop a recommender system implementing a 4-layered architecture. The first layer collects news from heterogeneous sources. The second layer aggregates the collected data and extracts potentially relevant events. In addition, semantic data collections are integrated in order to consider expert-defined events (such as birthdays or memorial days). The third layer computes recommendations based on the events relevant to the target group. In the 4th layer, the recommendations are enriched and optimized for presentation. Explanations are generated for improving the trust in the relevance of the recommendations. The architecture of the system is visualized in Fig. 1. In the next paragraphs, we explain the implemented components in detail.

3.1 Collecting Data for Detecting Events

The crawlers continuously collect data being the basis for the identification of events. In order to focus on the events relevant to the target group, we carefully select the sources. In our scenario, we are especially interested in the domains art house, festivals, and documentaries. We analyze the RSS feeds of portals reporting on the domains. In addition, we crawl the TWITTER messages of an expert-defined set of accounts (using the TWITTER streaming API). In addition, we collect tweets from the major news portals for tracking the most relevant topics in the domain of politics. The selection of sources grants us access to up-to-date knowledge from domain experts. These experts typically write about the most relevant events and current trends. In our system we monitor ≈ 800 TWITTER accounts and ≈ 15 RSS feeds.

3.2 Recognition of Relevant Events

We consider two types of events. "Static" events such as birthdays, anniversaries, and memorial days are imported from semantic data collections. "Dynamic" events such as won awards, politic events, or the death of a director are detected in the news streams.

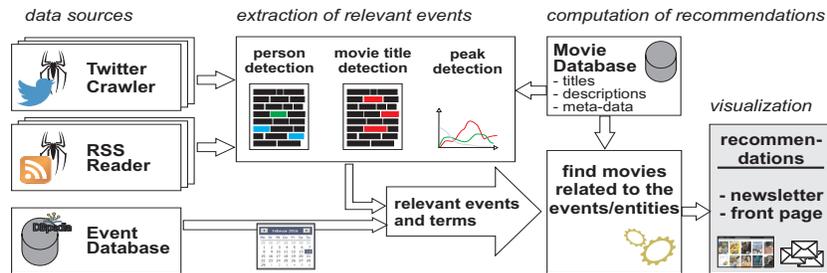


Fig. 1. The main components of our recommender system.

Knowledge Source for Events The “static” events are separated into two groups. The first group is formed by person- and movie-related events. The second group is built by holidays and memorial days typically related to specific keywords and genres.

Relevant Persons: Based on the movie catalog we know the persons related to the potentially relevant movies. We link the persons with DBPEDIA in order to collect all available birthdays of persons related to the movies. The same procedure is done for movie release days and awards won by the movie. The challenge in the task is the ambiguity of names and titles. We address this issue by computing a matching score taking into account context data. We only connect persons with DBPEDIA if the confidence score is above a threshold in order to prevent false positive matches. The score is calculated using the attributes of the entities (occupation, age, synopsis). User feedback is incorporated in order to correct and extend the automatically created links.

Relevant Holidays: In contrast to persons directly listed in the meta-data describing movies, the relations between holidays and movies are computed based on the textual description of the holidays. For this purpose we search the name of the holiday in the movie description and compute the textual similarity between the descriptions of the holidays (retrieved from DBPEDIA) and the synopsis of the movie. If the relatedness is above a threshold (optimized on a training dataset) the movie is linked to the holiday.

Discussion: Aggregating the different types of events, we find on average of about 20 events for each day of the year. This number of potentially relevant events allows us to filter out the most relevant events taking into account feedback from users and experts. In addition, the number of potentially relevant events allows us to ensure the diversity of events (e.g. with respect to actors, directors, composers as well as birthday, anniversaries). Static events are related for a specific day and typically recalled on the date of occurrence. However, some users may still be interested in the event a few days earlier or later (e.g. if they do not use the portal during the week). In order to make these recommendations available to those users, the relevance of static events degrades slowly over the course of 5 days.

Identifying Events in Tweets and RSS Feeds For recognizing events in news streams, we analyze how often a relevant person (listed in the movie catalog) is mentioned in

the news or tweets on a daily basis. An event is detected if a person is much more frequently mentioned than during an “average” day. Due to the large differences in popularity of movies and people, we implemented a 3-dimensional model: The popularity of a topic is identified by a long- and a short-term change in mentions as well as by the number of sources in that the topic is recognized. Since we do not compare topics against each other, each topic must fulfill criteria that are specific to its own time series. This leads to higher diversity in the recommended movies as well as to a broad spectrum of movies. In general, trend-detection for popular persons works more reliably than for unpopular persons. This is due to the fact that an increase of mentions of a popular topic is larger and thus easier to separate from noise.

Discussion The detection of the events and the linking of events with entities is the central component of the recommender. The recognition of static events is computed in advance when new movies are added to the catalog. The linking of dynamic events is done on a daily basis. The process is based on several text mining and similarity computations. A regular desktop computer suffices to complete the computations within minutes as the catalog is of limited size.

3.3 Computing Recommendations

Even though our database contains a large number of events, each type of event should optimally trigger its own set of recommended movies. On the other hand, some topics and events are very well connected in the movie database. In some cases, this leads to a number of recommended movies too large for a recommendation set. In other cases, too few movies are available to fill a type-specific set (i.e. recommendations based on birthdays). In those cases, our system mixes sets together based on the types’ similarity (e.g. birthdays and days of death) to achieve a suitable size for a single set. If we have too many candidates for a set (e.g. 20 birthdays on the same day), we lower the number of allowed movies per event or select the set of trigger events randomly for each user.

3.4 Presentation of results

In our VoD scenario, we present the recommendations computed based on the recognized topics on the front page of the VoD portal and in daily newsletters to registered users.

The topical recommendation of the front page: The landing page of the VoD service presents the sets of recommended movies. A header shows the type of event used for this set and creates the topical connection between each movie. To initially spike the users interest in a recommended movie, the trigger event is presented together with a short description of the event and, if available, context information on the related topic.

Theme-focused Newsletters: The VoD service already sends out a daily newsletter to registered users. This newsletter contains a set of 5 movies that have a deep topical connection. This connection is represented by a motto, for example, “Directors Inspired by Quentin Tarantino” or “Dream of a Better Life”. In order to build newsletter

automatically, we compute the most relevant event of the day and compute related movies. In the next step we try to fill templates created for the newsletters, such as “*Today is the birthday of <X>. His best movies here on <name of the portal>.*”

Discussion The implemented system is based on components and can be easily extended by integrating additional sources or by integrating components tailored to new types of events. The service interface allows the integration in existing recommender system.

4 Evaluation

We analyze the recommendations computed by our system. First, we evaluate the recommendations depending on the type of event used to trigger them. Secondly, we analyze the relevance of the recommendations and the acceptance of the suggested movies. The relevance of recommendations is analyzed based on the web server log of the VoD platform (currently only using content-based item-to-item recommendations) as well as feedback from experts (curators working for the VoD portal).

Recommendations based on News The system looks for trending topics by analyzing search log of the VoD system. Tab. 1 shows an overview of the number of events the recommender extracted based on trending topics for 9 days in January 2015. Analyzing detected events reveals that several trends in news feeds are correlated with “static” events. After the death of a popular actor, the time-series of mentions of that actor oftentimes spikes to an all-time high indicating that people are generally interested in this type of event. For other events, for example festivals and awards ceremonies, the number of mentions of the event increases during certain points in time. Awards ceremonies are often mentioned shortly after nominees for prizes are announced, during the award ceremony itself, and for a short period of time after the event, then oftentimes together winners of awards. Overall 20% of detected trends can be linked to “static” events. However, due to the large number of stored events, only 5% of events are covered by trends. Presenting trends together with a recommended movie proves to be difficult, if no knowledge is available on what event triggered the trend. Test users of our web application reacted positively to the most-popular approach, e.g. presenting the tweet with the highest “favorite” count.

Recommendations based on Static Events **Holidays and memorial days** days such as Veterans Day or Mother’s Day are linked to movies based on the similarity between the description of the holiday and the list of assigned term (retrieved from DBPEDIA) and the movie description. In our scenario, we considered 706 holidays related to at least one of the movies in our catalog. Analyzing the impact of these days on the user behavior, we observed a high variance for country-specific holidays: Displaying the trigger event together with recommended movie caused users to click 8 times more often. Confessional holidays increased clicks only by a factor of 1.7, while international holidays and memorial days increased clicks by a factor of 2.4.

Our **birthday** database contains 2,570 entries listing on average 7.02 birthdays per day. These events are related to 3,291 distinct movies. Tab. 1 shows the statistic of

relevant birthdays for the first days in January 2015. In order to evaluate the relevance of computed recommendations we check the recommendations against the spikes in the web server log. We found that 23% of recommendations derived from birthdays could be recognized by an increased movie related activity in the log file.

Table 1. The statistic of events recognized in news and semantic data.

date	number of relevant trends	number of related films	example person related to one of the trending topics	number of films related to the example person	number of relevant birthdays	number of related films	example person whose birthday this is	number of films related to this person
Jan 1, 2015	0	0	-	0	9	9	Snitz Edwards	1
Jan 2, 2015	2	2	Christian Bale	1	4	6	Lloyd Whitlock	2
Jan 3, 2015	3	4	Van Johnson	2	8	8	Thomas Morris	1
Jan 4, 2015	1	1	Forest Whitaker	1	3	6	August Diehl	4
Jan 5, 2015	6	8	Jessica Chastain	2	7	7	Shea Whigham	1
Jan 6, 2015	5	11	Ethan Hawke	4	13	14	Eddie Redmayne	2
Jan 7, 2015	4	4	Charlie Parker	1	5	6	Nicolas Cage	2
Jan 8, 2015	5	5	Marcus Vetter	1	6	7	Sarah Polley	2
Jan 9, 2015	8	13	Jodie Whittaker	3	11	43	Harun Farocki	32

Death-Days: Compare to birthdays, our database provides a significant smaller number of dates of death. The dataset contains 581 entries covering 286 days of the year. These events are related to 729 distinct movies. Similar to the birthday recommender, most death-days are not recognizable as peaks in the web server log. The death of a person (detected in the news stream) results in an increased user interest. The dates of death retrieved from DBPEDIA relate to dates several years in the past. This explains the different impact of death dates retrieved from the semantic database from death dates detected in the news.

Discussion We showed that our approach allows us to provide useful recommendations without having access to user profiles. The impact of the recommendations depends on the type of the identified event. In general, the relevance of dynamically recognized events (death of an actor, an award won by an actor) is more relevant than “static” events retrieved from a knowledge data base. “Big” birthdays of popular persons are more relevant than “usual” birthdays. Nevertheless “static” events are valuable since these events ensure that we reliably provide a fixed number of recommendations and diversify the result set.

Recommending movies based on events is often unexpected to users. In our discussions with users and the experts from the VoD portal we got positive feedback for the approach. In order to accept the recommendations it is important that users know or at least are interested in the events because the relevance of the events is crucial for the acceptance of the movie recommendations. On the other hand, our approach helps users to discover new content by recommending items based on events users usually would not be aware of.

5 Conclusions and Future Work

In this paper we present our system providing topical film recommendations based on different types of events. We discussed how to detect potentially relevant events from news and social media streams as well as the integration of semantic knowledge sources. In our analysis we found that birthdays of artists have only a very small influence on the user behavior. Events detected in news streams are better suited for recommending movies.

In contrast to traditional CF-based approaches, the developed approach helps users to discover new films. The relevance of suggestions is based on the similarity with current events instead of the similarity with entries in the user profile. We currently work on two personalization approaches. We combine the relevance scores computed using collaborative filtering ensuring that the identified events are matching the individual user preferences. In addition, we plan to allow users to add own sources (RSS feeds). This ensures that the news streams providing the basis for recognizing the relevant event meet the user needs.

Furthermore, we work on combining Named Entity Recognition algorithms and similarity computations based on semantic graphs for detecting movies related to current news. The weighted aggregation of several different relevance measures ensures a higher significance of recommended movies and provides the basis for more detailed explanations. The presented concept for recommending movies can be easily adapted for many additional scenarios, such as online shops. The use of the recent news (or weather data) is a promising new paradigm providing relevant recommendations without requiring detailed (sensitive) user profiles. A careful selection of sources analyzed for detecting events ensures that the recommendations are relevant for specific target groups. Based on the feedback we received for the implemented prototype there is a high potential in this approach.

Acknowledgments The work has been partially done in the EEGoF project supported by the German Federal Ministry for Economic Affairs and Energy. The research leading to these results was performed in the CrowdRec project, which has received funding from the EU 7th Framework Programme FP7/2007-2013 under grant agreement No. 610594.

References

1. J. Bobadilla, F. Ortega, A. Hernando, and J. Bernal. A collaborative filtering approach to mitigate the new user cold start problem. *Know.-Based Syst.*, 26:225–238, Feb. 2012.
2. W. Gao and F. Sebastiani. From classification to quantification in tweet sentiment analysis. *Social Network Analysis and Mining*, 6(1):1–22, 2016.
3. L. Hennig, D. Ploch, D. Prawdzik, B. Armbruster, H. Düwiger, E. W. De Luca, and S. Albayrak. SPIGA - multilingual news aggregator. *Procs. of GSCL 2011*, 2011.
4. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. (TOIS)*, 22(1):5–53, 2004.
5. A. Q. Macedo, L. B. Marinho, and R. L. Santos. Context-aware event recommendation in event-based social networks. In *Procs. of the 9th ACM RecSys Conf.*, NY, USA, 2015. ACM.
6. H. Steck. Item popularity and recommendation accuracy. In *Procs. of the 5th ACM Conf. on Recommender Systems*, pages 125–132, New York, NY, USA, 2011. ACM.

MapReduce Frameworks: Comparing Hadoop and HPCC

Work in Progress

Fabian Fier, Eva Höfer, Johann-Christoph Freytag

Humboldt-Universität zu Berlin, Institut für Informatik,
Unter den Linden 6, 10099 Berlin, Germany
{fabian.fier,eva.hoefer,freytag}@informatik.hu-berlin.de

Abstract. MapReduce and Hadoop are often used synonymously. For optimal runtime performance, Hadoop users have to consider various implementation details and configuration parameters. When conducting performance experiments with Hadoop on different algorithms, it is hard to choose a set of such implementation optimizations and configuration options which is fair to all algorithms. By fair we mean default configurations and automatic optimizations provided by the execution system which ideally do not require manual intervention. HPCC is a promising alternative open source implementation of MapReduce. We show that HPCC provides sensible default configuration values allowing for fairer experimental comparisons. On the other hand, we show that HPCC users still have to consider implementing optimizations known from Hadoop.

1 Introduction

In our research, we use MapReduce and its implementation Hadoop to experimentally compare the runtime of various scalable algorithms. Along these experiments, we identified practical issues that make a fair comparison and experimental reproducibility hard. Hadoop offers many configuration parameters that influence the runtimes of programs such as the number of Reducers or whether data compression is to be used between Map and Reduce. Furthermore, there are numerous possible optimizations in Hadoop programs, such as custom datatypes and very efficient byte-wise comparators. It is possible to “tweak” every implementation with a certain set of options and implementation optimizations. The same set of options and optimizations can lead to poor execution times for the implementation of different algorithms [1]. The literature barely discusses these configuration parameters and implementation details, although they are crucial for the validity of experimental results.

Another promising open source MapReduce implementation is HPCC (High Performance Computing Cluster) from LexisNexis [2]. Unlike Hadoop, HPCC hides many configuration details from the user. We are interested in how HPCC might replace Hadoop in our context and whether it allows for a fairer comparison when implementing MapReduce algorithms in it. As a running example,

we describe a common MapReduce-based textual similarity join algorithm. We introduce our implementation of this algorithm in Hadoop and its pitfalls concerning system configuration and code details. We compare this implementation to our corresponding implementation in HPCC and discuss our findings.

The paper is structured as follows. Section 2 describes HPCC and ECL. Section 3 contains the textual similarity join problem and the MapReduce-based approach we use as a running example. Section 4 introduces and compares our implementations of the algorithm in Hadoop and HPCC. The last section sums up our findings and gives an outlook on future research on this topic.

2 HPCC and ECL

HPCC is an open source parallel distributed system for compute- and data-intensive computations [2]. It contains a distributed file system. The main user interface of HPCC is ECL (Enterprise Control Language). ECL follows a dataflow-oriented programming paradigm and has declarative components. The following example code¹ computes a word count:

Define record structure “WordLayout” consisting of string “word”:

```
WordLayout := RECORD
  STRING word;
END;
```

Read given dataset into the variable “wordsDS”, apply WordLayout:

```
wordsDS := DATASET(['HPCC'], .., ['ANALYTICS'] ], WordLayout);
```

Define record structure “WordCountLayout” consisting of “word” and “count”. Note the count is defined by COUNT(GROUP) which implies that this record structure is to be applied to a grouped dataset:

```
WordCountLayout := RECORD
  wordsDS.word;
  wordCount := COUNT(GROUP);
END;
```

Apply WordCountLayout on dataset wordsDS and group by “word”:

```
wordCountTable := TABLE(wordsDS, WordCountLayout, word);
```

ECL allows to incorporate user-defined first-order functions written in C++ or Java [2]. These functions can be called from ECL functions. The semantics of the Map operator is represented in the ECL function PROJECT. It applies a user-defined function to each record in a given dataset. Reduce semantics can be emulated by partitioning and distributing the data with DISTRIBUTE, sorting it locally with SORT, and running a user-defined function on each group with ROLLUP. Although HPCC is not originally designed for the MapReduce programming paradigm, it is straightforward to adapt a MapReduce program to ECL. Thus, we regard HPCC as an alternative implementation of Hadoop.

¹ Adapted from <https://aws.hpccsystems.com/aws/code-samples/>

3 Textual Similarity Join

This section describes the textual similarity join problem and outlines the algorithmic approach from Vernica et al. [3] to compute it. We subsequently use this algorithm as an example to compare Hadoop to HPCC.

The textual all-pairs similarity join is a common operation that detects similar pairs of objects. Objects can either be strings, sets, or multisets. The similarity is defined by similarity functions such as Cosine or Jaccard similarity. Applications of this join are near-duplicate removal, document clustering, or plagiarism detection. Without loss of generality, we assume a self-join on sets.

Definition 1 (Similarity Join). *Given a collection of sets S , a similarity function sim , and a user-defined threshold δ , a similarity join finds all pairs with a similarity above the threshold: $\{\langle s_1, s_2 \rangle \mid sim(s_1, s_2) \geq \delta, s_1 \in S, s_2 \in S, s_1 \neq s_2\}$.*

A naive approach of computing this join is to compare each possible pair of objects. Due to its quadratic complexity, it is not feasible even for small datasets. More advanced approaches use a filter-and-verification framework. The framework consists of two steps. The first step computes candidate pairs, which are a superset of the result set. Due to the use of filters, the candidate set is much smaller than the cross product (assuming that a majority of pairs of objects of S is not similar). The second step computes the actual similarity for each candidate pair to verify if its similarity is above δ .

A prominent filter-and-verification MapReduce-based algorithm for set similarity joins is the VernicaJoin [3]. The main filtering idea is to only compare short prefixes of two objects to generate candidate pairs. Given a similarity function, a threshold δ and an object length $|s|$, we can compute a prefix length. It can be shown that two objects can only be similar if they have an overlap of at least 1 in their prefixes. One optimization is to sort the words in the objects by their global frequency in ascending order. This assures that the prefixes only contain the least frequent words which reduces the number of candidate pairs.

For our experimental comparison of similarity join algorithms, we adapted VernicaJoin to use already integer-tokenized input (instead of raw string input) and to output ID pairs of similar objects (instead of string pairs). These changes enable us to compare this algorithm to others. In the following, we describe our implementations of this adapted algorithm.

4 Comparison of Implementations

In this section, we introduce our implementations of the previously described algorithm in Hadoop and HPCC. We discuss the most runtime-relevant details concerning implementation and configuration. Due to space restrictions, we refer to the upcoming full version of this paper for experimental results.

The implementations consist of three steps. In the first step, we compute the global token frequency. In the second step, we sort the tokens in each object by this frequency and replicate each object for each token in its prefix. We group

all objects by their prefix tokens and verify for each pair in this group if it meets the threshold δ . The third step removes duplicates.

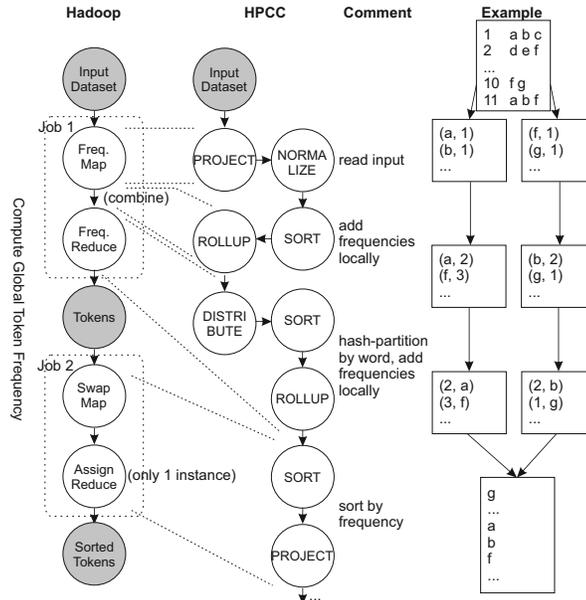


Fig. 1. Hadoop and HPCC Dataflows (First Step).

Figure 1 shows the dataflows in Hadoop and HPCC for the first step. Consider the example in the right column of the figure. The example assumes a parallelization degree of 2. The input has the form $\langle recordId, \langle tokenId, \dots, N \rangle \rangle$. For each token in each object, we create a new record $\langle tokenId, 1 \rangle$. The 1 represents the initial count of the token. In the following step, we add the token count for each partition. In the last two steps, we order the tokens according to their frequency.

Note that we use a *Combine* in Hadoop Job 1, which computes the token counts locally on the Map side. This significantly reduces network traffic and shuffle costs at the the Reduce side. We apply the same idea in HPCC by using ROLLUP and SORT in local mode. As in Hadoop, it is important to apply this concept. The ECL compiler does not automatically insert such an optimization.

In Hadoop, the *number of Map instances* is dependant on the HDFS block size by default. The default block size is either 64 or 128 MB. If the number of data blocks is smaller than the number of available Map instances, only a subset of the available Map instances is used by default. If in addition the first-order function is compute-intensive, this can lead to a longer runtime compared to executing Map on all available Map instances. This issue might be solved for

example by manually changing the input split size parameter of Hadoop. The distributed file system of HPCC splits the data at object borders and evenly distributes it amongst the available compute nodes. If the subsequent operator can operate on independent data chunks, it is executed on each data split.

In Hadoop, we manually set the *number of Reduce instances*. The default number is 1. If it is set too low, the computing nodes are under-utilized. If it is set too high, resources like main memory or network get overloaded. An optimal value is usually application- or even data-dependant. HPCC handles this parallelization implicitly.

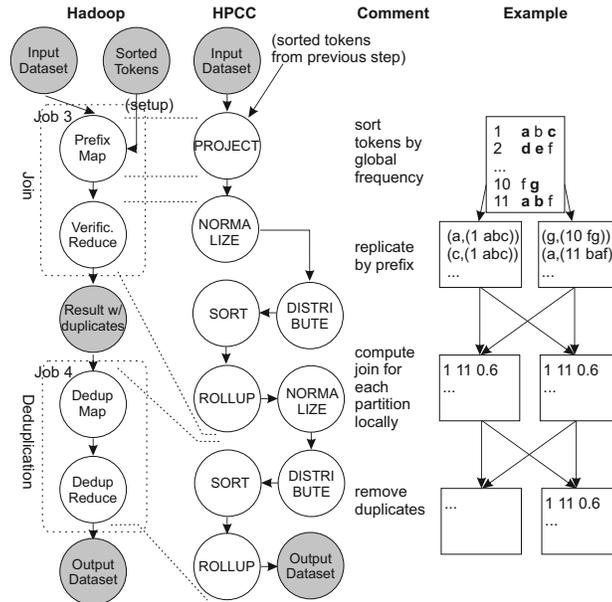


Fig. 2. Hadoop and HPCC Dataflows (Second and Third Step).

Figure 2 shows the dataflows for the second and third step of the implementations. Consider the data example in the right column. We read the input, sort the tokens by their frequency in ascending order, and replicate each object for each token in the prefix. We illustrate the prefix with bold numbers in the input. Since this replication can be computed on independent data partitions, we use two boxes for the resulting partitions. We group all records with the same tokens and compute their pairwise similarity. If two records share more than one common token in the prefix, the similarity of this pair is computed more than once. In the last step, we deduplicate the result.

The Map in Hadoop Job 3 uses a setup function, which initially reads the word frequencies from the first step. It sorts the words in each object according

to their global frequency and computes the prefix length. For each word in the prefix, it outputs the key-value pair $\langle\langle word, objectLength \rangle, object \rangle$. Note that the key consists of two integers, the word and the length. As proposed by Vernica et al. [3], we use a combined key consisting of the word and the length of the containing record. The word in the key partitions the data. The length is used additionally for local sort on the Reduce side. The Reducer retrieves the objects ordered by length. Since it performs a local nested-loop, we can *prune locally buffered objects* which cannot be similar anymore to all subsequent objects due to their length difference. We implemented a custom partitioner, sorter and grouper for this. This approach has an impact on the runtime. If the number of locally buffered objects exceeds memory boundaries, the computation becomes slow. In ECL, we implement the same approach by sorting the records by length within each partition. For each partition, we run a user-defined Java function that computes the similarity join locally. As in Hadoop, the user-defined function is stateful and can cause memory overflow.

5 Summary, Future Work

We were interested in how HPCC might be an alternative to Hadoop as an execution platform to allow for a fairer comparison when implementing MapReduce algorithms. Using the VernicaJoin to implement a textual similarity join, we showed that a complex MapReduce algorithm can be adapted to HPCC in a straightforward way. HPCC takes away some configuration details from the user like the parallelization degree (number of Map and Reduce instances). However, ECL still requires its users to carefully partition data so that intermediate buffers do not get overloaded. It is also necessary to explicitly implement optimizations such as local Combines. We plan to investigate further the influence of memory configuration on runtime. Especially in Hadoop, it is usually not clear to the user how its memory-related parameters impact performance. Furthermore, we plan to adapt this textual similarity join approach to use even more native ECL functions rather than user-defined “black box” code. This opens optimization possibilities which can potentially be integrated into the HPCC system.

Acknowledgements. This work was supported by the Humboldt Elsevier Advanced Data and Text (HEADT) Center.

References

1. Babu, S.: Towards Automatic Optimization of MapReduce Programs. In Proceedings of the 1st ACM symposium on Cloud computing. ACM (2010)
2. Middleton, A. M., Bayliss, D. A., and Halliday, G.: ECL/HPCC: A Unified Approach to Big Data. In: Furth, B. and Escalante, A.: Handbook of Data Intensive Computing, pp. 59–107. Springer, New York (2011)
3. R. Vernica, M. J. Carey, and C. Li. Efficient parallel set-similarity joins using mapreduce. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pages 495506. ACM (2010).

A Clustering Approach for Holistic Link Discovery (Project overview)

Markus Nentwig, Anika Groß, and Erhard Rahm

Database Group, Department of Computer Science, University of Leipzig

Abstract. Pairwise link discovery approaches for the Web of Data do not scale to many sources thereby limiting the potential for data integration. We thus propose a holistic approach for linking many data sources based on a clustering of entities representing the same real-world object. Our clustering approach utilizes existing links and can deal with entities of different semantic types. The approach is able to identify errors in existing links and can find numerous additional links. An initial evaluation on real-world linked data shows the effectiveness of the proposed holistic entity matching.

1 Introduction

Linking entities between sources has been a major effort in recent years to support data integration in the so-called Web of Data. A large number of tools for semi-automatic link discovery has been developed to facilitate the generation of new links (mostly of type `owl:sameAs`) [5]. Repositories such as BioPortal [7] or LinkLion [6] collect numerous links for many sources to improve their availability and re-usability without having to repeatedly determine such links for new applications and use cases.

Despite the advances made, there are significant limitations in the achieved inter-linking of data sources and in the current approaches for link discovery. First, the degree of inter-linking is still low and automatically generated links are wrong in many cases [2]. Current approaches for link discovery only match two data sources at a time (pairwise linking) resulting in a poor scalability to many sources [8]. This is because the number of pairwise mappings increases quadratically with the number of sources, e.g., one would need almost 20,000 mappings to fully interconnect 200 sources.

Most of the current link discovery approaches process only two data sources at a time restricting scalability for many sources while few approaches natively support multiple data sources. In [3] the quality of joins on Linked Open Data is improved by determining highly connected entity groups in a set of given links using metrics such as edge betweenness. The joint entity matching approach in [1] aims at finding links between multiple data sources based on an iteratively adopted matrix of pairwise similarity values. Existing approaches to determine `owl:sameAs` links also focus on entities of the same type while many sources contain entities of different types (bibliographic datasets contain publication and author entities, geographical datasets contain numerous kinds of entities such as countries, lakes, etc.). Furthermore, existing links are hardly utilized when additional links need to be determined.

The need for holistic approaches to integrate many data sources has been outlined in [8] with the suggestion to use clustering-based approaches to link and fuse matching

entities for improved scalability. We are working on such clustering-based approaches for the Web of Data [4] and summarize the approach and initial evaluation results in this short project overview. The approach utilizes already existing links and supports the integration of entities of different semantic types. All matching entities from different sources are grouped into a single cluster thereby supporting a much more compact representation of match results than with binary links. Furthermore, the cluster-based approach facilitates the integration of additional sources and entities since they only need to be matched with the set of already existing clusters rather than adopting a pairwise linking with numerous different sources.

We consider a set of k data sources containing entities of different types. Each entity e is referenced by an URI and has a set of describing semantic properties (i.e., RDF vocabulary). Two entities of different sources can be connected by a `owl:sameAs` link if they were found to represent the same real-world object. All same-as links between two sources S_i and S_j ($1 \leq i, j \leq k$) constitute a binary equivalence mapping $M_{i,j} = \{(e_1, e_2, sim) | e_1 \in S_i, e_2 \in S_j, sim \in [0, 1], i \neq j\}$. Link discovery tools can assign a similarity value sim to indicate the strength of a connection with 1 denoting equality (highest similarity). For k data sources, there can be up to $\frac{k \cdot (k-1)}{2}$ such equivalence mappings. For holistic entity clustering, we use a set of existing mappings $\mathcal{M} = \bigcup_{i,j=1}^k M_{i,j}$ and the set of associated entities \mathcal{E} of the k data sources as input. The goal is to compute a set of n clusters $\mathcal{C} = \{c_1^r, \dots, c_n^r\}$ such that each cluster only includes matching entities (denoting the same real-world object) and that different clusters represent different entities. In this paper, we consider duplicate-free data sources, such that a cluster can contain at most k entities. For each cluster we determine a cluster *representative* r derived from the cluster entities to simplify the comparison between clusters.

The following Sec. 2 will describe and illustrate the workflow for the proposed holistic entity clustering. We then present preliminary evaluation results in Sec. 3 and conclude.

2 Holistic Clustering

Our holistic clustering approach utilizing existing links consists of four main steps: preprocessing, initial clustering based on connected components, cluster splitting and iterative cluster merging. We illustrate the approach in Fig. 1 for partially linked geographical entities from four data sources. Due to space restrictions, linked entities (and corresponding properties) are shortened to their IDs and clusters are represented by thick bordered boxes. While our algorithm is generic, it can be customized to specific domains by providing appropriate background knowledge, similarity functions and thresholds to determine relevant entities and clusters. For the considered geographical domain, the similarity function determines a combined similarity from the string (trigram) similarity on normalized labels, the similarity of the semantic entity type and the normalized geographical distance. The details of the workflow will be described in the rest of this section.

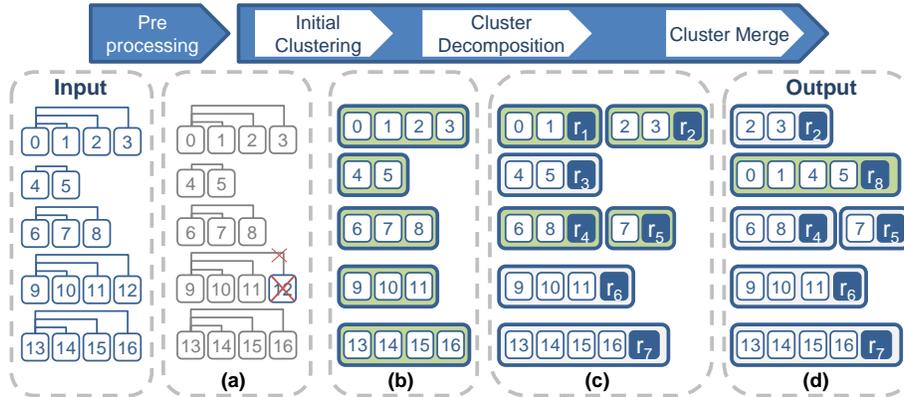


Fig. 1: Application of holistic clustering to running example.

2.1 Preprocessing

During preprocessing we normalize property values needed for the similarity computation, i.e., we simplify entity labels, harmonize information about the semantic types of entities and check that the input mappings do not violate the assumption of duplicate-free data sources.

Information about the semantic type of entities differs substantially between sources or may be missing. For instance, DBpedia uses *City* and *Town* whereas Freebase has a type *citytown* and other related types. To overcome such differences, we use background knowledge about the equivalence and comparability of entity types of different sources to harmonize the type information. We manually determined this type mapping for our geographical sources although it could be constructed with the help of ontology matching approaches. Based on the type mapping we simplified numerous types to more general ones, e.g., the types *city* or *suburb* are treated as type *Settlement*. After harmonizing the type information, we remove all links where the linked entities have incompatible types. Note that we do not exclude links to entities with missing type information.

With the assumption of duplicate-free data sources in place we check if all input mappings comply with the restriction. In Fig. 1, entities 11 and 12 come from the same source so that the links (9–11) and (9–12) violate the 1:1 assumption. In such cases, we only keep the best-matching link (9–11) and drop weaker links (9–12) as shown in Fig. 1a.

2.2 Initial Clustering

Using the preprocessed entities and mappings we first identify a set of initial clusters by computing all connected components as the transitive closure from the given links. Each resulting connected component builds an initial cluster C covering all entities that are directly or indirectly connected via a same-as link in \mathcal{M} . In our running example, we create five different clusters covering 2-4 entities (see Fig. 1 b).

2.3 Cluster Decomposition

The initially created clusters can contain entities that should actually be separated, e.g., due to wrong input links or because of an insufficiently high transitive similarity between entities. For this reason we decompose clusters (1) based on incompatible semantic types and (2) exclusion of entities based on intra-cluster similarity values. Finally, for each resulting cluster a cluster representative is created.

Type-based Grouping: While we eliminate links with incompatible semantic types during preprocessing, there can be entities without type information (e.g., entity (0)) leading to clusters with entities of different types during the initial clustering. We split such clusters into several smaller sub-clusters with entities of the same type. Entities without semantic types are then added to the sub-cluster of their most similar neighbor using computed similarities between cluster members. For the considered cluster of our example, we first build sub-clusters (2, 3) and the singleton cluster (1) of different types (e.g. *Settlement* vs. *BodyOfWater*). The untyped entity (0) is assigned to the cluster with the more similar (geographically closer) entity (1) resulting in sub-cluster (0, 1).

Similarity-based Refinement: We further split clusters based on the computed intra-cluster similarity between entities. For each entity, we determine the average similarity of its links to other cluster members and separate an entity if the average similarity is below a given threshold t_s . This process is executed iteratively as long as the average similarity of an entity is smaller than t_s . In the merge phase, such separated entities may be added to other more similar clusters. As shown in Fig. 1 c we separate entity (7) from the cluster (6, 7, 8) since the entity had a low label similarity to (6) and (8).

Cluster Representative: For each resulting cluster we create a cluster representative to (1) facilitate the computation of inter-cluster similarities in the merge step and (2) to efficiently match new entities, e.g., from additional data sources. We create the representative by combining the properties from all entities in a cluster and select a preferred value for each property, e.g., based on a majority consensus, the maximal length of labels or pre-determined source priorities (for geo-coordinates). We also keep track of the data sources represented in the cluster to avoid unnecessary merges for already covered data sources.

2.4 Cluster Merge

Lastly we merge similar clusters below the maximal possible cluster size k . Therefore we determine the similarity between clusters by applying the domain-specific similarity function on the cluster representatives. Given the typically large number of clusters, this is an expensive operation if we consider all pairs of clusters (quadratic complexity). We avoid unneeded comparisons by not considering pairs of clusters with incompatible types, overlapping data sources and clusters with $> k$ resulting elements. The cluster mapping \mathcal{CM} computed for the remaining cluster pairs is restricted to the most similar pairs of clusters with a similarity exceeding the merge similarity threshold t_m .

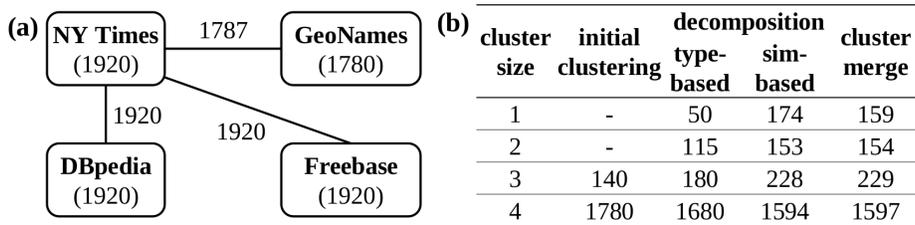


Fig. 2: (a) Data set structure: number of entities and links, (b) Cluster sizes in workflow phases.

Cluster merging is an iterative process that continues as long as there are merge candidates in \mathcal{CM} . In each iteration we select the pair of clusters (c_1, c_2) with the highest similarity from \mathcal{CM} , merge it into a new cluster c_m and compute a new representative for it. c_1 and c_2 are removed from \mathcal{C} and appropriate cluster pairs are removed from \mathcal{CM} . Furthermore, c_m is added to \mathcal{C} and \mathcal{CM} is extended by similar cluster pairs for the new cluster c_m obeying the restriction for new cluster pairs. The termination of the loop and the merge step is guaranteed since we reduce the number of clusters in each iteration. Applying the approach to our example leads to the merging of $(0, 1, r_1)$ and $(4, 5, r_3)$ into the new cluster $(0, 1, 4, 5, r_8)$ (see Fig. 1 c,d) due to a high similarity of all properties.

For the given example, we clustered entities from four different data sources thereby finding previously unknown links and eliminating wrong existing links for improved data quality. The six resulting clusters (Fig. 1 d) implicitly represent 17 pairwise entity links compared to 12 initially given links from which 3 turned out to be incorrect. In particular, we could now identify matches between previously unconnected sources.

3 Initial Evaluation

We evaluate our holistic clustering approach using the location subset of the OAEI 2011 Instance Matching benchmark with links of presumed high quality. Fig. 2 a shows the number of links between the four geographical data sources and the number of entities that are interconnected by these links. We retrieved additional entity properties via SPARQL endpoints or REST APIs in the respective sources in 2015. Still, the geo-coordinates were missing for 1009 entities (13.4%) and the type information even for 2525 entities (33.5%), including all entities from the NYTimes dataset. We use the similarity function described in Sec. 2; the similarity thresholds t_s, t_m are set to 0.7.

We first evaluate the resulting cluster sizes for the different phases of our holistic clustering approach applied to these datasets (Fig. 2 b). During the preprocessing (not shown in the Fig.), we already removed seven wrong NYT-GeoNames links based on the one-to-one cardinality restriction; the missing type information for NYT did not allow removal of type-incompatible links during preprocessing. The initial clustering results only in clusters of sizes 3 and 4 since each NYT entity is linked with an entity in Freebase and DBpedia. Applying the type-based grouping and similarity-based re-

finement results in a significant number of cluster splits and clusters of size 1 and 2 due to incompatible entity types and partially low intra-cluster similarity. During the merge phase some of the smaller clusters can be merged into larger ones leading to more clusters of sizes 3 and 4. In particular, 15 singleton clusters could be merged into clusters of size 2 and 3. Overall, the resulting clusters represent 9510 links with 4596 new links and 713 deleted links compared to the input link set. In particular, we could cluster many entities from the previously unconnected sources GeoNames, DBpedia and Freebase.

4 Conclusion

We proposed a new holistic approach for clustering-based link discovery for many data sources. The approach utilizes existing links and can match entities of different semantic types. The determined entity clusters facilitate the integration of more data sources without having to individually link them to each other data source. An initial evaluation for linked data from the geographical domain confirmed that the new approach holds great promise as it can identify wrong links and many additional links even between previously unconnected sources. In the future, we will evaluate the scalability and quality of our approach on larger datasets and more sources from different domains based on a parallel Hadoop-based implementation that is currently under development. We will also study specific aspects such as improving the quality of current mapping collections like BioPortal and the incremental extension of entity clusters when integrating new data sources.

References

1. Christoph Böhm, Gerard de Melo, Felix Naumann, and Gerhard Weikum. LINDA: Distributed Web-of-Data-Scale Entity Matching. In *Proc. of the 21st ACM CIKM*, pages 2104–2108. ACM, 2012.
2. Daniel Faria, Ernesto Jiménez-Ruiz, Catia Pesquita, Emanuel Santos, and Francisco M Couto. Towards annotating potential incoherences in BioPortal mappings. In *The Semantic Web—ISWC 2014*, pages 17–32. Springer, 2014.
3. Jan-Christoph Kalo, Silviu Homoceanu, Jewgeni Rose, and Wolf-Tilo Balke. Avoiding chinese whispers: Controlling end-to-end join quality in linked open data stores. In *ACM Web Science 2015*, 2015.
4. Markus Nentwig, Anika Groß, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. Holistic Entity Clustering for Linked Data. Technical report, 2016. submitted for publication.
5. Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. A Survey of Current Link Discovery Frameworks. *Semantic Web J.*, 2016.
6. Markus Nentwig, Tommaso Soru, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. LinkLion: A Link Repository for the Web of Data. In *ESWC 2014 Posters & Demo session*.
7. Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res*, 37:W170–W173, 2009.
8. Erhard Rahm. The Case for Holistic Data Integration. In *Proc. ADBIS*. Springer LNCS, 2016.

Query-driven Data Integration (Short Paper)

Peter K. Schwab, Andreas M. Wahl, Richard Lenz, and Klaus Meyer-Wegener

Friedrich-Alexander-Universität Erlangen-Nürnberg,
Technische Fakultät, Department Informatik,
Lehrstuhl für Informatik 6 (Datenmanagement),
Martensstr. 3, 91058 Erlangen, Germany
{peter.schwab, andreas.wahl, richard.lenz, klaus.meyer-wegener}@fau.de
<https://www6.cs.fau.de>

Abstract. The paper describes an ongoing project that pursues the idea of query-driven data integration. Instead of first creating a common global schema and fetching, transforming, and loading the data to be integrated, we start with the queries. They are taken as a specification of information need and thus as the overall purpose of integration. Two repositories are being developed, one for all information related to the queries and one for potential data sources to which those queries may refer. Queries may have very different forms, and thus there are many different ways how they can be used to make the integration effort more efficient.

Keywords: Database, Query, Data integration

1 Introduction

Data integration [5] is a task found in many enterprises, causing tremendous amounts of repeated work. Often, an additional data source is considered interesting and thus a complex process of data integration is started. A heavy-weight example is the merger of two companies. Usually, some kind of ETL (“extract – transform – load”) process is created with great effort, so that the data from one system can be made available in the other.

Since the effort is substantial and in some cases even prohibiting, other approaches like “pay as you go” have been proposed [4]. They postpone the integration effort to the time when the data are actually needed. However, they still focus on the (incremental) creation of a common global schema.

We continue this line by proposing to look at the queries first. This is our understanding of the term “query-driven”. It does not mean that we ignore the other information on the data that may already be available, but the queries are in the focus. A query is regarded as a specification of information need. Ideally, it is given in formal notation already, but we accept other forms as well. The query is written “as if” the new source(s) had already been integrated. So it is not yet executable. The purpose of data integration is then slightly modified to making that query executable and to do only the part of the integration process that is required to achieve that.

There are many variations and options involved in this scenario: the form of the query, the annotations or hints pointing to potential data sources, a repository of data sources found and used in the past, another repository of queries that have already been used for integration purposes, and many more. The purpose of this paper is to define some of these steps and to propose methods that could be used in them. It is assumed that the whole field is by far too large to be handled in just one project, let alone in one paper.

2 Scenarios

We have been asked to help in data-integration tasks many times. Our insistence to first name a few queries caused some confusion in the beginning, because even the users often had the idea to collect lots of data first and to only then think about possible accesses and analyses. It was not so difficult, however, to convince them that taking the queries into account would help to focus the development, and save time and resources.

One of these scenarios is the Aroma-Research Database. Here, we can show an example of a real “query”, see Fig. 1, that has been given to us in the very beginning. The elements of this data sheet are all coming from different data sources, some even from remote sites. The details are not important here. Excel is used so far to enter the data into the system, and the users would like to continue to do so. So tools extracting data from Excel tables and formatting them appropriately [7, 10] must be used. The knowledge of the queries allows to restrict the effort to the data that are known to be included in the result.

253 ER / FR CAS-Nr.:

Lösungen:

Nr.	Konz.	Erstellt am	Von	Standort
RS-253	2033,5 µg/ml	19.10.2010	FK	ER
RS-253b	1789 µg/ml	23.2.2012	JN	FR
ES-253	47540 µg/ml	3.1.2013	JS	ER
IS-039-A2	958,8 µg/ml	4.5.2012	FK	ER

Retention Indices:

capillary	RI	max.	min.	n
DB-5	1261	1274	1244	7
DB-FFAP	2029	2040	2009	4
DB-1701	1432	1444	1424	5

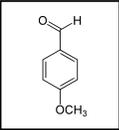
Quality and OT:

Quelle	Matrix	Qualität	Konz	OT
ER	Wasser	Waldmeister, süß, Marzipan	4753 µg/L	-
FR	Wasser	Sweet woodruff-like, sweet	3870 µg/L	-
Web	?	Sweet, powdery, vanilla, anisic, woody, coumarin and creamy with a spicy nuance	-	-
ER	Luft	Waldmeister	-	500 ng/L

Isotopic labelled standards: ²H₂-p-Anisaldehyde IS-039-A2 [Link zum Datenblatt](#)

Qualitäts-sicherung:

Kriterium	OK/Erg.	Anmerkung	Name, Dat.
Gehalt	>99%	Verunreinigung: o-Anisaldehyde (ca. 0,7%) Link zum Chromatogramm	CH, 2.2.2009
MS	ok	Abgleich mit NIST, Link zum Spektrum (MS-ES)	CH, 2.2.2009
Quant. IS-039-A2	928,4 µg/ml	Über Me-Octanoat, Link zur Auswertung	AL, 17.3.2012



[Link Safety Data Sheet](#)

Properties

Molecular formula: C₉H₁₀O₂

Molar mass: 136.15 g/mol

Density: 1.119 g/cm³ at 15 °C

Melting point: 0 °C

Boiling point: 248 °C

Purity: 98%

Supplier: Aldrich (Steinheim, Germany)

order nr.: [A88847](#)

delivery date: 28.01.2009

price: 43,10€

amount: 100 g

Fig. 1. A “query” given to us by the users of the Aroma-Research Database

A second scenario is currently called “Walhalla DB”. The purpose is to allow music researchers to view different aspects of an opera (we use Richard Wagner’s “Walküre” – Valkyrie – as an example) simultaneously. These are: a video of some performance, the libretto, and the music sheets. The idea is to allow for flexible navigation in any of the views, and then synchronizing the other views. Here, the queries are defined by the given user interface. It is still not fixed, but already defines a set of accesses to the underlying data sets, which may be calls of a database API as well as database queries. The relational database at the moment has only one table (which is actually the data-entry mode preferred by the music researchers) and will be re-designed in a proper way on the base of the given queries. Images of music sheets are stored separately. And the performance video is accessed via YouTube. The results of the design together with the final queries will be published in due course.

In co-operations with industrial partners, we often found the wish to associate own data with publicly available data, e. g. weather information or tweets. It is interesting how the weather may have influenced sales numbers, or how positive or negative tweets had such an effect. This strongly suggests to take the integration of linked open data [8, 1] into account, as it has already been demonstrated by the OPEN project [2].

So the demand for integration is obvious, and in all these scenarios we envision a benefit from looking at the queries early.

3 Query formulation

Apart from the world of formal query languages, there is also another world even more relevant to many users: result tables or reports, as in the scenario of the Aroma-Research Database. Users can easily sketch a form or report with the data they would like to combine, each field potentially coming from a different source. The data sources are often known, but provide only semi-structured formats (e. g. Excel, PDF, HDFS). They may be located in other departments or even enterprises.

Since experts must transform these specifications into more formal languages anyway, the world of query languages must be investigated, too. Here, queries may now include elements that are not yet found in the given schema as in DrillBeyond [6]. They may be accompanied with hints on the potential sources for these elements. This triggers a process of searching, downloading, transformation, and storage, which is quite usual. The difference now is that the query is already given, so the effort can be tailored to it.

We plan to look at different query languages here, but to get started, we will concentrate on SQL first.

4 Integration with Execution

The degree to which the queries identify data sources may vary. Sometimes the sources are well-known, as in the cases of the Aroma-Research Database and

the Walhalla DB. Sometimes there is the mere hope that appropriate open data sources can be found the World-wide Web. Then searching for data sources must be the first step. Since this may be a significant effort, we propose to keep the information on data sources that have been found and have proven useful in answering former queries. Knowing the query may help further, e. g. if it includes a join with local tables. In this case the join attribute—name as well as type, and values—can give hints on the characteristics of the data source being sought.

Once the source has been found, execution of the query may begin, doing integration on the fly. This can mean fetching data from the source and transforming them to the format required for further processing. It may as well mean to ship some sub-query to the source mostly to select and filter the data before fetching them. The latter requires that the source can execute these sub-queries—and of course the knowledge of the query in the first place. It is a commonplace that this can reduce the amount of data to be transmitted substantially.

In the end, the data coming from the source needs to be transformed to the format required. Again, we see a benefit in having the query at hand and not just creating some part of a common (global) schema, because we transform exactly those parts that will be included in the final processing or the query result.

As far as the query result contributes to the construction of a global schema, it may of course be used for that, too. This, however, is done *after* the execution, so that it does not postpone the delivery.

5 Repository

We are currently designing a repository for queries and any information related to them¹. SQL as a query language will be dominant in the beginning, but we try to remain open for other query languages and forms. Not surprisingly, the amount of information associated with a single query can be quite substantial. It includes:

- the query expression in text form,
- a query specification in form of example results,
- the abstract syntax tree of the query,
- the relevant parts of the query as separate features (relations used, attributes returned, selection predicates, grouping, etc.),
- a summary of the query result,
- frequency of query execution,
- importance of query execution²,
- reference to the software that invokes the query,
- execution cost or time of the query,
- the query execution plan in some system at hand,
- and many others.

¹ It is understood that this design takes the queries to be executed on the repository into account.

² Some kind of “emergency” query may be very rare, yet very important once executed.

The repository will be designed with evolution in mind, that is, beginning with a rather small subset of attributes and features, but with a potential to grow.

The other repository keeps track of the data sources. It makes sense to remember the data sources that have already been found as well as the experience with them. Their usefulness in at least some context should be saved. This includes a notion of trust and reliability. Also, the mapping and transformation of data elements from that source should be kept.

Both repositories will be linked. This allows to find the data sources that have previously been used by a query. Still, the user may decide to replace them by better sources found in the meantime. And it allows to find all the queries that have used a given data source in the past. Their owners may be informed about changes in that source, up to the fact that it may no longer be available.

6 Related Work

The literature on data integration is tremendous. The book by Doan et al. [5] may be considered the standard work now. Hardly any of the approaches, however, are query-driven. Most of them focus on the construction of a global schema; not necessarily before using the system, but at least before executing a query.

The work on incremental integration is also helpful when doing it query-driven. The Data Tamer system [12] (now a product named “Tamer”) includes many techniques for that. The already-mentioned systems OPEN and DrillBeyond [2, 6] are query-driven to some extent, and provide very useful mechanisms for the integration of external data sources, which can be adapted to work with our approach. Query-driven schema expansion as described in [11] is similar to our approach, but concentrates on a rather specific form of external data, namely ratings based on crowd sourcing. The example used is the rating of movies. Database reverse engineering and SQL tracking [3] can be used to feed a query repository. Collaborative query management [9] also creates a repository, but for a different purpose: User are supported in writing queries by searching for similar queries.

7 Summary and Outlook

The impetus of the project described in this paper is to take queries into account when designing a data management system that integrates heterogeneous data sources. This has many aspects that still need to be investigated in more detail. Queries can be specified in many different ways with different properties, and their execution can be prepared and finally done along many different paths. It is our belief that it is worthwhile to follow at least some of them.

We have already taken a first step as documented in [13]. The repository can be initialized with the help of query logs. The article proposes a particular approach to evaluate query logs in the context of data integration. The information required to do this will also be available in the repository sketched here. The query log is used to extract knowledge about data sources and thus contributes

to the contents of the second repository. This knowledge can then be utilized by the data scientists interacting with the data-integration system. The queries may contain fictional tables and attributes, and the system helps to find appropriate data sources. This generates a kind of kernel for the first repository.

Please note that it is not necessary to take a puristic view. The query-driven approach can easily be combined with the classical approach (that we may now call schema-driven). The message here is simply to consider the queries as well. Whatever knowledge about the data and their preliminary structure is available, it should certainly not be ignored.

Acknowledgement: The authors would like to thank the anonymous reviewers for their valuable remarks.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data – the story so far. *Int. Journal on Semantic Web & Information Systems* 5(3), 1–22 (2009)
2. Braunschweig, K., Eberius, J., Thiele, M., Lehner, W.: OPEN - enabling non-expert users to extract, integrate, and analyze open data. *Datenbank-Spektrum* 12(2), 121–130 (2012)
3. van den Brink, H.J., van der Leek, R.C.: Quality metrics for SQL queries embedded in host languages. In: *Proc. Special Session on System Quality and Maintainability (SQM, March 20) in conjunction with 11th European Conf. on Software Maintenance and Reengineering: “Software Evolution in Complex Software Intensive Systems” (CSMR, Amsterdam, the Netherlands, March 21-23)*. p. 2 (2007)
4. Das Sarma, A., Dong, X., Halevy, A.: Bootstrapping pay-as-you-go data integration systems. In: *Proc. SIGMOD*. pp. 861–874. ACM, New York, NY, USA (2008)
5. Doan, A., Halevy, A., Ives, Z.: *Principles of Data Integration*. Morgan Kaufmann, Waltham, MA, USA (2012)
6. Eberius, J., Thiele, M., Braunschweig, K., Lehner, W.: DrillBeyond: Processing multi-result open world SQL queries. In: *Proc. 27th Int. Conf. on SSDBM*. pp. 16:1–16:12. ACM (2015)
7. Eberius, J., Werner, C., Thiele, M., Braunschweig, K., Dannecker, L., Lehner, W.: DeExcellerator: a framework for extracting relational data from partially structured documents. In: *Proc. CIKM*. pp. 2477–2480 (2013)
8. Heath, T., Bizer, C.: *Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan & Claypool (2011)
9. Khoussainova, N., Balazinska, M., Gatterbauer, W., Kwon, Y., Suciu, D.: A case for a collaborative query management system. In: *Proc. CIDR* (2009)
10. Le, V., Gulwani, S.: FlashExtract: A framework for data extraction by examples. In: *Proc. PLDI (Edinburgh, United Kingdom, June 9-11)*. pp. 542–553 (2014)
11. Selke, J., Lofi, C., Balke, W.T.: Pushing the boundaries of crowd-enabled databases with query-driven schema expansion. *PVLDB* 5(6), 538–549 (2012)
12. Stonebraker, M., Bruckner, D., Ilyas, I.F., Beskales, G., Cherniack, M., Zdonik, S.B., Pagan, A., Xu, S.: Data curation at scale: The Data Tamer system. In: *Proc. CIDR* (2013)
13. Wahl, A.M.: A minimally-intrusive approach for query-driven data integration systems. In: *Proc. IEEE 32nd ICDE Workshops (ICDEW)* (2016)

SABER: Window-Based Hybrid Stream Processing for Heterogeneous Architectures^{*}

Alexandros Kolios[†], Matthias Weidlich^{†‡}, Raul Castro Fernandez[†],
Alexander L. Wolf[†], Paolo Costa[‡], and Peter Pietzuch[†]
{akolios, mweidlic, rc3011, alw, costa, prp}@imperial.ac.uk

[†]Imperial College London [‡]Humboldt-Universität zu Berlin [‡]Microsoft Research

Abstract

Stream processing systems found wide-spread application in domains such as credit fraud detection, urban traffic management, and click stream analytics. These systems process continuous streams of input data in an online manner, aiming at maximising processing *throughput* while staying within acceptable latency bounds. Heterogeneous architectures that combine multi-core CPUs with many-core GPGPUs have the potential to improve the performance of stream processing engines. Yet, a stream processing engine must execute streaming SQL queries with sufficient data-parallelism to fully utilise the available heterogeneous processors, and decide how to use each processor in the most effective way.

Addressing these challenges, we present SABER, a *hybrid* high-performance relational stream processing engine for CPUs and GPGPUs. It executes window-based streaming SQL queries following a hybrid execution model. Specifically, SABER incorporates the following innovations:

- It features a *hybrid stream processing model* based on query tasks, each comprising a batch of stream data and a query operator. Instead of relying on offline performance models to select the processor on which to run a query operator, SABER employs an adaptive *heterogeneous lookahead scheduling* strategy to balance the load on the different types of processors.
- It provides *window-aware task processing*, supporting sliding window semantics in the presence of fixed-sized batches. SABER ensures result correctness after the out-of-order processing of tasks by first buffering and then incrementally releasing the results as tasks finish execution.
- It exploits *pipelined stream data movement* to the GPGPU that interleaves data movement and task execution, thereby maintaining high utilisation of the PCIe bandwidth.

An experimental comparison against state-of-the-art engines shows that SABER increases processing throughput while maintaining low latency for a wide range of streaming SQL queries with both small and large window sizes.

^{*} Published as: A. Kolios, M. Weidlich, R. C. Fernandez, A. L. Wolf, P. Costa, and P. Pietzuch. SABER: Window-based hybrid stream processing for heterogeneous architectures. In F. Özcan, G. Koutrika, and S. Madden, editors, Proceedings of the 2016 SIGMOD Conference, San Francisco, CA, USA, June 26 - July 01, 2016, pages 555–569, ACM.

Graph n -grams for Scientific Workflow Similarity Search

David Luis Wiegandt, Johannes Starlinger, and Ulf Leser

Humboldt-Universität zu Berlin, Institut für Informatik
Unter den Linden 6, 10099 Berlin, Germany
{wigandtd, starling, leser}@informatik.hu-berlin.de
<https://www.informatik.hu-berlin.de>

Abstract. As scientific workflows increasingly gain popularity as a means of automated data analysis, the repositories such workflows are shared in have grown to sizes that require advanced methods for managing the workflows they contain. To facilitate clustering of similar workflows as well as reuse of existing components, a similarity measure for workflows is required. We explore a new structure-based approach to scientific workflow similarity assessment that measures similarity as the overlap in local structure patterns represented as n -grams. Our evaluation shows that this approach reaches state-of-the-art quality in scientific workflow comparison and outperforms some established scientific workflow similarity measures.

Keywords: Scientific Workflows, DAGs, n -grams, graph theory

1 Introduction

Over the last years, *scientific workflows (SWFs)* have emerged as a useful means of data analysis, particularly in the life sciences. SWFs are shared in (online) repositories to optimally support non-experts who cannot develop SWFs by themselves. A problem that arises with the growth of such repositories is the occurrence of (partial) duplicates. To avoid duplicates in the first place and enforce the reuse of existing components instead, and to allow for grouping of workflows that fulfil similar tasks (e.g. for result presentation), a similarity measure for SWFs is required. The problem of defining adequate similarity measures has been attacked in various ways, either based on the SWFs' graph structure or on their associated description and annotations in repositories. In this work, we focus on graph-based approaches as they are in principal applicable to arbitrary workflows, in contrast to approaches that rely on the presence of annotations.

Graphs can be compared using either global or local measures. A systematic approach to categorising and evaluating various established measures [14] yields that algorithms that consider a graph's full structure are too strict in the context of SWF similarity assessment. On the contrary, algorithms that fully ignore any structural information and only perceive graphs as sets of vertices are too loose, as two workflows that share a high amount of common tasks but with very different interconnections, should not be declared as similar. Instead, approaches that retain substructures only to a certain degree yield promising results [1,15,17].

Since it is our aim to measure the similarity of SWFs that are stored in repositories, we particularly focus on algorithms that decompose graphs into smaller units. These units, encoding a certain amount of structure, can be computed once when the workflow is added to the repository, and be reused in subsequent similarity assessments, making the computational complexity of the decomposition process negligible. In information retrieval, one way to represent such units are n -grams, a concept which has been applied to graphs as well.

For instance, in [17], workflows are perceived as finite state automata and the automata's states are encoded into fixed-size words called n -grams. Two workflows' similarity is measured by the overlap in their n -gram-sets. This approach was proven to outperform other similarity measures in the more general field of workflow similarity assessment, however, it cannot easily be transferred to SWFs, as, in general, SWFs lack transition labels that are required to convert them to FSAs. A similar approach that aims at similarity assessment of trees is presented in [1]. In [19], an n -gram-based approach to measuring two graphs' similarity in the context of graph similarity joins was presented as an alternative to the computationally expensive graph edit distance. It is among the state-of-the-art algorithms in this field [3].

Since n -gram-based approaches have proven to be successful in various domains, in this work, we transfer the concept to SWF similarity assessment. To this end, we particularly focus on n -grams that represent sub-paths of length n . For our evaluation, we make use of the gold standard corpus of manually retrieved similarity ratings introduced in [14].

In the following, we start with formal definitions of graphs and workflow graphs in Section 2, that we refer to in the subsequent review of existing solutions in Section 3. In Section 4, we introduce our novel approach which we evaluate in Section 5. In Section 6, we make specific propositions on how to improve our approach in the context of possible future work and close with some concluding remarks.

2 Preliminaries

2.1 Graphs

A *directed graph* $G = (V, E)$ is a tuple consisting of a vertex set V and an edge set $E \subseteq V \times V$. With $[n] := \{i \in \mathbb{N} : 1 \leq i \leq n\}$ we usually assume $V = [n]$ for some $n \in \mathbb{N}$. A *path* of length n is a sequence of vertices $v_1, \dots, v_n \in V$ that are connected by edges, i.e. $(v_{i-1}, v_i) \in E, i = 2 \dots n$. If $v_1 = v_n$, the path is called a *cycle*. Directed graphs without cycles are called directed acyclic graphs (*DAGs*). For the rest of this work, when talking about graphs we always mean DAGs if not explicitly stated otherwise.

With $\hat{E} := \{\{v, w\} : (v, w) \in E\}$, we call $\hat{G} := (V, \hat{E})$ the *underlying undirected graph* of G . A graph G is *connected* if there exists a path between any two vertices in V and *weakly connected* if only \hat{G} is connected. A *vertex-labelled* graph is a quadruple $G = (V, E, \Sigma, l)$ with labels over a finite alphabet Σ and a function $l : V \rightarrow \Sigma^*$, although we usually do not explicitly state the alphabet and the labelling function. An *edge-labelled* graph is defined analogously. The ratio of $|E|$ and the graph's maximum possible amount of edges is called its *density*.

Given any $v \in V$, we call $suc(v) = \{w \in V : (v, w) \in E\}$ its set of *successors* and $pre(v) = \{w \in V : (w, v) \in E\}$ its set of *predecessors*. Furthermore, we denote a vertices' *in-degree* by $\deg^-(v) = |pre(v)|$ and its *out-degree* by $\deg^+(v) = |suc(v)|$. We refer to $b(G) = \frac{|E|}{|V|}$ as the *average branching factor* of G , being the average out-degree (and in-degree) of any vertex $v \in V$. For some $S \subset V$ we call $G[S] = (S, E[S])$ with $E[S] = \{(v, w) \in E : v, w \in S\}$ the vertex-induced subgraph of G .

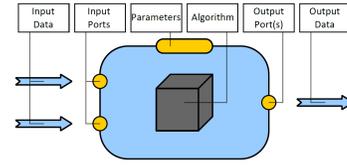
2.2 Workflow Graphs

Prodan et al define an SWF recursively as a tuple $G = (V, E_c, E_d, IN, OUT)$ [10, 204 ff.] with IN being the workflow's set of input ports, OUT being its set of output ports and V being its set of *tasks*. Each task $v \in V$ computes a function $v : IN^{(v)} \rightarrow OUT^{(v)}$. By connecting one task's output ports to another task's input ports, *control* and *data* flow dependencies $E_c, E_d \subseteq V \times V$ are introduced.

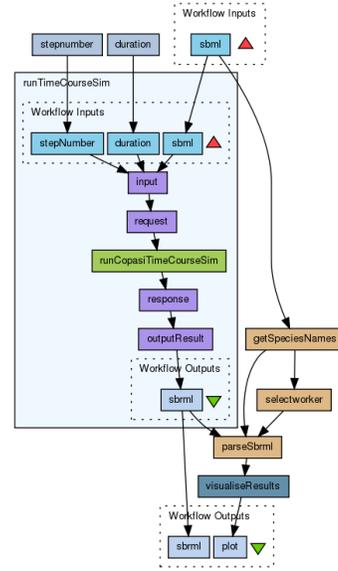
Following *Prodan et al*, we distinguish *atomic* and *composite* tasks, the latter being a set of nested sequential and parallel tasks that can, in turn, be composite and atomic tasks. Figure 1a shows an atomic task, whereas Figure 1b shows a complete SWF (itself a composite task) consisting of multiple atomic tasks and a composite task `runTimeCourseSim` we call a *sub-workflow*. Relations $\prec_c, \prec_d \subseteq V \times V$ are defined with \prec_c being the control flow precedence and \prec_d being the data flow precedence. Given two activities $v_1, v_2 \in V$, we write $v_1 \prec_c v_2$ iff v_1 can not be executed before v_2 finishes. Similarly, we write $v_1 \prec_d v_2$ iff v_1 can not be executed without being provided with data by v_2 . In contrast to *Prodan et al*, in this study we do not require $v_1 \prec_d v_2 \implies v_1 \prec_c v_2$ as we focus on the *Apache Taverna* Workflow Management System, which allows pipelining and streaming of data between tasks [16].

The set V of a workflow can be further subdivided into disjoint sets A (activities), J_{AND} (AND-Joins), J_{OR} (XOR-Joins), S_{AND} (AND-Splits) and S_{OR} (XOR-Splits) [5], allowing for conditional branching and parallel processing. This workflow model can be further extended, e.g. by adding a labelling function.

We make use of the fact that most *scientific workflow management systems* (such as *Apache Taverna*) prohibit looping, i.e. cycles in the workflow graph. This allows us to



(a) A labelled task of an SWF [2]



(b) An SWF with a sub-workflow [7]

Fig. 1: SWF components

represent SWFs as DAGs [16]: We set $E = E_c \cup E_d$ and omit the explicit mentioning of IN and OUT , because we are only interested in the underlying DAG's structure, which can be described by its set of vertices and edges as introduced in the previous section.

3 Related Work

The problem of workflow similarity assessment has been attacked in various different ways. As we focus on graph-based approaches, we group the approaches outlined in this section wrt. their preservation of topology and structure and distinguish (1) *local* approaches that completely disregard the graphs' structure, (2) *global* approaches that preserve the graphs' structure (almost) entirely and (3) *alternative approaches* that aim to compromise between local and global by preserving *substructures*.

Local Measures. Considering only the workflows' tasks, in [13], the Vector Space Model (VSM) as known from information retrieval [12] is adapted for workflow comparison. Given two vertex-labelled workflow graphs $G_i = (V_i, E_i, \Sigma_i, l_i)$ for $i = 1, 2$, each workflow is represented as a k -dimensional vector $D_i = (d_{i,1}, \dots, d_{i,k})$ with $T = l_1[V_1] \cup l_2[V_2] = \{t_1, \dots, t_k\}$ being the set of all vertex labels in both workflows and $d_{i,j} = |\{v \in V_i : l(v) = t_j\}|$, $j = 1..k$ being the amount of vertices in the i -th workflow that are labelled with t_j . In a last step, the similarity is computed as the cosine similarity between both workflows' vectors.

Another approach called *Module Sets (MS)*, that is similar to the VSM in that it completely disregards substructures by perceiving a workflow graph only as a set of tasks (called *modules* in the referenced work), proved to be comparably reliable and fast [14]: Given SWF graphs $G_i = (V_i, E_i)$ for $i = 1, 2$, all pairs of tasks in $V_1 \times V_2$ are compared (e.g. with regard to their label's string edit distances).

$$sim_M : V_1 \times V_2 \rightarrow [0, 1] \subset \mathbb{R} \quad (1)$$

Based on sim_M , a *maximum weight* matching $mw \subseteq V_1 \times V_2$ is computed and the similarity score is given by eq. (2).

$$nnsim_{MS}(V_1, V_2) = \sum_{(v_1, v_2) \in mw} sim_M(v_1, v_2) \quad (2)$$

In a last step, the similarity score is normalised over the workflows' size using a modified Jaccard index:

$$sim_{MS}(V_1, V_2) = \frac{nnsim_{MS}(V_1, V_2)}{|V_1| + |V_2| - nnsim_{MS}(V_1, V_2)} \quad (3)$$

Global Measures. An alternative examined in the same work [14] that turned out to provide more stable results than module sets is to retain substructures by comparing *path sets*. Given a DAG $G = (V, E)$, we denote the set of all paths of length n that start at a node with no inbound edges and end in a node with no outbound edges by $P_n(G)$

and the set of all such paths of arbitrary length by $PS(G)$:

$$P_n(G) = \{(v_1, \dots, v_n) \in V^n : (v_{i-1}, v_i) \in E, i = 2 \dots n, \deg^-(v_1) = \deg^+(v_n) = 0\} \quad (4)$$

$$PS(G) = \bigcup_{n=1}^{|V|} P_n(G) \quad (5)$$

Given SWF graphs G_1, G_2 and their path sets $PS_1 := PS(G_1), PS_2 := PS(G_2)$, the computation of the pairwise similarity score between paths (sim_P), the non-normalised ($nnsim_{PS}$) and the normalised similarity score (sim_{PS}) are defined analogous to the module sets approach by using paths instead of modules. To compute the similarity of two paths $P_1 = v_1, \dots, v_n, P_2 = w_1, \dots, w_m$, a *maximum weight non-crossing (mwnc) matching* is computed on their vertices, i.e. for all $(v_i, w_j), (v_k, w_l) \in mwnc \subseteq P_1 \times P_2$ it holds that $k > i \implies l > j$ and $k < i \implies l < j$ [8].

Another common graph distance measure is the graph edit distance. Given graphs G_1, G_2 and the three edit operations *insert*, *delete* and *substitute* (all applicable to edges and nodes), each with a certain cost assigned, the graph edit distance is defined as the minimum cost sequence of edit operations to transform G_1 into G_2 or vice versa [11]. Despite being optimal in terms of accuracy in finding structural differences between graphs, the graph edit distance's time complexity is exponential, making it inappropriate for large graphs. Further, it turned out to be too strict for SWF similarity assessment and yields comparably bad results [14].

Alternative Approaches. In [14], it was shown that graph comparison algorithms that consider the graphs full structure appear to be too strict for fine grained assessment of SWF similarity, whereas the comparison of SWFs by their module sets turned out to provide convincing results. Comparing substructures and focusing on the execution order of the workflows' tasks, on the other hand, turned out to lead to more reliable results, but is computationally expensive.

In [15], a new approach to SWF similarity search, that aims to compromise between local and global approaches, is introduced. *Layer Decomposition (LD)* performs a topological decomposition of a given SWF graph, i.e. it decomposes the graph into disjoint sets of vertices depending on the vertices relative position within their data flow strand. This is done by repeatedly removing all vertices with no inbound edges and their associated outbound edges. Given a graph $G = (V, E)$, the i -th layer L_i is defined recursively:

$$L_1(G) = \{v \in V : \deg^-(v) = 0\} \quad (6)$$

$$L_i(G) = \{v \in G[L_{i-1}(G)] : \deg^-(v) = 0\} \quad (7)$$

The ordered set $LD(G)$ that contains all layers is called the graph's layer decomposition. To compare two graphs G_1, G_2 by their layer decompositions $LD_1 := LD(G_1), LD_2 := LD(G_2)$, the approach of computing $nnsim_{LD}$ and sim_{LD} is analogous to the path and module sets, except for the use of a maximum weight non-crossing matching between layers and a maximum weight matching between two layers' vertices. LD is able to outperform other algorithms for SWF similarity assessment, including path sets, module sets and the graph edit distance [15].

The distinctive feature of layer decomposition and path set comparison is that the matching between two workflows' tasks within the retrieved substructures is performed

as a part of the similarity assessment, while the other approaches assume a global matching to be computed beforehand. Two workflows' tasks are matched with respect to their execution order, because tasks in one workflow can only be mapped to tasks that are within the matched layer or path in the other workflow. Since the matching is non-crossing, the overall similarity depends on the extent to which both workflows overlap in their module sets but also on the similarity of the order the tasks appear in. This is important as two workflows might share a great amount of common tasks (i.e. they have similar module sets) but due to their differing execution order, the workflows might implement different functionalities.

In the related area of automata theory, an n -gram-based approach was introduced [17] in which a workflow with transition labels in Σ^* is perceived as a *finite state automaton (FSA)* $M = (Q, \Sigma, \delta, q_0, F)$ with states Q , an alphabet Σ , a transition function $\delta : Q \times \Sigma \rightarrow Q$, an initial state $q_0 \in Q$ and final states $F \subseteq Q$. We define $\hat{\delta}(q, w)$ as the state of M after reading a word w , starting in state q :

$$\hat{\delta} : Q \times \Sigma^* \rightarrow Q \quad (8)$$

An n -gram is defined as a word $w \in \Sigma^n$ with $w = w_1 w_2 \dots w_n$, s.t. there exists a state from which on the word w is recognized, ending in a state $q \in Q$:

$$S_n(q) = \{w \in \Sigma^n : \exists r \in Q : \hat{\delta}(r, w) = q\} \quad (9)$$

The idea is to represent an automaton by the union of all states' n -grams and to measure two FSAs distance by the minimum sum of edit distances (the minimum cost sequence of insertions, deletions and substitutions for transforming one given string into another [18]) across all pairs of n -grams.

It was shown in [18] that this approach empirically outperforms certain structure- and language-based algorithms.

In summary, it can be observed that it is important to partially retain a workflow's structural topology and thus the execution order, but the increase in computational complexity should not be too high. Module and path set comparison, as well as the graph edit distance and layer decomposition have all been evaluated in the context of SWF comparison. In the following, we investigate the adaption of the concept of n -grams for SWFs.

4 Novel Approach

We propose a new approach that takes local graph patterns into account to compare graphs by measuring their commonalities regarding certain patterns. First, we introduce the notion of n -grams as used in our approach (similar to [17]). We assume \mathbb{S} to be the set of all SWF graphs.

Definition 1 (n -gram). *Given a graph $G = (V, E) \in \mathbb{S}$, we define the function $NGS_n : \mathbb{S} \rightarrow \mathbb{P}(V^n)$ to compute a set of all paths of length n as*

$$NGS_n(G) = \{(v_1, v_2, \dots, v_n) \in V^n \mid (v_i, v_{i+1}) \in E, i = 1 \dots n-1\} \quad (10)$$

with $NGS_n(G)$ being a set of n -grams of vertices.

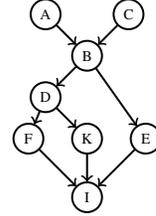
Algorithm 2: Path n -gram

input : Graph $G = (V, E)$, n
output: n -grams $\subseteq V^n$

```
1 procedure compute_paths( $G = (V, E)$ ,  $v \in V$ ,  $depth$ )
2   if  $depth = 0$  then
3     return  $\{(v)\}$ ;
4   else
5      $paths \leftarrow \emptyset$ ;
6     foreach  $w \in \text{suc}(v)$  do
7        $tmp \leftarrow \text{compute\_paths}(G, w, depth - 1)$ ;
8       foreach  $(u_1, \dots, u_{depth}) \in tmp$  do
9          $paths \leftarrow paths \cup \{(v, u_1, \dots, u_{depth})\}$ ;
10      end
11    end
12    return  $paths$ ;
13  end
14 end
15  $n\text{-grams} \leftarrow \emptyset$ ;
16 foreach  $v \in V$  do
17    $n\text{-grams} \leftarrow n\text{-grams} \cup \text{compute\_paths}(G, v, n - 1)$ ;
18 end
19 return  $n\text{-grams}$ ;
```

Example 1 (n -grams). We illustrate the concept of n -grams by calculating all 3-grams for the graph on the right. To simplify the notation, we denote the tuples by concatenating their vertices. We obtain the following 3-gram set:

$$NGS_3(G) = \{ABD, ABE, BDF, BDK, BEI, CBD, \\ CBE, DFI, DKI\}$$



For the computation of a graph's set of n -grams we propose algorithm 2, based on a modified *iterative deepening depth-first search (IDDFS)* [6]. The decision to propose an IDDFS-based algorithm instead of a Floyd-Warshall-based algorithm was made because the Floyd-Warshall algorithm is less efficient for graphs with a low density, which is, in turn, correlated with the graph's average branching factor. It has been shown in [9], that most workflows have a rather low average branching factor, which also implies a low density. Our observations regarding our evaluation corpus, where the median density is approximately 0.3, support these findings. Our evaluation corpus (introduced in [14]) contains 1483 of the 2123 SWFs in the *myExperiment* repository, the largest public SWF repository to date.

To derive a measure from two graphs' sets of n -grams, we first introduce a method to compare two n -grams. For the remainder of this section, we assume two graphs $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$ and refer to their accompanying n -gram sets by $NGS_1 := f(G_1) \subseteq V_1^n$, $NGS_2 := f(G_2) \subseteq V_2^n$ with $f \in \bigcup_{n \in \mathbb{N}} NGS_n$.

Since an n -gram $a = (a_1, \dots, a_n) \in NGS_n(G)$ for some graph $G = (V, E)$ is a vector in the n -dimensional vector space V^n , we refer to the i -th component of an n -gram by a_i .

Definition 2 (n -gram similarity). Given two graphs $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$, a similarity function $sim : V_1 \times V_2 \rightarrow [0, 1]$ and n -gram sets NGS_1, NGS_2 , we define the similarity of two n -grams $a \in NGS_1, b \in NGS_2$ as

$$sim_{NG}(a, b) = \frac{1}{n} \sum_{i=1}^n sim(a_i, b_i) \quad (11)$$

In our case, as well as in [15], the sim -function is derived from the normalised string edit distance between the tasks' labels a_i, b_i . To retrieve a set of mappings of n -grams that are present in both sets wrt. the global maximum weight matching, we introduce the n -gram mapping operator:

Definition 3 (n -gram mapping operator). Given two n -gram sets NGS_1, NGS_2 and a global maximum weight matching $mw \subseteq V_1 \times V_2$ between both workflows' tasks, we define the set of common n -grams as

$$NGS_1 \otimes^{mw} NGS_2 := \{(a, b) \in NGS_1 \times NGS_2 : (a_i, b_i) \in mw, i = 1..n\} \quad (12)$$

The similarity of two n -gram sets NGS_1, NGS_2 is calculated as the sum of the similarities of all n -gram mappings in $NGS_1 \otimes^{mw} NGS_2$:

Definition 4 (n -gram set similarity measure). Given two n -gram sets NGS_1, NGS_2 , we define the non-normalised similarity as

$$nnsim_{NGS}(NGS_1, NGS_2) := \sum_{(a,b) \in NGS_1 \otimes^{mw} NGS_2} sim_{NG}(a, b) \quad (13)$$

To normalise this similarity value, we use a modified Jaccard index as described in [14] and already used for layer decomposition [15]:

Definition 5 (Normalisation). We define the normalised similarity of two n -gram sets NGS_1, NGS_2 as

$$sim_{NGS}(NGS_1, NGS_2) := \frac{nnsim_{NGS}(NGS_1, NGS_2)}{|NGS_1| + |NGS_2| - nnsim_{NGS}(NGS_1, NGS_2)} \quad (14)$$

We also considered another type of n -grams that encode each vertices' neighbourhood (i.e. its predecessors, successors and itself) similar to [1]. Due to the low average branching factor in SWFs, this type of pattern turned out to be too restrictive and its detailed analysis is omitted here.

5 Evaluation

5.1 Setup

For the evaluation of our novel approach, we make use of the gold standard of similarity ratings utilised in [15]: From a corpus of 1485 Apache Taverna workflows, 24 *query*

workflows were chosen, each having another 10 workflows associated with it. 15 SWF experts from 6 different institutions were shown a query workflow along with one of its 10 associated workflows and had to decide whether they are *very similar*, *similar*, *related* or *dissimilar*. A fifth option was to choose *unsure*, though these ratings were not further considered in the evaluation. The result were multiple rankings of 10 workflows, all ordered by their similarity to the respective query workflow.

In a last step, a *consensus* ranking was computed from these rankings to aggregate the single experts' ratings. We use this consensus to compare it to the rankings retrieved by our algorithm using the measures introduced in [4]: We represent a consensus ranking and a ranking obtained by running our algorithm as partial orders \sqsubset_* and \sqsubset , respectively, and call a pair of workflows W_1, W_2 *concordant* if it appears in the same order in both rankings, i.e.:

$$C = \{(W_1, W_2) : (W_1 \sqsubset W_2 \wedge W_1 \sqsubset_* W_2) \vee (W_2 \sqsubset W_1 \wedge W_2 \sqsubset_* W_1)\} \quad (15)$$

If the order in our predicted ranking differs from the consensus ranking, we call it *discordant*, i.e.:

$$D = \{(W_1, W_2) : (W_1 \sqsubset W_2 \wedge W_2 \sqsubset_* W_1) \vee (W_2 \sqsubset W_1 \wedge W_1 \sqsubset_* W_2)\} \quad (16)$$

To compare a ranking obtained using our algorithm to the respective consensus ranking, we count the amount of concordant and discordant pairs and use the definition of ranking *correctness* (eq. (17)) and *completeness* (eq. (18)):

$$CR(\sqsubset, \sqsubset_*) = \frac{|C| - |D|}{|C| + |D|} \quad (17) \quad CP(\sqsubset) = \frac{|C| + |D|}{|\sqsubset_*|} \quad (18)$$

The value of correctness is in $[-1, 1]$, with -1 indicating a perfectly negative correlation between \sqsubset_* and \sqsubset as in this case $|C| = 0$. Conversely, $+1$ indicates a perfectly positive correlation. On the other hand, completeness measures the number of pairs whose elements can be distinguished by the experts but not by our algorithm.

5.2 Results

Figure 2a shows the results: The coloured bars show the algorithms' ranking correctness, whereas the small black squares indicate ranking completeness. The black error bars visualise the standard deviation of ranking correctness. All approaches present in the diagrams use matchings of tasks based on the edit distance of the tasks' labels.

For $n \in \{3, 4, 5\}$, n -grams increasingly outperform path sets in terms of correctness, for $n = 5$ even with a slightly lower variance as can be seen in Figure 2a. Contrariwise, there is an overall decrease in completeness as n grows, making n -grams the only approach with a completeness far below 1. This is due to our definition of an n -gram: As opposed to e.g. path sets, we require both graphs to contain at least one path of length n , whereas in path sets, where graphs are compared by paths starting at a source and ending in a sink without any constraints on the paths' lengths, even a single vertex per graph is sufficient to conduct a similarity assessment. The same applies to module sets, graph edit distance and layer decomposition, as neither of them imposes any requirements on the graphs' size.

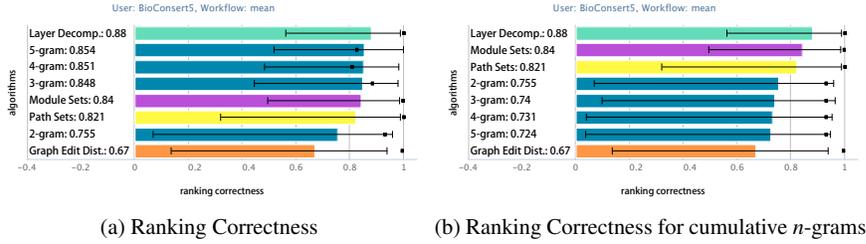


Fig. 2: Mean Ranking Correctness for Different Algorithms and Setups

To analyse the impact of this design specificity, we first define *coverage* as the ratio of a graph's vertices, that are part of at least one n -gram. The histogram in Figure 3 shows the share of workflows in our corpus whose coverage is within different ranges for n -grams of size $n = 2..5$. For instance, in 45% of the workflows in our corpus, between 80% and 100% of the vertices are covered by 4-grams, while 48% do not even contain a single 4-gram, as can be seen in the green and light blue bars of p4. Further, only 67% contain at least one 3-gram, which implies that only $\approx 45\%$ of all possible pair-wise comparisons can be conducted using 3-grams. For 4-grams, the share further decreases to $\approx 27\%$.

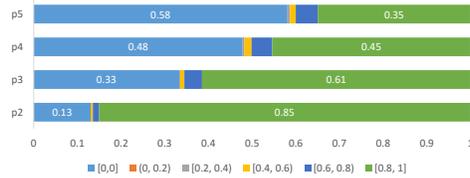


Fig. 3: A histogram depicting the coverage of the graphs in our corpus by n -grams of different sizes.

Observations regarding our corpus of 1483 SWFs yield that the median number of (a) vertices is 5, (b) edges is 4 and (c) average branches per vertex is 0.8, i.e. the majority of SWF graphs are comparably small, prohibiting a further increase in correctness by the use of larger n -grams. This is the major problem that is inherent to our approach, as we have to *reject* a comparison whenever at least one of the workflows that are to be compared does not contain an n -gram of the required size, leading to a lower completeness.

An obvious remedy could be to let the value of n depend on the involved graphs' size, instead of requiring both graphs to contain at least one path of a fixed length n . To this end, we define the maximum n that a graph contains at least one n -gram for as:

$$\max n(G) = \max\{n \in \mathbb{N} : NGS_n(G) \neq \emptyset\} \quad (19)$$

Given workflow graphs G_1, G_2 , we let $n_1 = \max n(G_1), n_2 = \max n(G_2)$ and compute the similarity as the (weighted) sum of similarity scores for all $2 \leq n \leq n_*$ with $n_* = \min(n_1, n_2)$:

$$\text{csim}(G_1, G_2) := \frac{2}{n_*(n_*+1)-2} \sum_{i=2}^{n_*} i \cdot \text{sim}_{NGS}(NGS_i(G_1), NGS_i(G_2)) \quad (20)$$

In eq. (20), the weighted sum is normalised by dividing it by the sum of all integers from 2 to n_* , with $\frac{2}{n_*(n_*+1)-2}$ being derived from the Gaussian summation formula.

Unfortunately, the great increase in completeness (≈ 0.92 for all $n \in \{2, 3, 4, 5\}$) comes at the cost of bad correctness and variance as can be seen in Figure 2b. In fact, as noted in [4], there exists a negative correlation between correctness and completeness, that is due to the definitions of the evaluation measures, i.e. the correctness can be increased at the cost of more rejections and, in turn, a lower completeness. As a result, our initial approach performs better as it simply rejects workflow pairs it can not compare, instead of falling back to a lower n which would yield a less reliable similarity value.

6 Conclusion

In this work, we introduced n -grams in the context of SWF comparison as fixed-size local graph patterns that cover a vertices context to different extents, depending on the value of n . To compare two SWFs by their sets of n -grams, we adapted the similarity measure from [14,15] and were able to show that n -grams for $n \in \{3, 4, 5\}$ outperform path and module sets and are close to layer decomposition in terms of the mean correctness. However, we still see potential for various optimisations based on these first results on n -gram based SWF similarity.

First of all, we see the evaluation of other local graph patterns as a possible direction of further research. For instance, another interesting approach could be to compare only data-flow splits or to retrieve similarity values for multiple patterns and to assign a weight to each one. Regarding the approach presented in this work, we propose optimisations specifically dedicated to either correctness or completeness:

Correctness. To further increase correctness, the matching strategy could be adjusted. In particular, a more local matching such as a maximum weight non-crossing matching *within* n -grams combined with a maximum weight matching *across* n -grams should lead to an increase in correctness as the similarity measure would become more sensitive to partly similar n -grams and thus would allow for a more fine-grained similarity assessment.

Also, some modules in an SWF perform trivial tasks with a low specificity for a certain context [14], which is particularly a problem for higher n , which yield smaller n -gram sets. To this end, an optimisation the correctness can be expected to benefit from is *importance projection* introduced in [14], where modules performing trivial tasks are removed from the graph prior to the similarity assessment. As the workflow graphs are possibly reduced in size by the use of importance projection, the n -gram retrieval algorithm (Algorithm 2) would also benefit in terms of run time, as its run time complexity has an exponential correlation with n .

Completeness. The most obvious way to increase the completeness would be to relax the definition of the n -gram similarity measure from Definition 2 to instead compare n -grams of possibly different sizes. Formally, we compute the n -gram set for some $n_1 = \min(\maxn(G_1), n)$ (with \maxn as defined in eq. (19)) for workflow G_1 and some n_2 for G_2 respectively. The overall comparison process becomes similar to the path set comparison as presented in Section 3: To preserve the execution order within the compared n -grams, a maximum weight non crossing matching is computed within the

n -grams and the overall similarity is the maximum weight matching of both workflows' n -gram sets as described before regarding the correctness.

Altogether, we have shown n -grams to be a powerful technique for SWF similarity assessment, that still leaves room for further refinements to take on in future work.

References

1. Augsten, N., Böhlen, M., Gamper, J.: Approximate matching of hierarchical data using pq-grams. In: PVLDB 2005. pp. 301–312. VLDB Endowment (2005)
2. Bux, M., Leser, U.: Parallelization in scientific workflow management systems. arXiv preprint arXiv:1303.7195 (2013)
3. Chen, Y., Zhao, X., Xiao, C., Zhang, W., Tang, J.: Efficient and scalable graph similarity joins in mapreduce. *TheScientificWorldJournal* 2014, 749028 (2014)
4. Cheng, W., Rademaker, M., De Baets, B., Hüllermeier, E.: Predicting partial orders: ranking with abstention. In: ECML-PKDD. pp. 215–230. Springer (2010)
5. Kiepuszewski, B., ter Hofstede, A., van der Aalst, W.: Fundamentals of control flow in workflows. *Acta Informatica* 39(3), 143–209 (2003)
6. Korf, R.E.: Depth-first iterative-deepening: An optimal admissible tree search. *Artificial intelligence* 27(1), 97–109 (1985)
7. Li, P.: Copasi time simulation of sbml model. <http://www.myexperiment.org/workflows/1202/versions/1/previews/full>, accessed: 05.03.2016
8. Malucelli, F., Ottmann, T., Pretolani, D.: Efficient labelling algorithms for the maximum noncrossing matching problem. *Discrete Applied Mathematics* 47(2), 175–179 (1993)
9. Ostermann, S., Prodan, R., Fahringer, T., Iosup, R., Epema, D.: On the characteristics of grid workflows. In: CoreGRID Symposium-Euro-Par. vol. 2008, pp. 1–12 (2008)
10. Prodan, R., Fahringer, T.: *Grid Computing*, vol. 4340. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
11. Riesen, K., Bunke, H.: Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision Computing* 27(7), 950–959 (2009)
12. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Communications of the ACM* 18(11), 613–620 (1975)
13. Santos, E., Lins, L., Ahrens, J.P., Freire, J., Silva, C.T.: A first study on clustering collections of workflow graphs. In: *Provenance and Annotation of Data and Processes*, pp. 160–173. Springer (2008)
14. Starlinger, J., Brancotte, B., Cohen-Boulakia, S., Leser, U.: Similarity search for scientific workflows. *Proceedings of the VLDB Endowment* 7(12), 1143–1154 (2014)
15. Starlinger, J., Cohen-Boulakia, S., Khanna, S., Davidson, S.B., Leser, U.: Layer decomposition: An effective structure-based approach for scientific workflow similarity. In: *10th International Conference on e-Science*. vol. 1, pp. 169–176. IEEE (2014)
16. Talia, D.: *Workflow systems for science: concepts and tools*. ISRN Software Engineering 2013 (2013)
17. Wombacher, A.: Evaluation of technical measures for workflow similarity based on a pilot study. In: *On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE*, pp. 255–272. Springer (2006)
18. Wombacher, A., Li, C.: Alternative approaches for workflow similarity. In: *Services Computing (SCC), 2010 IEEE International Conference on*. pp. 337–345. IEEE (2010)
19. Zhao, X., Xiao, C., Lin, X., Wang, W.: Efficient graph similarity joins with edit distance constraints. In: *2012 IEEE International Conference on Data Engineering (ICDE 2012)*. pp. 834–845 (2012)

Discovering Data Transformations in Web Resources (Abstract)

Ziawasch Abedjan¹ John Morcos² Ihab F. Ilyas²
Mourad Ouzzani³ Paolo Papotti⁴ Michael Stonebraker⁵

¹ TU Berlin ² University of Waterloo ³ Qatar Computing Research Institute
⁴ Arizona State University ⁵ MIT CSAIL
abedjan@tu-berlin.de {jmorcos,ilyas}@uwaterloo.ca
mouzzani@qf.org.qa ppapotti@asu.edu stonebraker@csail.mit.edu

Abstract. In data integration, data curation, and other data analysis tasks, users spend a considerable amount of time converting data from one representation to another. For example US dates to European dates or airport codes to city names. In practice, data scientists have to code most of the transformation tasks manually, search for the appropriate dictionaries, and involve domain experts. In a previous vision paper, we presented the initial design of **DataXFormer**, a system that uses web resources to assist in transformation discovery [1]. Specifically, **DataXFormer** discovers possible transformations from web tables and web forms and involves human feedback where appropriate. We demonstrated the system at SIGMOD 2015 and deployed an open version of the system, which helped us to increase our initial workload from 50 to 120 transformations [3]. At the same time we extended **DataXFormer** with new algorithms to find

1. transformations that entail multiple columns of input data,
2. indirect transformations that are compositions of other transformations,
3. transformations that are not functions but rather relationships, and
4. transformations from a knowledge base of public data.

We report on experiments with the collection of 120 transformation tasks, and show our enhanced system automatically covers 101 of them by using openly available web resources [2].

References

1. Z. Abedjan, J. Morcos, M. Gubanov, I. F. Ilyas, M. Stonebraker, P. Papotti, and M. Ouzzani. DataXFormer: Leveraging the web for semantic transformations. In *CIDR*, 2015.
2. Z. Abedjan, J. Morcos, I. F. Ilyas, P. Papotti, M. Ouzzani, and M. Stonebraker. DataXFormer: A robust data transformation system. In *ICDE*, 2016.
3. J. Morcos, Z. Abedjan, I. F. Ilyas, M. Stonebraker, P. Papotti, and M. Ouzzani. DataXFormer: An interactive data transformation tool. In *SIGMOD*, 2015.

Scalable Detection of Emerging Topics and Geo-spatial Events in Large Textual Streams

Erich Schubert, Michael Weiler, and Hans-Peter Kriegel

Institut für Informatik, LMU Munich, Germany
{schube,weiler,kriegel}@dbs.ifi.lmu.de

Social media are a popular source for live textual data. This data poses several challenges due to its size, velocity, and heterogeneity. Existing methods for emerging topic detection often are only able to detect events of a global magnitude such as natural disasters, or they can only monitor user-selected keywords or a curated set of hashtags. Interesting emerging topics may, however, be of much smaller magnitude and may involve the combination of two or more words that are not yet known in beforehand.

We present several contributions introduced in previous work [1, 2]:

- (i) A significance measure that can detect emerging topics early, long before they evolve into “hot tags”, by drawing upon experience from outlier detection.
- (ii) An efficient online algorithm to track these statistics for all words and word-pairs with only a fixed amount of memory, and without predefined keywords.
- (iii) The clustering of the detected co-trends into larger topics, because a single event will cause multiple word combinations to trend at the same time.
- (iv) How to incorporate location information into this process to both allow reporting the locality of events as well as detecting local-only geo-textual patterns.

The significance score provides an estimated frequency and standard deviation of words, word-pairs, and word-location information on the data stream at minimal cost. It allows for normalization across location, culture, and language and enables the detection of change events both in already frequent and not previously seen combinations. In contrast to earlier work, it can monitor every word at every location with only a fixed amount of memory, compare the values to statistics from earlier data, and immediately report significant deviations with minimal delay. The algorithm is capable of reporting “Breaking News” in real-time as they happen in social media around the world. Location is modeled at different granularities, such that events can be detected at a city, country, or global level by incorporating OpenStreetMap data, or at particular coordinates.

References

- [1] E. Schubert, M. Weiler, and H.-P. Kriegel. “SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds”. In: *Proc. 20th ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*. 2014, pp. 871–880.
- [2] E. Schubert, M. Weiler, and H.-P. Kriegel. “SPOTHOT: Scalable Detection of Geo-spatial Events in Large Textual Streams”. In: *Proc. 28th Int. Conf. on Scientific and Statistical Database Management (SSDBM)*. 2016.

Approaches for Annotating Medical Documents

Victor Christen, Anika Groß, Erhard Rahm

Department of Computer Science, University of Leipzig, Germany
{christen,gross,rahm}@informatik.uni-leipzig.de

Abstract. Annotations are useful to semantically enrich documents and other datasets with concepts of ontologies. In the medical domain, many documents are not annotated at all and manual annotation is a difficult process making automatic annotation methods highly desirable to support human annotators. We propose a linguistic-based and a reuse-based approach annotating medical documents by concepts from an ontology. The reuse-based approach utilizes previous annotations to annotate similar medical documents. The approach clusters items in documents such as medical forms according to previous ontology-based annotations and uses these clusters to determine candidate annotations for new items.

1 Introduction

The annotation of data with concepts of standardized vocabularies and ontologies has gained increasing significance due to the huge number and size of available datasets as well as the need to deal with the resulting data heterogeneity. Annotations of medical documents such as Electronic Health Records (EHR) that are used to document the history of patients can also support advanced analyses and searches. For instance, they can be used to identify significant co-occurrences between the use of certain drugs and negative side effects in terms of occurring diseases [5]. Moreover, case report forms are used for examining clinical trials, e.g. to ask for the medical history of probands. To enable an efficient search for medical documents, annotations can be used to semantically look for a certain set of forms, e.g., in the MDM repository of medical data models [2] and to design new forms with a similar topic.

To improve the value of medical documents for analysis, reuse and data integration it is thus crucial to annotate them with concepts of ontologies. Since the number, size and complexity of medical documents and ontologies can be very large, a manual annotation process is time-consuming or even infeasible. Hence, automatic annotation methods become necessary to support human annotators with recommendations for manual verification. The goal of an annotation method is the identification of annotations for a collection of medical documents. An annotation is an association between a document and a concept from an ontology, where the concept covers the semantics of the document. Therefore, a document might be annotated with more than one concept to precisely describe the content of the document. The use of annotations enables a standardized representation,

since an ontology is a unified set of concepts and a set of relationship interrelating the ontology concepts by certain relationship types, e.g. *is - a*, *part - of* or domain-specific relationships such as *is - located - in*. The annotation of documents by using concepts of an ontology is related to the entity-linking problem that is a well studied field [6]. Moreover, there exist different annotation methods such as MetaMap [1] that annotates medical documents with concepts of UMLS by applying a linguistic-based approach.

In our recent work, we realized different annotation methods to identify annotations for medical forms based on concepts of UMLS. We initially start with a linguistic-based annotation approach [4]. A crucial part of an annotation method is the identification of annotation candidates in terms of effectiveness and efficiency. In general, a medical document or a collection of medical documents cover topically a subset of an ontology. Moreover, the quality of annotation candidates depends on the quality of synonyms and labels for a concept. We overcome such issues by creating a reuse repository for utilizing verified annotated documents [3]. We are able to build more compact and preciser representatives for a concept based on the verified documents than the synonyms and labels for a concept. Moreover, the reuse of the generated representatives to annotate a set of medical documents is more efficient than using the whole ontology.

2 Linguistic-Based Annotation Approach

The workflow consists of a preprocessing, a candidate identification and a selection step (see Fig. 1). The input of the workflow is a set of forms \mathcal{F} , an ontology \mathcal{O} , and a similarity threshold δ . This kind of documents consists of a set of question that we want to annotate with a set of concepts. In our case, we use concepts from the Unified Medical Language System (UMLS) that is an integrated knowledge system including several biomedical ontologies. First, we normalize the labels and synonyms of ontology concepts by removing stop words, transforming all string values to lower case and removing delimiters. The same preprocessing steps are applied for each form F_i . We identify an intermediate annotation mapping $\mathcal{M}_{F_i, \mathcal{O}}$ by lexicographically comparing each question with the labels and synonyms of ontology concepts. For this purpose, we apply three string similarity measures, namely trigram, TF/IDF as well as a longest common sequence string similarity approach. We keep an annotation (q, c, sim) for a question q and a concept c , if the maximal similarity sim of the three string similarity approaches exceeds the threshold δ . Finally, we select annotations from the intermediate result by not only choosing the concepts with the highest similarity but also by considering the similarity among the concepts. For

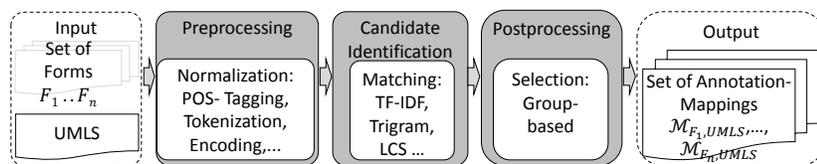


Fig. 1. Workflow of the linguistic-based annotation approach

this purpose, we group the concepts associated with a question based on their mutual similarity and only choose the concept with the highest similarity per group in order to avoid the redundant selection of highly similar concepts. This group-based selection proved to be quite effective in [4] albeit it only considers the string-based (linguistic) similarity between questions and concepts, and among concepts.

3 Reuse-based Annotation Approach

The workflow for the reuse-based annotation approach is shown in Figure 2. Its input includes a set of verified annotation mappings containing the annotations for reuse. The result is a set of annotation mappings $\mathcal{M}_{\mathcal{F},\mathcal{O}}$ for the unannotated input forms \mathcal{F} w.r.t. ontology \mathcal{O} . In the first step, we use the verified annotations to determine a set of *annotation clusters* $\mathcal{AC} = \{ac_{c_1}, ac_{c_2}, \dots, ac_{c_m}\}$. For each concept c_i used in the verified annotations, we have an annotation cluster ac_{c_i} containing all questions that are associated to this concept. To calculate the similarity between an unannotated question and the questions of an annotation cluster we determine for each cluster a *representative* (feature set) $ac_{c_i}^{fs}$ consisting of relevant term groups in this cluster. A relevant term group is either a frequently co-occurring term group in the questions of the cluster or the maximized overlap between the terms of a question and the synonyms or the label of a concept, i.e., we do not use term groups that build a subset of another frequently occurring term group. As an example, Figure 3 shows the resulting annotation cluster $ac_{C0023467}$ for UMLS concept $C0023467$ about the disease *Acute Myeloid Leukaemia*. In the UMLS ontology, this concept is described by a set of 32 synonyms (Figure 3 left). The annotation cluster also contains 25 questions associated to this concept in the verified annotation mappings. Most questions only relate to some of the synonym terms of the concept while other synonyms remain unused. So the abbreviation 'AML' that is a part of some synonyms is often used but the abbreviation 'ANLL' does not occur in the medical forms used to build the annotation clusters. For this example, we generate only 9 relevant term groups, i.e., the representative feature set of the cluster is much more compact than the free text questions and large synonym set.

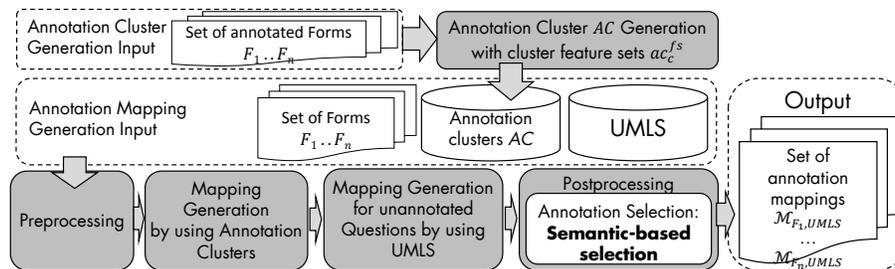


Fig. 2. Workflow of the reuse-based annotation approach

C0023467	$Q_{C0023467}$	$ac_{C0023467}^{fs}$
ANLL, AML, Acute myelocytic leukaemia, AML - Acute myeloid leukaemia, acute myelogenous leukemia (AML) ⋮	1. Previous induction-type chemotherapy for MDS or AML 2. Relapsed or treatment refractory AML 3. Patients with relapsed AML 4. Patients older than 60 years with acute myeloid leukemia according to FAB (>30 % bone marrow blasts) not qualifying for, or not consenting to, standard induction chemotherapy or immediate allografting	AML, acute myeloid leukemia, acute promyelocytic leukemia, acute myelodysplastic leukaemia ⋮
32 synonyms	25 questions	9 term groups

Fig. 3. Sample annotation cluster $ac_{C0023467}$ for UMLS concept $C0023467$ with its set of associated questions $Q_{C0023467}$ and feature set $ac_{C0023467}^{fs}$.

After these initial steps we determine the annotation mapping for each unannotated input form F_i . We first preprocess a form and the ontology as in the base approach (see Fig. 1). Then we determine an annotation mapping $\mathcal{M}_{F_i, \mathcal{O}}^{Reuse}$ for the form based on the annotation clusters. Depending on the degree of reusable annotations the determined mapping is likely to be incomplete. We thus identify all questions that are not yet covered by the first mapping. For these questions we apply the base algorithm to match them to the whole ontology and obtain a second annotation mapping. We then take the union of the two partial mappings to obtain the intermediate mapping $\mathcal{M}'_{F_i, \mathcal{O}}$. Finally, we apply a context-based selection strategy to determine the annotations for the final mapping $\mathcal{M}_{\mathcal{F}, \mathcal{O}}$. The input for the selection of annotations is a set of grouped candidate concepts for each question in the medical forms \mathcal{F} . To determine the final annotations per question, we rank the candidate concepts within each group based on a combination of both linguistic and context-based similarity among the candidate concepts. For this purpose, we consider two criteria for a set of candidate concepts of a certain question: first, the degree to which concepts co-occurred in the annotations for the same question within the verified annotation mapping, and second, the degree of semantic (contextual) relatedness of the concepts w.r.t. the ontological structure. The goal is to give a high contextual similarity (and thus a high chance of being selected) to frequently co-occurring concepts and to semantically close concepts. To determine a context-based similarity, we construct a *context graph* $G_q = (V_q, E_q)$ for each question q . The vertices V_q represent candidate concepts that are interconnected by two kinds of edges in E_q to express that concepts have co-occurred in previous annotations or that concepts are semantically related within the ontology. In both cases we assign distance scores to the edges that will be used to calculate the context similarity between concepts.

4 Evaluation

We evaluate the proposed annotation approaches for medical forms and compare it with the MetaMap tool. Our evaluation uses medical forms about eligibility criteria (EC) and about quality assurance (QA) w.r.t cardiovascular procedures

from the MDM platform [2]. To evaluate the quality of automatically generated annotations, we use manually created reference mappings from the MDM portal. These reference mappings might not be perfect ("a silver standard") since the huge size of UMLS makes it hard to manually identify the most suitable concepts for each item. To analyze the quality of the resulting annotation mappings, we compute precision, recall and F-measure using the union of all annotated form items in the evaluation dataset. Table 4 shows the number of forms, items and verified annotations for the reuse and evaluation datasets.

dataset	ECRD1	ECRD2	ECeval	QARD1	QARD2	QAEval
#forms	200	100	25	16	32	23
#items	3125	1638	310	453	795	609
#annotations	13027	6911	578	694	1054	668

Table 4. Statistics on the reuse and evaluation datasets for EC and QA

Figure 5 shows the results for the two datasets and different configurations. Our reuse-based approach outperforms MetaMap in terms of mapping quality for each dataset. For the EC dataset, F-Measure is improved by $\sim 4\%$ (EC_{RD1}) and $\sim 8.6\%$ (EC_{RD2}) indicating that the the computed annotation clusters allow a more effective identification of annotations than with the original concept definition. In addition, our approach benefits from using the ontological relationships for selecting annotations resulting in a much better precision than using MetaMap (54.5% for EC_{RD2} than compared to 43.1%). While MetaMap achieved a better F-Measure than the baseline approach for the EC dataset it performed poorly for the QA dataset where its best F-Measure of 44.8% was much lower for the baseline approach and reuse-based approaches (57.5 and 59%), mainly because of a very low recall for Metamap.

A positive side of MetaMap is its high performance due to the use of an indexed database for finding annotations. Its runtimes were up to 13 times faster than for the baseline approach and it was also faster than the reuse-based approach. In future work we will study whether the use of MetaMap in combination with the reuse approach, either as an alternative or in addition to the baseline approach, can further improve the annotation quality.

5 Conclusion

We proposed a linguistic-based and a reuse-based approach to semantically annotate medical documents such as EHRs with concepts of an ontology. The linguistic-based approach identifies an annotation mapping between a form and an ontology by comparing each question of the form with the synonyms or labels of each concept from an ontology. The reuse-based approach avoids the comparison of each concept by utilizing already found and verified annotations for similar CRFs. It builds so-called annotation clusters combining all previously annotated questions related to the same medical concept. New questions are matched with the identified cluster representatives to find candidates for annotating concepts.

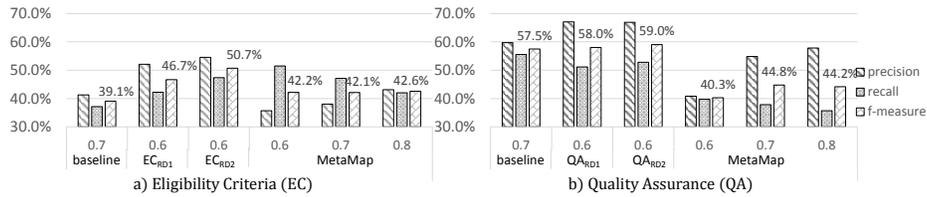


Fig. 5. Comparison of the quality for the resulting annotation mappings from the baseline approach, reuse-based approach and MetaMap.

To identify the most promising annotations, we proposed a context-based selection strategy based on the semantic relatedness of concept candidates as well as known co-occurrences from previous annotations. We compared our approaches with MetaMap and showed that the reuse-based approach outperforms the annotation method of MetaMap in terms of quality. However, the efficiency is lower than MetaMap, since it uses an indexed database.

For future work, we plan to use different annotation frameworks for generating more candidates and to get more evidence for correctness. We also plan to build a reuse repository covering annotation clusters and their feature sets for different medical subdomains. Such a repository can be used to identify annotations for new medical documents. It further enables a semantic search for existing medical document annotations. This can be useful to define new medical forms by finding and reusing suitable annotated items instead of creating new forms from scratch.

References

1. A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proc. AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
2. B. Breil, J. Kenneweg, F. Fritz, et al. Multilingual medical data models in ODM format—a novel form-based approach to semantic interoperability between routine health-care and clinical research. *Appl Clin Inf*, 3:276–289, 2012.
3. V. Christen, A. Groß, and E. Rahm. A reuse-based annotation approach for medical documents. In *Submitted for: International Semantic Web Conference (ISWC)*, 2016.
4. V. Christen, A. Groß, J. Varghese, M. Dugas, and E. Rahm. Annotating medical forms using UMLS. In *Data Integration in the Life Sciences (DILS)*, volume 9162 of *LNCS*, pages 55–69. 2015.
5. P. LePendou, S. Iyer, C. Fairon, N. H. Shah, et al. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. *Journal of Biomedical Semantics*, 3(S-1):S5, 2012.
6. W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2015.

Robust Query Processing in Co-Processor-accelerated Databases

Sebastian Breß^{1,2*}, Henning Funke², Jens Teubner², Volker Markl^{1,3}

DFKI GmbH, Intelligente Analytik für Massendaten, Alt-Moabit 91c, 10559 Berlin¹
Technische Universität Dortmund, FG DBIS, Otto-Hahn-Straße 14, 44227 Dortmund²
Technische Universität Berlin, FG DIMA, Einsteinufer 17, 10587 Berlin³
`sebastian.bress@dfki.de`, `henning.funke@tu-dortmund.de`,
`jens.teubner@tu-dortmund.de`, `volker.markl@tu-berlin.de`

Abstract. Technology limitations are making the use of *heterogeneous computing devices* much more than an academic curiosity. In fact, the use of such devices is widely acknowledged to be the only promising way to achieve application-speedups that users urgently need and expect. However, building a robust and efficient query engine for heterogeneous co-processor environments is still a significant challenge.

In our latest work [1], we identify two effects that limit performance in case co-processor resources become scarce. *Cache thrashing* occurs when the working set of queries does not fit into the co-processor’s data cache, resulting in performance degradations up to a factor of 24. *Heap contention* occurs when multiple operators run in parallel on a co-processor and when their accumulated memory footprint exceeds the main memory capacity of the co-processor, slowing down query execution by up to a factor of six.

We propose solutions for both effects. *Data-driven operator placement* avoids data movements when they might be harmful; *query chopping* limits co-processor memory usage and thus avoids contention. The combined approach—*data-driven query chopping*—achieves robust and scalable performance on co-processors. We validate our proposal with our open-source GPU-accelerated database engine CoGaDB and the popular star schema and TPC-H benchmarks.

Acknowledgments. The work has received funding from the Deutsche Forschungsgemeinschaft (DFG), Collaborative Research Center SFB 876, project C5, from the European Union’s Horizon2020 Research & Innovation Program under grant agreement 671500 (project “SAGE”), and by the German Ministry for Education and Research as Berlin Big Data Center BBDC (funding mark 01IS14013A).

References

1. S. Breß, H. Funke, and J. Teubner. Robust query processing in co-processor-accelerated databases. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 1891–1906. ACM, 2016.

* Work done when author was working at TU Dortmund.

On the Evaluation of Outlier Detection: Measures, Datasets, and an Empirical Study Continued

Guilherme O. Campos¹, Arthur Zimek², Jörg Sander³,
Ricardo J. G. B. Campello¹, Barbora Micenková⁴, Erich Schubert⁵,
Ira Assent⁴, and Michael E. Houle⁶

¹ University of São Paulo, {gocampos,campello}@icmc.usp.br

² University of Southern Denmark, zimek@imada.sdu.dk

³ University of Alberta, jsander@ualberta.ca

⁴ Aarhus University, {barbora,ira}@cs.au.dk

⁵ Ludwig-Maximilians-Universität München, schube@dbis.lmu.de

⁶ National Institute of Informatics meh@nii.ac.jp

The evaluation of unsupervised outlier detection algorithms is a constant challenge in data mining research. Little is known regarding the strengths and weaknesses of different standard outlier detection models, and the impact of parameter choices for these algorithms. The scarcity of appropriate benchmark datasets with ground truth annotation is a significant impediment to the evaluation of outlier methods. Even when labeled datasets are available, their suitability for the outlier detection task is typically unknown. Furthermore, the biases of commonly-used evaluation measures are not fully understood. It is thus difficult to ascertain the extent to which newly-proposed outlier detection methods improve over established methods. We performed an extensive experimental study [1] on the performance of a representative set of standard k nearest neighborhood-based methods for unsupervised outlier detection, across a wide variety of datasets prepared for this purpose. Based on the overall performance of the outlier detection methods, we provide a characterization of the datasets themselves, and discuss their suitability as outlier detection benchmark sets. We also examine the most commonly-used measures for comparing the performance of different methods, and suggest adaptations that are more suitable for the evaluation of outlier detection results.

We present the results from our previous publication [1] as well as additional observations and measures available at the outlier benchmark data repository: <http://www.dbs.ifi.lmu.de/research/outlier-evaluation/>

References

- [1] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. “On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study”. In: *Data Mining and Knowledge Discovery* 30 (4 2016), pp. 891–927. DOI: 10.1007/s10618-015-0444-8.

Finding Trees in Mountains – Outlier Detection on Polygonal Chains

Michael Singhof, Daniel Braun, and Stefan Conrad

Heinrich-Heine-Universität Düsseldorf, Institut für Informatik,
Universitätsstr. 1, 40225 Düsseldorf, Germany
{singhof, braun, conrad}@cs.uni-duesseldorf.de

Abstract. In this work, we present an approach to the detection of outliers in certain polygonal chains. These originate from images of mountains which are first segmented in order to extract the mountains silhouette. In general, the aim of our framework is to recognise the mountain in the image in order to overcome the problem of large amounts of images on the internet that are not tagged and thus cannot be searched in a sensible fashion.

The appearance of outliers in our case is specific by them either being obstacles in the image that are in front of the mountain or they are due to problems during the silhouette extraction step. In this work, we show, how outliers are defined in our context, namely as sub-sequences that are found by a double threshold technique. Therefore, we describe how the anomaly scores for single vertices in the polygonal chains are computed via a histogram distance based approach. We also introduce an improved way to compute reference data and outlier scores and show that this changes allow for significant better outlier detection results.

1 Introduction

Social networks and image sharing platforms enable users to easily share their photos, with millions of pictures uploaded every day. However, most of these photos are not properly tagged. Because of this, they are not easily accessible since they cannot be simply searched for. The aim of our framework presented in [3] therefore is, to be able to recognise mountains in a given image. We hope to overcome this problem for the example of mountain recognition by being able to automatically tag them.

Since our idea is to be able to annotate every image of a mountain, even without GPS information, we have to extract much more precise silhouettes than approaches that will work on GPS tagged images, only. Therefore, it is of imminent importance to correct errors that occur during the initial segmentation step. In this work, we present an enhancement to our outlier detection step from [3], that uses clustering on the reference silhouettes in order to enable a more precise outlier detection.

2 Related Work

There are approaches for automatic tagging, which provide solutions for other motifs than mountains [13]. However, this work is based on image features, which are very similar for different mountains and not feasible for mountain recognition.

The area of mountain recognition is rather small. Baatz et al. [1] use human interaction to correct silhouettes that are first computed by an algorithm. They also have released their data set, which is used in this work. However, our approach is intended to be able to recognise a mountain without human intervention. Other approaches to mountain annotation, such as [2, 6, 12], rely on GPS tags in order to estimate the position of the mountain in the picture. Thus, the number of mountains to compare with is relatively small and because of this, the extraction of the silhouette from the mountain does not have to be as exact as in our case. To our knowledge, our approach is the only approach that uses adaptive correction techniques in order to enhance the quality of the extracted silhouette.

A silhouette is essentially a two dimensional polygonal chain. However, there is not much work on outlier detection on polygonal chains. A closely related problem is outlier detection in time series. There are many approaches to this, such as Hot Sax [8] and its variants [11, 4, 9]. These approaches search for discords rather than outliers in our sense. A discord is defined as the sub-sequence of a given time series that has the greatest distance to all other sub-sequences of the same length. These methods have advantages in cases, where periodic events are measured with a time series, since the length of the discord is given as a parameter. To our problem, this is not directly applicable because neither do we know the length of outliers in the silhouette nor do we want to find the most unusual outlier, only.

Another, but less closely related, approach to outlier detection in time series is change point detection [5, 7]. A change point is a point in a time series, that is unusual for the part of the time series up until that point but introduces a new behaviour to the time series that might be repeated later on. While this method is useful on time series it cannot be adapted easily to polygonal chains, since the latter do not have a strong time dimension where one can assume that earlier events affect later ones.

3 AdaMS Framework and Mountain Identification

In this section we shortly introduce our adaptive mountain silhouette (AdaMS) framework that initially has been presented in [3] to put the outlier detection task presented in this work into perspective. The task of mountain recognition can be divided in two large parts: first, the silhouette extraction from an input image and second, the identification of that silhouette by mapping it to a known silhouette of a mountain.

AdaMS is our approach to solve the first of these problems. It uses a grid based segmentation algorithm for the initial silhouette extraction. This silhouette

is then, as a second step, searched for outliers which then get classified. Regions around correctable outliers are then resegmented with a different parameter set. The resulting silhouette is again searched for outliers, until no more correctable outliers are found or a maximum number of iterations has been reached.

The identification of the silhouette itself can be divided into two problems, namely the mapping to a known silhouette and the creation of the reference data, i.e. the known silhouettes. Due to the great number of mountains, the mapping of an extracted silhouette to a labelled silhouette has to be precise and with as little computing cost as possible. Therefore it seems advisable to use a step wise process that uses fast to compute distance measures in a first step in order to exclude reference data with low resemblance. As the number of relevant silhouettes decreases, more precise similarity measures can be applied.

The creation of the reference data is a task, that could theoretically be carried out by humans. However, that would be an immense workload because one labelled silhouette per mountain would not suffice to support a reliable identification. This is because mountains look different from different angles. It seems much more feasible to automatically extract the reference silhouettes from digital elevation maps.

4 Enhancing Outlier Detection in AdaMS

For finding outliers, we first have to define outliers for our application case. Based on this definition, we show the basic AdaMS outlier detection and then introduce recent improvements.

4.1 Outlier Definition

A silhouette is a two dimensional polygonal chain that can be converted to a relative silhouette:

Definition 1 A relative silhouette $RS = (v_1, \dots, v_n)$, $n > 0$, is a polygonal chain with $v_i = (l_i, a_i)$ for all $1 \leq i \leq n$, where $l_i > 0$ is the length of a line segment and $a_i \in (-180^\circ, 180^\circ]$ is the angle relative to the x-axis. [3]

On such a relative silhouette we want to find unusual parts that are caused by obstacles such as trees or segmentation faults due to low contrast. So an outlier $o = (v_i, \dots, v_j)$, $1 \leq i < j \leq n$, is a part of a relative silhouette, where the combination of vertices v_i, \dots, v_j does not fit the usual patterns in relative silhouettes extracted from mountain images. This is introduced formally in the following definition.

Definition 2 Let $l > 0$, and $RS = (v_1, \dots, v_n)$ be a relative silhouette.

We call $o = (v_i, \dots, v_j)$ an l -outlier if the following is true:

1. For all v_k , $i \leq k \leq j$, it holds that v_k is a weak anomaly.
2. There exist $m_1, m_2 \in \{i, \dots, j\}$ such that $m_2 - m_1 \geq l$ and for all v_k , $m_1 \leq k \leq m_2$, it holds that v_k is a strong anomaly. [3]

An outlier $o = (v_i, \dots, v_j)$ is called a maximum l -outlier if and only if neither (v_{i-1}, \dots, v_j) nor (v_i, \dots, v_{j+1}) are l -outliers.

Definition 2 mentions strong and weak anomalies. These, in contrast to outliers, are single vertices that have unusual properties, that is, high anomaly scores. We therefore show how these concepts can be defined. Anomaly scores are computed via histograms of parts of relative silhouettes.

Definition 3 Given a relative silhouette RS , then $H_{RS}(s, l)$ denotes the histogram consisting of the points v_s, \dots, v_{s+l-1} of RS and H_{RS} denotes the histogram over all vertices of RS . [3]

Based on these histograms and the distance to the reference histogram H_{ref} , we are now able to introduce the vertices' anomaly scores:

Definition 4 Given a relative silhouette $RS = (v_1, \dots, v_n)$ and a reference histogram H_{ref} .

The anomaly score

$$an(v_i) := \frac{1}{l} \sum_{j=i-l+1}^i d_j$$

of vertex v_i is the average of the distances $d_j = \text{dist}(H_{RS}(j, l), H_R)$.

With the anomaly scores we can now finally introduce the different kinds of anomalies used in definition 2:

Definition 5 Let $RS = (v_1, \dots, v_n)$ be the silhouette of an image with corresponding anomaly scores $an(v_i)$ for vertex v_i , reference anomaly score distribution mean μ and standard deviation σ and two thresholds $0 < \tau_{out} < \tau_{in}$.

Then we call v_i a weak anomaly if

$$an(v_i) \geq \mu + \tau_{out} \cdot \sigma$$

and a strong anomaly if

$$an(v_i) \geq \mu + \tau_{in} \cdot \sigma. [3]$$

This shows, that an outlier in our case is an application of a double threshold technique. The idea here is that often, within an obstacle or a segmentation fault, only small parts of the outlier consist of vertices with unusually high anomaly scores. The rest of the outlier consists of vertices whose anomaly scores are still high, but on their own would not suffice to identify an outlier. Figure 1 shows such an example.



Fig. 1. An outlier – strong anomalies are marked red, weak pink and the silhouette yellow.

4.2 SingleRef Outlier Detection and Reference Data Computation

As in [3], our first version outlier detection algorithm computes the vertices anomaly via a sliding window approach, following the idea outlined in the previous section. More than one window length can be used and the same window length may be used multiple times to induce a weighting to the distances computed by the different window lengths. This is necessary, because a longer window length gives a single vertex more distance scores than a shorter one and thus is represented stronger in the anomaly score.

Given the anomaly values $an(v_i)$ for the vertices, finding the maximum l -outliers according to definition 2 is straightforward. In the first step, we search for sequences of vertices with a length of at least l that all are strong anomalies. Once we have found such a sequence, we let it grow by adding vertices that are neighbours of the currently detected outlier and weak anomalies on both sides.

In regard to reference data computation, the algorithm gets a selection of outlier free silhouettes that were chosen by hand and computes a single reference histogram H_{ref} of all those silhouettes. We therefore refer to this version of the algorithm as SingleRef outlier detection. The statistical properties μ and σ are then computed by using the same window lengths as in the actual outlier detection to compute the anomaly score of every vertex in the reference silhouettes.

4.3 MultiRef – Improving SingleRef

Computing one histogram, in a certain manner, aggregates the data represented by the histogram. Due to the differences in details, a stronger aggregation results in higher standard deviation of the single data points to this aggregates than

a weaker aggregation. This, in turn, leads to obscured real anomalies, since distances to the reference data are in general rather high. As the evaluation chapter shows, this leads to either high rates of false positives or relatively low detection rates for SingleRef.

A solution to this problem would be the usage of multiple reference histograms. Intuitively, one would choose one histogram per silhouette, resulting in Histograms $H_{ref}^1, \dots, H_{ref}^n$ if we assume n reference silhouettes. In order to use more than one reference histogram, however, we have to adjust the distance computation. The base distance used in AdaMS is the following:

Definition 6 *Let $G = (g_1, \dots, g_n)$, $H = (h_1, \dots, h_n)$ be histograms with n buckets.*

The above average distance of G to H is defined by

$$\text{dist}(G, H) := \max(|\text{aab}(G)|, |\text{aab}(H)|) - |\text{aab}(G) \cap \text{aab}(H)|,$$

where

$$\text{aab}(F) := \left\{ i \in \{1, \dots, n\} \mid f_i \geq \frac{1}{n} \sum_{i=1}^n f_i \right\}$$

with $F = (f_1, \dots, f_n)$ being a histogram with n buckets.

Essentially, by the above average distance, the number of above average buckets that are the same in both histograms is subtracted from the higher number of above average filled buckets.

Theorem 7 *The above average distance from definition 6 is a pseudometric.*

For the proof of this theorem see appendix A.

Now, when comparing a histogram with not just one reference histogram but several, we compute the above average distance to all reference histograms and then choose the minimum of that distances.

Definition 8 *Given a histogram H and reference histograms $H_{ref}^1, \dots, H_{ref}^n$, then*

$$\text{dist}_{mr}(H) := \min_{1 \leq i \leq n} \{ \text{dist}(H, H_{ref}^i) \}$$

is called the min-ref distance.

As the number of reference silhouettes is potentially large and it is beneficial to add further silhouettes free of outliers to the reference data, it is clear that the usage of a reference histogram per reference silhouette is not feasible and the number of histograms has to be reduced. On the other hand, we want to minimise the loss of detail due to aggregation. We therefore utilise a k -means clustering [10] on the reference silhouettes' histograms.

By this, we are able to reduce the number of reference silhouettes to any given k while ensuring, that we loose as little detail as possible, because the subadditivity holds for the above average distance. This means, the distance

to the cluster representative is an upper bound on the distance to the closest silhouette’s histogram. We ensure small distances between representatives and members of the cluster by clustering with random start representatives 1000 times and choosing the clustering with the smallest quadratic distance. It is also noteworthy, that by setting $k = 1$ we only get one reference silhouette that is identical to the reference silhouette of the SingleRef method.

By this, as the evaluation shows, we are able to achieve better outlier detection without changing our outlier detection algorithm as such nor do we need more reference data.

4.4 Further Steps

The next steps after identifying the outliers are to classify them into obstacles in the picture and errors in the segmentation step, for example due to low contrast between parts of the mountain and the sky. As described in [3], we use four classes of outliers, namely obstacles, segmentation errors where the silhouette is too high, segmentation errors where the silhouette is too low and false positives. For the classification we utilise a k -nearest neighbour approach on the outliers’ histograms.

5 Evaluation

In order to evaluate the approaches introduced in the previous section, we manually annotated a test set of 111 outliers from 14 silhouettes, that, in total, consist of 3580 vertices. The silhouettes have been automatically extracted from the images by a variant of the the segmentation algorithm presented in [3], but without the outlier detection steps and therefore without the adaptive correction.

The outlier detection algorithm has been trained with 48 silhouettes that are mostly free of outliers and have been extracted with the same mechanism as described above. These have been clustered 1000 times per number of clusters and the clustering with the lowest overall quadratic distance of histograms in respect to their cluster representative has been chosen.

We first evaluated the precision and recall based on detected outlier vertices for different values for k , the number of reference silhouette clusters and thus reference histograms. As parameter set we chose the minimum length for an inner outlier $l = 3$. The inner and outer thresholds have been set to $\tau_{in} = 2$ and $\tau_{out} = 1$. The results of this are shown in table 1. Note here, that for $k = 48$ no clustering is used, but here one reference histogram per reference silhouette is used.

Precision and recall are computed by counting the vertices that have been declared as parts of outliers correctly and dividing that number by the total number of detected outlier vertices respectively the total number of annotated outlier vertices. The last number is shown in the last row of the table. It can be seen here, that MultiRef, for every tested value of k shows better results than SingleRef. Especially recall is much better than with SingleRef and gets higher

Method	Value	Silhouette														Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	
SingleRef	Precision	86	0	18	0	0	0	82	0	58	0	27	98	48	90	72
	Recall	54	0	15	0	0	0	65	0	25	0	5	43	47	46	33
$k = 3$	Precision	87	0	69	42	0	42	79	22	51	88	52	95	72	86	73
	Recall	66	0	41	29	0	66	72	20	64	8	17	58	69	47	46
$k = 5$	Precision	82	0	68	37	0	70	82	43	44	80	56	89	73	89	74
	Recall	68	0	39	37	0	66	71	43	52	19	28	65	75	62	54
$k = 7$	Precision	89	0	57	40	0	100	82	42	45	79	61	97	71	89	77
	Recall	68	0	17	28	0	60	71	40	50	19	27	53	68	58	49
$k = 10$	Precision	85	0	59	41	0	100	78	42	45	79	73	95	72	89	77
	Recall	65	0	18	29	0	60	71	40	51	19	28	60	63	56	49
$k = 20$	Precision	88	0	57	40	0	65	79	39	45	82	74	94	72	89	74
	Recall	68	0	18	37	0	64	88	41	52	18	28	49	70	56	49
$k = 30$	Precision	84	0	78	41	0	65	89	41	46	76	61	89	76	90	76
	Recall	70	0	59	48	0	64	91	39	59	24	38	68	79	71	61
$k = 48$	Precision	84	0	78	41	0	65	71	45	46	67	66	89	78	91	75
	Recall	70	0	59	48	0	64	91	39	59	24	38	69	79	71	61
#Outlier vertices		305	30	71	234	46	47	144	121	147	344	192	682	315	902	3580

Table 1. Evaluation on detected outlier points. Precision and Recall are given in percent. Thresholds are $\tau_{in} = 2$ and $\tau_{out} = 1$, $l = 3$.

overall with increasing k . Interestingly, while there is a huge improvement in recall from $k = 20$ to $k = 30$, recall does not get any better when using one histogram per reference silhouette. Precision, too, is higher for the MultiRef variants in respect to SingleRef, but it is noteworthy here, that, instead of rising with the number of clusters, a maximum is reached for $k = 7$ respectively $k = 10$ and for bigger values, a slight decrease can be noticed. Note here, that the results in table 1 are given without carrying out the classifying of outliers. By that step, some outliers will be classified as false positives and thus precision after classification should increase.

Table 1 also shows, that for silhouettes 2 and 5, no correct outliers have been found at all. The outliers in both silhouettes are rather short. The three outliers in silhouette 2 have a length of 8, 9 and 13 vertices and the four outliers in silhouette 5 have a length of 9, 10, 12 and 15 vertices.

Based on this observation, table 2 shows the number of outliers that have been hit by the detection algorithm. The same parameter set as above has been used for this. An outlier is counted as being hit by the detection algorithm, if at least one vertex of it has been detected as being part of an outlier. In context with AdaMS, if this happens, the silhouette will be recomputed in this region and can thus be corrected. It is clear, that all algorithms have problems with smaller outliers. For the extraction of good silhouettes, it is more important to find huge outliers, and our detection rates for those are promising, in general. The detection rate increases with the length of the outlier for every variant. On

Method	$ out \leq 5$	$5 < out \leq 10$	$10 < out \leq 20$	$20 < out \leq 50$	$50 < out $	Total
SingleRef	0	1	2	5	10	18
$k = 3$	2	5	7	14	12	40
$k = 5$	5	7	9	15	13	49
$k = 7$	4	5	7	14	13	43
$k = 10$	4	5	8	14	13	44
$k = 20$	4	8	8	15	12	47
$k = 30$	6	10	11	16	15	58
$k = 48$	6	10	11	16	15	58
#Outliers	18	38	19	21	15	111

Table 2. Number of detected outliers. The length of outliers is denoted by $|out|$. Thresholds are $\tau_{in} = 2$ and $\tau_{out} = 1$, $l = 3$.

the other hand, with increasing k , the detection rate in general increases, too. Interestingly, for $k = 5$ more outliers have been detected than for surrounding values of k . However, due to the relatively small numbers of outliers used in our evaluation, this might be coincidence.

τ_{in}	τ_{out}	l	$ out \leq 20$	$20 < out \leq 50$	$50 < out $	Total	Precision	Recall
2	1	3	21	15	13	49	74	54
1.5	1	3	25	17	14	56	69	58
2.5	1	3	12	13	12	37	77	49
2	0.75	3	24	15	13	52	71	59
2	1.25	3	20	15	13	48	76	50
2	1	1	21	15	13	49	72	55
2	1	5	17	14	13	44	76	53

Table 3. Effect of parameter changes based on MultiRef with $k = 5$.

The results in table 3 show the results of our investigation of the effects of parameter changes. Essentially, lowering τ_{in} or l results in a greater number of detected outliers, while raising that values reduces the number of outliers. Changes to τ_{out} affect the size of detected outliers. The lower τ_{out} becomes, the bigger are the resulting outliers. Due to overlaps with the real outliers, lowering of one of the values leads to higher recall and decreased precision. Increasing them raises precision but induces losses to recall.

In summary, the results show that even for the worst choice of k , the number of hit outliers is more than two times higher than that detected by SingleRef, while at the same time precision and recall are increased.

6 Conclusion and Future Work

In this work, we have presented our definition of outliers and our approach to make the detection of outliers in polygonal chains that represent silhouettes extracted from pictures of mountains more effective. Our results show, that the MultiRef variant introduced in this work greatly improves the outlier detection results. The number of detected outliers is increased by a factor of two to three, depending on the number of clusters, while precision and recall are also increased.

However, there are some points that we want to address in the future. One of the main questions is, whether our distance function as given in definition 6 is ideal or if a more elaborate histogram distance function such as the earthmover's distance [14] will yield better results. Also, we plan to use additional data for the outlier detection. The contrast strength seems to be a good measure, since segmentation faults usually occur in regions of low contrast.

A Appendix

The fact of the above average distance being a pseudometric is of importance since it ensures that the triangle equation is satisfied by that construct. From this it can be derived that the greater the distance between two histograms is, the greater the difference between those is and there are no short cuts by using intermediate histograms.

Proof. In order to show that the above average distance is a pseudometric, four properties have to be shown. Let g, h, k be histograms with the same number of buckets.

Non-negativity $\text{dist}(g, h) \geq 0$. This is trivial since $|aab(g)| \geq |aab(g) \cap aab(h)|$ and $|aab(h)| \geq |aab(g) \cap aab(h)|$, thus

$$\max(|aab(g)|, |aab(h)|) \geq |aab(g) \cap aab(h)|.$$

Identity of indiscernibles

$$\begin{aligned} \text{dist}(g, g) &= \max(|aab(g)|, |aab(g)|) - |aab(g) \cap aab(g)| \\ &= |aab(g)| - |aab(g)| = 0. \end{aligned}$$

Symmetry

$$\begin{aligned} \text{dist}(g, h) &= \max(|aab(g)|, |aab(h)|) - |aab(g) \cap aab(h)| \\ &= \max(|aab(h)|, |aab(g)|) - |aab(h) \cap aab(g)| \\ &= \text{dist}(h, g) \end{aligned}$$

since both $\max(\cdot, \cdot)$ and the intersection of sets are symmetric functions.

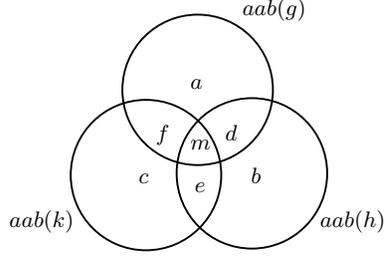


Fig. 2. Sets for proof of subadditivity.

Subadditivity In order to prove the subadditivity an auxiliary construction is necessary. As shown in figure 2, the sets $aab(g)$, $aab(h)$ and $aab(k)$ are split in four disjoint sets each, such that $aab(g) = a \cup d \cup f \cup m$, $aab(h) = b \cup d \cup e \cup m$ and $aab(k) = c \cup e \cup f \cup m$.

Now, without loss of generality, let $|aab(g)| \geq |aab(h)|$. Then $\text{dist}(g, h) = |aab(g)| - |aab(g) \cap aab(h)| = |a| + |f|$.

Case 1 Now, let $|aab(h)| \geq |aab(k)|$. Then it holds that $\text{dist}(k, h) = |b| + |d|$ and $\text{dist}(g, k) = |a| + |d|$, since $|aab(g)| \geq |aab(h)| \geq |aab(k)|$. Thus, in this case

$$\begin{aligned} & \text{dist}(g, k) + \text{dist}(k, h) - \text{dist}(g, h) \\ &= |a| + |d| + |b| + |d| - |a| - |f| \\ &= |b| + 2|d| - |f| \geq 0, \end{aligned}$$

because

$$\begin{aligned} & |aab(h)| \geq |aab(k)| \\ \Rightarrow & |b| + |d| + |e| + |m| \geq |c| + |e| + |f| + |m| \\ \Rightarrow & |b| + |d| \geq |c| + |f| \geq |f|. \end{aligned}$$

Case 2 Assume now, that $|aab(h)| < |aab(k)|$, so $\text{dist}(k, h) = |c| + |f|$.

Case 2.1 Let $|aab(g)| \geq |aab(k)|$. It follows that

$$\begin{aligned} & \text{dist}(g, k) + \text{dist}(k, h) - \text{dist}(g, h) \\ &= |a| + |d| + |c| + |f| - |a| - |f| \\ &= |d| + |c| \geq 0. \end{aligned}$$

Case 2.2 The last case to be considered occurs if $|aab(g)| < |aab(k)|$ which results in $\text{dist}(g, k) = |c| + |e|$. Then

$$\begin{aligned} & \text{dist}(g, k) + \text{dist}(k, h) - \text{dist}(g, h) \\ &= |c| + |e| + |c| + |f| - |a| - |f| \\ &= 2|c| + |e| - |a| \geq 0. \end{aligned}$$

This is because of a similar argument to case 1, because

$$|aab(k)| \geq |aab(g)| \Rightarrow |c| + |e| \geq |a| + |d| \geq |a|.$$

□

References

1. Baatz, G., Saurer, O., Köser, K., Pollefeys, M.: Large Scale Visual Geo-Localization of Images in Mountainous Terrain. In: *Computer Vision - ECCV 2012* (2012)
2. Baboud, L., Čadík, M., Eisemann, E., Seidel, H.P.: Automatic Photo-to-terrain Alignment for the Annotation of Mountain Pictures. In: *Proc. of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (2011)
3. Braun, D., Singhof, M., Conrad, S.: AdaMS: Adaptive Mountain Silhouette Extraction from Images. In: *Proc. of MLDM 2016* (2016)
4. Buu, H.T.Q., Anh, D.T.: Time Series Discord Discovery Based on iSAX Symbolic Representation. In: *Third International Conference on Knowledge and Systems Engineering* (2011)
5. Fawcett, T., Provost, F.: Activity Monitoring: Noticing Interesting Changes in Behavior. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (1999)
6. Fedorov, R., Fraternali, P., Tagliasacchi, M.: Mountain Peak Identification in Visual Content Based on Coarse Digital Elevation Models. In: *Proc. of the 3rd ACM International Workshop on Multimedia Analysis for Ecological Data* (2014)
7. Kawahara, Y., Sugiyama, M.: Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation. In: *Proc. of 2009 SIAM International Conference on Data Mining (SDM2009)*, (2009)
8. Keogh, E., Lin, J., Fu, A.: Hot Sax: Efficiently Finding the Most Unusual Time Series Subsequence. In: *Fifth IEEE International Conference on Data Mining (ICDM'05)* (2005)
9. Khanh, N.D.K., Anh, D.T.: Time Series Discord Discovery Using WAT Algorithm and iSAX Representation. In: *Proceedings of the Third Symposium on Information and Communication Technology* (2012)
10. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In: *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (1967)
11. Pham, N.D., Le, Q.L., Dang, T.K.: HOT aSAX: A Novel Adaptive Symbolic Representation for Time Series Discords Discovery. In: *Asian Conference on Intelligent Information and Database Systems* (2010)
12. Porzi, L., Buló, S.R., Valigi, P., Lanz, O., Ricci, E.: Learning Contours for Automatic Annotations of Mountains Pictures on a Smartphone. In: *Proc. of the International Conference on Distributed Smart Cameras* (2014)
13. Rischka, M., Conrad, S.: Image Landmark Recognition with Hierarchical K-Means Tree. In: *Database Systems for Business, Technology and Web (BTW 2015)* (Mar 2015)
14. Rubner, Y., Tomasi, C., Guibas, L.J.: A Metric for Distributions with Applications to Image Databases. In: *Computer Vision, 1998. Sixth International Conference on. IEEE* (1998)

SemRes: A System for Creating and Searching Semantic Documentation for Conservators.

Ernesto William De Luca^{1,2}

¹ Georg-Eckert-Institute. Leibniz-Institute for international Textbook Research.
Celler Straße 3. D-38114 Braunschweig

² Potsdam University of Applied Sciences. Kiepenheuerallee 5. 14469 Potsdam,
Germany

Abstract. Documentation is very time consuming; conservators have to categorize photos, give detailed description of the procedures and justify decisions about materials. Every conservator delivers his/her own documentation to the respective monument administrations that archive it in their shelves, without any possibility to access them digitally. In this case, the access to already conducted procedures, as well as the provision of the knowledge about the methods and materials used are very difficult, often almost impossible. The SemRes System provides conservators with semantic technologies in order to better structure and retrieve their knowledge and supports them in decision-making for conservation and conservation procedures.

Keywords: Information Retrieval, Semantic Retrieval, Linked Open Data, Digital Humanities

1 Introduction

In conservation the documentation is an integral part of the restorers work to describe their restored objects. It is important and indispensable to understand the phases of conservation and to allow decision-making for conservation and conservation procedures. The information they contain about the conducted conservation procedures, as well as the methods and the materials used are an immeasurable source of knowledge for this discipline and build the basis of the knowledge used for the conservation of a given object.

On the other hand, ontologies are the central components of semantic tools and have been present in the art and cultural area for some time. For instance, the CIDOC Conceptual Reference Model (CRM) supports the integration, communication and the exchange of differently structured information from the cultural heritage field. The field of conservation is, in this context, still underrepresented. The only semantic contribution in this domain is the Australian project The Twentieth Century in Paint¹, which investigates paint materials used in art works of the 20th Century in the Asian-Pacific region. One goal of the project

¹ <http://www.20thcpaint.org/>

is to develop a suitable ontology. This was developed for a specific sub-domain and can be reused, but represents only a very small area of conservation. To develop a generic conservation ontology, we have to consider the numerous measures and materials, as well as the damage patterns and causes of loss, which can vary greatly depending on the materiality and exposure of the object to be restored. To achieve a high acceptance in the professional world, conservation experts must be involved in this development process. Ontologies in the field of conservation documentation have not yet been developed. This is a big gap in the research infrastructure of conservation and other fields of cultural heritage.

2 The SemRes System

The presented demonstrator² bases on an own ontology for conservation documentation that includes all these semantic relations. The ontology is developed as a basis for semantic processing of documentation for conservators. The Triple-Store is based on Erlangen CIDOC Conceptual Reference Model (CRM³) and has been extended with VIAF⁴ and Geonames⁵.

This research is a joint effort of conservators, information and computer scientists in order to help conservators in organizing and retrieving their own documentation semantically. It aims to build on European initiatives and efforts that seek to free access of all cultural heritage information. The prior objective of this work is the preparation of data for the restorers that are automatically linked in the Linked Open Data Cloud. In this way, the access and reuse of scientific data on restoration is made possible not only for conservators, but also for architects, historians, archeologist and interested people. The SemRes system uses the information stored in a relational database and in a triple store that are mapped to another. The triples that have been automatically built represent the relations between artifact (object), the materials, the date of origin and the conservation phases. These semantic relations allow conservators to search and retrieve documentation about a given object that is semantically related to another one. We can find other documentations using the semantic relations through conservators, or materials that have been already used for some conservation phases or through the date of origin that could show how objects are historically related to another. The conservation phases show how documentations (that could have been written by different conservators) are related to each other and can be retrieved and semantically connected. In the final demonstration, we will show in more details how the ontology has been built and will give some more concrete examples, in order to make clear how the system uses the information stored in the relational database and in the triple store. The resources will be presented, as well as the classification of 539 material classes (related mostly to stone conservation).

² <http://semres.de/>

³ <http://erlangen-crm.org/current-version/>

⁴ <http://viaf.org/>

⁵ <http://www.geonames.org/>

Case-Based Decision Support on Diagnosis and Maintenance in the Aircraft Domain

Pascal Reuss^{1,2} and Klaus-Dieter Althoff^{1,2} and Wolfram Henkel³
pascal.reuss@dfki.de
klaus-dieter.althoff@dfki.de
Wolfram.Henkel@airbus.com

¹Intelligent Information Systems Lab, University of Hildesheim

²Competence Center Case Based Reasoning, German Center for Artificial Intelligence,
Kaiserslautern

³Airbus Operations GmbH, Hamburg

Abstract. Aircraft diagnosis is a highly complex topic. Many knowledge sources are required and have to be integrated into a diagnosis system. This paper describes the instantiation of a multi-agent system for case-based aircraft diagnosis based on the SEASALT architecture. This system will extend an existing rule-based diagnosis system, to make use of the experience on occurred faults and their solutions. We describe the agents within our diagnosis system and the knowledge modeling for the case-based reasoning systems. In addition we give an overview over the current implementation.

1 Introduction

Technical domains can be very complex and the diagnosis of machines belonging to these domains is not easy. In most cases, dozens of relations between individual parts have to be considered and a high number of constraints can have an effect on the measurable symptoms and the diagnoses. An aircraft is one of the most complex machines built by humans and therefore the diagnosis and maintenance of aircraft is very resource consuming. Finding the root cause for an occurred fault may cost several hours, because it can be caused by a single part, the interaction between parts, or even the communication infrastructure between these parts. For example, if a monitor does not display the status of a system, the monitor could be broken, one or more parts of the system sending incorrect or no information, or the communication cable to the monitor may be broken. Therefore the use of experience from past faults can be very helpful to get a more precise diagnosis and reducing the time for finding and repairing the root cause. In this paper we describe an approach to use Case-Based Reasoning (CBR) within a decision support system to integrate experience knowledge into the existing diagnosis approach. The decision support system is a multi-agent system (MAS) that contains several CBR systems to provide experience knowledge. In the next section we give an overview of the OMAHA research project and the SEASALT architecture. Section 2 contains related work to our approach and Section 3 describes the multi-agent system for decision support in more detail. The last Section gives a summary and an outlook on ongoing and future work.

1.1 OMAHA project

The decision support system is under development in the context of a research project called OMAHA. The project tries to develop an **Overall Management Architecture for Health Analysis** for civilian aircrafts. Several topics are addressed within the project like diagnosis and prognosis of flight control systems, innovative maintenance concepts, and effective methods of data processing and transmission. A special challenge of the OMAHA project is to integrate not only the aircraft and its subsystems, but also systems and processes in the ground segment like manufacturers, maintenance facilities, and service partners into the maintenance process. Several enterprises and academic and industrial research institutes take part in the OMAHA project: the aircraft manufacturer Airbus, the system manufacturers Diehl Aerospace and Nord-Micro, the aviation software solutions provider Linova and IT service provider Lufthansa Industrial Solutions as well as the German Research Center for Artificial Intelligence and the German Center for Aviation and Space. In addition, several universities are included as subcontractors. The project started in 2014 and will last until the end of March, 2017. [7]

1.2 SEASALT architecture

The SEASALT (**Shared Experience using an Agent-based System Architecture LayouT**) architecture is a domain-independent architecture for extracting, analyzing, sharing, and providing experiences [4]. The architecture is based on the Collaborative Multi-Expert-System approach [1],[2] and combines several software engineering and artificial intelligence technologies to identify relevant information, process the experience and provide them via an interface. The knowledge modularization allows the compilation of comprehensive solutions and offers the ability of reusing partial case information in form of snippets. The SEASALT architecture consists of five components: knowledge sources, knowledge formalization, knowledge provision, knowledge representation, and individualized knowledge. The *knowledge sources* component is responsible for extracting knowledge from external knowledge sources like databases or web pages and especially Web 2.0 platforms. The *knowledge formalization* component is responsible for formalizing the extracted knowledge into a modular, structural representation. The *knowledge provision* component contains the so-called Knowledge Line. The basic idea is a modularization of knowledge analogous to the modularization of software in product lines. The modularization is done among the individual topics that are represented within the knowledge domain. For each topic a so-called Topic agent is responsible. If a Topic Agent has a CBR system as knowledge source, the SEASALT architecture provides a Case Factory for the individual case maintenance [4],[3],[11]. The knowledge representation component contains the underlying knowledge models of the different agents and knowledge sources. The individualized knowledge component contains the web-based user interfaces to enter a query and present the solution to the user.

1.3 Application domain

The aircraft domain is a very complex technical domain. An aircraft consists of hundreds of components (e.g, Communication and Ventilation Control), which consist of

dozens of systems (e.g, Cabin Intercommunication System and Air Conditioning), that in turn contain dozens of individual parts (e.g, Flight Attendant Panel and Cabin Air Filter) called Line Replacement Units (LRU). These systems and LRUs are interacting with and rely on each other. Therefore, it is not easy to identify the root cause of an occurred fault, because it can either be found within a single LRU, within the interaction of several components of a system, within the interaction of LRUs of different systems, or even within the communication infrastructure of different LRUs. Finding cross-system root causes is a very difficult and resource expensive task. The existing diagnosis system onboard an aircraft can track root causes based on causal rules defined for the LRUs. These rules are not always unambiguous, because the diagnosis approach is effect-driven. Based on a comprehensible effect (visible, audible, or smellable) in the cockpit, the cabin, or the cargo bay, the diagnosis system tries to determine the system behavior that belongs to the effect and traces the root cause back through the defined rules. Based on the error messages and the identified root causes, so-called PFR items are created. This PFR items contains up to 50 error messages, that are associated with the same root cause. For each error messages up to three LRUs could be accused. This means, a PFR item could contain 150 different LRUs in the worst case. The mechanic at the aircraft has to check the LRUs until he finds the fault. The use of CBR for the diagnosis can help to clear ambiguous diagnosis situations with the help of experience knowledge from successfully solved problems, especially with cross-system root causes. At least CBR could help to ranked the LRUs based on experience. Therefore, we are developing a decision support system based on multiple software agents and CBR systems. The multi-agent system will not replace the existing rule-based system, but will extend the diagnosis approach to confirm or disagree with a diagnosis and the associated root cause.

In the next section we present some related work and compare it to our approach. In Section 3 we describe the problem and the use case of our case-based decision support system (Section 3.1), the instantiation of the multi-agent system based on the SEASALT architecture (Section 3.2), and the case structure and the similarity modeling of our CBR systems (Section 3.3). Finally we give a summary and an outlook on ongoing and future work of the decision support system.

2 Related work

Decision support for diagnosis (and maintenance) in the aircraft domain means that a lot of engineering knowledge is available to support this process. In the past various diagnostic approaches tried to improve diagnosis and maintenance in this domain: among others CBR, rule-based reasoning, model-based reasoning, Bayesian belief networks, Fuzzy inference, neural networks, fault trees, trend analysis, and a lot of combinations. For OMAHA, that is OMAHA work package 230, the exploitation of available experiences as supplementation to other already used knowledge sources is of high priority. See also the work of Reuss et al.[11] for relating our approach with a selection of related other experience reusing diagnostic approaches: the British research project DAME [10] dealing with fault diagnosis and prognosis based on grid computing , Dynamic CBR [13] learning also through statistic vectors containing abstract knowledge

condensed from groups of similar cases, and the hybrid approach of Ferret and Glasgow [9] combining model-based reasoning and CBR.

In addition to other specific characteristics of our approach one property differentiating it from many other (CBR) approaches is the fact that we develop a multi-agent system that applies a lot of CBR agents (among other ones). The following approaches have in common that they also combine a multi-agent system approach with CBR. Some researchers also dealing with CBR from different perspectives and trying to combine the specific insights to an improved overall approach are [15]. Of course, what makes our approach different here is that we are concerned with the development of a concrete framework with existing applications. Corchado et al.[8] present in their work an architecture for integrating multi-agent systems, distributed services, and an application for constructing Ambient Intelligence environments. Besides addressing a different domain and task this approach appears to be more open concerning the potential tasks agents can take over, while our approach is more focused in applying software engineering strategies for decomposing problems into sub-problems resulting in a distributed knowledge-based system. Zouhaire and his colleagues[16] developed a multi-agent system using dynamic CBR that learns from traces and is applied for (intelligent) tutoring. Our approach does not learn from traces but instead has to deal with a lot of technical knowledge and in addition has to solve critical problems. Srinivasan, Singh and Kumar[14] share with our approach that they develop a conceptual framework for decision support systems based on multi-agent systems and CBR systems. Our approach appears to be more on the side of integrating software engineering and artificial intelligence methods implementing concrete application systems, while the authors discuss how their framework influences decision support system in general.

3 Case-based decision support for diagnosis and maintenance

In this section we describe our approach to use experience knowledge through CBR systems to enhance the diagnosis and maintenance of aircraft. The decision support system is based on the SEASALT architecture and we present the instantiation of the components in the context of our system. We also describe the problem that should be addressed by our approach and the application use case. In addition, we present the case structure and similarity measures of our CBR systems.

3.1 Problem description and use case

The current diagnosis approach of Airbus aircraft is effect-driven. A Central Maintenance System (CMS) tries to correlate failure messages from LRUs to effects like red blinking lamps or a displayed message by using causal rules and time data. For every failure message a fault item is generated. The CMS tries to correlate new failure messages and effects to open fault items. If they can not be correlated, a new failure item is created. For each failure item, a root cause is determined. But the rules can determine several different root causes for a given failure item. This way a fault item can have up to ten root causes and for every root cause up to three LRUs can be accused. In the worst case thirty possible cases have to be considered when repairing the fault. The

maintenance technician uses his experience to filter the list of root causes and LRUs to identify the most probable starting point. Because not every technician has the same experience, the use of CBR to store and share experience and enhance the diagnosis with this experience could be a viable way to improve diagnosis and maintenance with a decision support system.

Several use cases of the decision support system were discussed in context of the OMAHA project: application in daily operations in the Operations Control Center for diagnosis during flight, application in the Maintenance Control Center to support unscheduled maintenance tasks, and application at Line Maintenance at the aircraft to support the maintenance technician directly. The second use case is identified as a viable use case for application of the system in the project context. The Customer Service of Airbus, as a part of the Maintenance Control Center (MCC), can use the decision support system to recommend maintenance actions based on successfully applied maintenance actions to similar problems in the past. The system will not replace the existing rule-based diagnosis, but will enhance it with experience knowledge to identify a root cause and the responsible LRU quicker and more precisely.

3.2 SEASALT instantiation within OMAHA

Our multi-agent decision support system is an instantiation of the SEASALT architecture. We describe the software agents in the individual components and their tasks within the decision support workflow. The central component of our system is the *knowledge provision*, where the Knowledge Line is located. The Knowledge Line is responsible for retrieving similar problems for a given fault situation and providing a diagnosis and the performed maintenance actions. Therefore several software agents are used to receive a query and retrieve a solution. A communication agent receives the input from a user and sends it to the coordination agent. The coordination agent is responsible for distributing the query to the relevant topic agents. Each topic agent has access to a CBR system, performs a retrieval, and delivers the found cases to the coordination agent. The knowledge is distributed among the CBR systems of the topic agents and is discriminated between aircraft types (e.g. A320, A350, or A380) and aircraft systems (cabin, ventilation control, hydraulic). Each system is identified by the so-called ATA chapter, a number of four or six digits. This way, one CBR system contains cases for A320 cabin faults, another CBR system contains cases for A380 ventilation control faults. The approach of having individual agents for each aircraft type and ATA chapter is based on the idea to split the knowledge among CBR systems to decrease the modeling, retrieval, and maintenance effort for each CBR system. The coordination agent decides which topic agents are required to find a solution for the query. This decision is based on the aircraft type and the ATA chapter. Nevertheless, the cases in the other case bases may contain useful information as well. Especially when the primary topic agents cannot provide a sufficient solution. Therefore the query can be distributed to the other topic agents as well, because faults and their maintenance recommendations may be similar in different aircraft types. The last agent in the *knowledge provision* is the so-called query analyzer agent. This agent is responsible for analyzing the query and identifying new concepts, which are not part of the vocabulary of the CBR systems. If any new concepts are found, a maintenance request is sent to a Case Factory[11].

The Case Factory derives appropriate maintenance actions and notifies a knowledge engineer about the changes. To analyze the query and performing the derived changes, parts of a workflow for knowledge transformation are used. These tasks combine natural language processing techniques and CBR mechanisms to identify new knowledge and transform it to be used by the CBR systems[12]. The user interface is located in the *individualized knowledge* component. It is a web interface, which provides options to send a query, perform a retrieval, present the solutions, enter new cases, and browse the case bases. Another interface in the component links to a data warehouse, where fault information is stored, which can be used as input for our decision support system. Via the interface, a query can be received and the solutions sent back to the data warehouse. The *knowledge formalization* component transforms structured, semi-structured, and unstructured data into a modular, structural knowledge representation used by all CBR systems. This way the knowledge is represented in the same way all over the multi-agent system. The complete version of the workflow for knowledge transformation is used by a so-called case base input analyzer. The complete workflow consists of eight steps: At first, information extraction methods are used to extract keywords and collocations and to find synonyms and hypernyms for the extracted keywords. Then the input data is analyzed to find associations within the allowed values of an attribute as well as across different attributes. This way we want to generate completion rules for query expansion. The keywords, synonyms, hypernyms, and collocations are added to the vocabulary and initial similarity values for keywords and their synonyms are set. The keywords and their hypernyms can be used to generate taxonomies for similarity measures. After the vocabulary extension, cases are generated from the input data and stored in the case bases. The last step is to perform a sensitivity analysis on the stored cases to determine the weighting for the problem description attributes[12]. In the *knowledge sources* component a collector agent is responsible for finding new data in the data warehouse, via web services or in existing knowledge sources of Airbus. New data could be new configurations or operational parameters, new synonyms or hypernyms, or complete new cases. The *knowledge representation* component contains the generated vocabulary, similarity measures and taxonomies, completion rules, and constraints provided for all agents and CBR systems.

3.3 Knowledge modeling and implementation

Based on our initial data analysis at the beginning of the OMAHA project, we decided to use the structured CBR approach and present the knowledge as attribute-value pairs. Much knowledge is stored in databases or CSV files and has a unique column-value correlation. Therefore it can easily be transformed into attribute-value pairs. But during the progress of the project, knowledge in form of free text became more and more important, because these free texts contain many relevant experience. A pure textual CBR approach is not viable, because the structured information is also important for a diagnosis. Therefore an approach was required that considers the structured information as well as the free texts. Because the knowledge from the cases in our CBR systems should be stored in the data warehouse as well, we decided to stay at the structured CBR approach and try to transform the information of the free text into a structured representation. We modeled a case structure with 68 attributes and distributed them

among problem description, solution, quality information, and additional information. The problem description of a case consists of 22 attributes like aircraft type, aircraft model, ATA chapter, fault code, and engine type. Seven attributes are modeled for the problem description extracted from free texts: system, function, status, location, time amount, time unit, and complete description. We use the workflow for knowledge transformation to analyze a free text and map the found information to the attributes. Splitting the information over several attributes allows us to reduce the modeling effort for each attribute, especially the similarity modeling based on taxonomies. All problem attributes are symbolic attributes and are based on a list of allowed values. These values are organized in taxonomies that are used to compute the similarity during retrieval. The current knowledge model contains more than 30.000 different values among all problem attributes. The solution consists of 10 attributes that contain maintenance actions, documentation references, root causes, and comments. The quality information is stored in two attributes that count the number of correct and false retrievals of a case based on user feedback. The other 34 attributes are used to store additional information that may be helpful for the operator in the MCC or for the maintenance technician. The current implementation of the decision support system covers a prototypical multi-agent system and a stand-alone version of the workflow. An interface was implemented to load the results from the workflow into the decision support system. The stand-alone version is used for extensive testing by our project partners and therefore not fully integrated into the decision support system. The workflow itself is fully implemented, but the single tasks should be improved. The multi-agent system contains eleven software agents and seven CBR systems. The multi-agent system and the workflow are implemented with JADE[6] and the CBR systems with myCBR[5]. All CBR systems are using the same case structure, but partially different vocabulary and similarity measures. Over all CBR systems we currently have more than 500 cases, some based on manual input, but the most generated with the workflow based on given data sets. The result an evaluation scenario with 20 queries is that an average of 78 percent of the retrieved cases have an appropriate diagnosis. For each query this number differs slightly. For some queries all retrieved cases were appropriate, for other queries only a few cases were appropriate. Not only the cases itself were checked, but also the ranking of the cases. An average of 18 percent of the retrieved cases were ranked wrong from an expert point of view.

4 Summary and Outlook

In this paper we describe the instantiation of our multi-agent system for case-based decision support on diagnosis and maintenance. We give an overview of the individual components and describe the case structure and similarity assessment of our CBR system. At the moment, we focus our work on the improvement of the workflow for knowledge transformation to get results with higher quality and thus improving the competence of our decision support system. In addition the workflow will be fully integrated into the decision support system. After the improvement of the workflow, a data set of more than 65.000 cases will be processed. Future work will be the implementation of the Case Factory approach for knowledge maintenance into the decision support sys-

tem and extending the learning capabilities for similarity improvements. Furthermore we will extend our CBR tool myCBR to improve the usage in the OMAHA project.

References

1. Althoff, K.D.: Collaborative multi-expert-systems. In: Proceedings of the 16th UK Workshop on Case-Based Reasoning (UKCBR-2012), located at SGAI International Conference on Artificial Intelligence, December 13, Cambridge, United Kingdom. pp. 1–1 (2012)
2. Althoff, K.D., Bach, K., Deutsch, J.O., Hanft, A., Mänz, J., Müller, T., Newo, R., Reichle, M., Schaaf, M., Weis, K.H.: Collaborative multi-expert-systems – realizing knowledge-product-lines with case factories and distributed learning systems. In: Baumeister, J., Seipel, D. (eds.) KESE @ KI 2007. Osnabrück (Sep 2007)
3. Althoff, K.D., Hanft, A., Schaaf, M.: Case factory - maintaining experience to learn. Advances in Case-Based Reasoning Lecture Notes in Computer Science 4106/2006, 429–442 (2006)
4. Bach, K.: Knowledge Acquisition for Case-Based Reasoning Systems. Ph.D. thesis, University of Hildesheim (2013), dr. Hut Verlag München
5. Bach, K., Sauer, C., Althoff, K.D., Roth-Berghofer, T.: Knowledge modeling with the open source tool mycbr. In: Nalepa, G.J., Baumeister, J., Kaczor, K. (eds.) Proceedings of the 10th Workshop on Knowledge Engineering and Software Engineering (KESE10). Workshop on Knowledge Engineering and Software Engineering (KESE-2014), located at 21st European Conference on Artificial Intelligence, August 19, Prague, Czech Republic. CEUR Workshop Proceedings (<http://ceur-ws.org/>) (2014)
6. Bellifemine, F., Caire, G., Greenwood, D.: Developing multi-agent systems with JADE. Jon Wiley & Sons, Ltd. (2007)
7. BMWI, A.L.: Luftfahrt 2020: Die deutsche luftfahrtforschung, partner im globalen wettbewerb (Bonn 2001)
8. Corchado, J.M., Tapia, D.I., Bajo, J.: A multi-agent architecture for distributed services and applications. International Journal of Innovate Computing 8, 2453–2476 (2012)
9. Feret, M., Glasgow, J.: Combining case-based and model-based reasoning for the diagnosis of complex devices. Applied Intelligence 7, 57–78 (1997)
10. Jackson, T., Austin, J., Fletcher, M., Jessop, M.: Delivering a grid enabled distributed aircraft maintenance environment (dame). Tech. rep., University of York (2003)
11. Reuss, P., Althoff, K.D., Henkel, W., Pfeiffer, M.: Case-based agents within the omaha project. In: Case-based Agents. ICCBR Workshop on Case-based Agents (ICCBR-CBRA-14) (2014)
12. Reuss, P., Althoff, K.D., Henkel, W., Pfeiffer, M., Hankel, O., Pick, R.: Semi-automatic knowledge extraction from semi-structured and unstructured data within the omaha project. In: Proceedings of the 23rd International Conference on Case-Based Reasoning (2015)
13. Saxena, A., Wu, B., Vachtsevanos, G.: Integrated diagnosis and prognosis architecture for fleet vehicles using dynamic case-based reasoning. In: Autotestcon 2005 (2005)
14. Srinivasan, S., Singh, J., Kumar, V.: Multi-agent based decision support system using data mining and case based reasoning. International Journal of Computer Science Issues 8, 340–349 (2011)
15. Sun, Z., Han, J., Dong, D.: Five perspectives on case based reasoning. In: 4th International Conference on Intelligent Computing. pp. 410–419 (2008)
16. Zouhair, A., En-Naimi, E.M., Amami, B., Boukachour, H., Person, P., Bertelle, C.: Incremental dynamic case based reasoning and multi-agent systems (idcbr-mas) for intelligent touring system. International Journal of Advanced Research in Computer Science and Software Engineering 3, 48–56 (2013)

Building an integrated CBR-Big Data Oriented Architecture based on SEASALT

PhD Proposal

Kareem Amin

German Research Center for Artificial Intelligence, Knowledge Management, Trippstadter Straße 122, 67663 Kaiserslautern,
kareem.amin@dfki.de

Abstract. The growth of intensive data-driven decision-making is now being recognized broadly. In this paper I propose a CBR – Big Data oriented architecture based on the SEASALT architecture. SEASALT will be enhanced to be compliant with Big Data frameworks. I will use the state-of-the-art/best practices approaches for managing Big Data and CBR. I will go through the process starting from gathering data stage till building a CBR system that is able to answer streams of questions and come up with accurate retrieved results in a reasonable time.

Keywords: Case-Based Reasoning, Big Data, Distributed Architecture, SEASALT, Multiple Agents

1 Introduction

Data is being generated extremely fast — a process that never stops like the data generated from social media networks. Facebook, the most active of social network, with over 1.4 billion active monthly users, generates the most amount of social data – users like over 4 million posts every minute – 4,166,667 to be exact, which adds up to 250 million posts per hour [3]. With the volume of data growing at unprecedented rate; Case-Based Reasoning (CBR) has a new challenge to deal with this large amount of data. Over the last years, CBR has proved its efficiency in different domains. CBR’s rule of thumb is based on reusing the experiences and by this avoiding having to “re-invent the wheel”. It is using the natural human reasoning of looking back to find the most similar cases that could help to solve the new issues. Since all companies now have a lot and different data sources, over days the size of data is getting bigger and the need to deal with this amount of data seamlessly is also increasing (e.g., electronic manuals, history of failure, electronic medical records) [10,12]. Therefore, the key factor for the next generation of CBR applications is the ability to deal with the complex and large amount of data that is generated every day and every moment. I need to use new distributed technologies to be able to scale CBR solutions up, to find new ways to

overcome the cases retrieval latency problems that might be crucial in critical systems (e.g., security systems, condition monitoring, sensors data, etc.).

This research proposal aims to enhance the SEASALT architecture [13, 14] by combining the Multiple Agent CBR System benefits with Big Data processing capabilities, in order to get a generic coherent architecture that can be applied with the future CBR systems.

2 Background

This section describes briefly the motivation for working with Big Data and Case-Based Reasoning and gives a short overview of the related work that has been carried out to achieve a similar goal.

2.1 Big Data, Big Challenges

Big Data management has very big challenges and opportunities, since almost everything today is generating data. Big Data management is not anymore driven by computer scientists or researchers, it is a must for companies today to benefit from their data, or they will not be able to survive in the current fast changing business environment [18]. Big Data management plays a major role in the competition between companies.

Big enterprises like SAP, IBM, Google, Microsoft, SAS and EMC are running now with maximum speed to build the most advanced Big Data platforms, not only from the software side, but also from the hardware one. They aim to attract new customers and help companies to gain the maximum benefit from their data.

The new amount of data requires new, innovative technologies to be able to give the results in a reasonable time [1].

The Big Data term refers to dynamic, large, structured and unstructured volumes of data generated from [2]:

- Traditional data sources – includes the transactional data created from ERP systems, CRMs, web store transactions, etc.
- Machine generated data – includes sensors data, smart meters, web logs, etc.
- Social data – includes data generated from social networks like Facebook, Twitter, LinkedIn, etc.

2.2 Distributed Case-Based Reasoning

CBR needs efficient techniques to manage its subtasks such as collecting and formatting data, case base maintenance, cases retrieval, cases adaptation and retaining new cases. From this point of view, the need to build distributed CBR systems for maximum efficiency increased. Multiple Agent CBR systems are widely used and very well known in Distributed CBR systems area, a lot of frameworks and related work have

been carried out elaborating different architectures and techniques to manage the CBR sub-tasks.

Most researches in Distributed CBR concentrated more on distributing resources within the CBR architecture but not on distributing the case base itself. One of the successful distributed CBR platforms is jCOLIBRI [4]. It supports the development of wide range of CBR software, it provides the required infrastructure to implement CBR systems [5, 6]. jCOLIRBI is depending on multiple agents to perform the subtasks associated with CBR. Multi Agent Systems (MAS) distribute the case base itself and/or some aspects of the reasoning among several agents [7]. I can categorize the research efforts in the area of distributed CBR using two criteria [8]:

1. How knowledge is organized/managed within the system (i.e. single vs. multiple case bases).
2. How knowledge is processed by the system (i.e. single vs. multiple processing agents).

Another example of successful CBR tools is myCBR. myCBR is a joint effort of the Competence Centre CBR at DFKI, Germany, and the School of Computing and Technology at UWL, UK (see <http://mycbr-project.net/>). myCBR Workbench [9] provides user-friendly graphical user interfaces for modelling various kinds of attribute-specific similarity measures and for evaluating the resulting retrieval quality. In order to reduce also the effort of the preceding step of defining an appropriate case representation, myCBR Workbench includes tools for generating the case representation automatically from existing raw data. The accompanying Software Development Kit (SDK) allows for integration into other applications and extension to specific requirements such as additional similarity calculations. Agent-based systems technology has generated lots of excitement in recent years because of its promise as a new paradigm for conceptualizing, designing, and implementing software systems [17]. In [13, 14] the SEASALT architecture is an application-independent architecture to work with heterogeneous data repositories and modularizing knowledge to be structured. It was proposed based on the CoMES approach to develop collaborative multi-expert systems. SEASALT aims to provide a coherent multi-agent CBR architecture that can define the outlines and interactions to develop multi-agent CBR systems. The SEASALT team has applied it in [14] to travel medicine as part of the docQuery project. It was as a textual CBR application domain to showcase how SEASALT could be used. In [15] Albert Pla et al. have provided a user friendly tool for medical prognosis (eXiT*CBRv2). They proposed an innovative multi-agent system architecture, in which they have a coordinator agent that is responsible for receiving new cases, then pass it to n agents. Each agent is connected with case base to retrieve cases based on different retrieval calculations. Afterwards, they all pass the results again to the coordinator agent to assess and compare results and at the end it gives the final results. They illustrated the use of the tool through several experiments carried out with a breast cancer database and they show how easy it is to compare distributed approaches that maintain naturally distributed clinical organization, compared to centralized systems.

Generally, CBR and MAS have proved efficiency with different successful distributed CBR systems. Currently systems are getting more complex and agents need to be

smarter to be able to deal with its environments. MAS bring a lot of advantages and benefits to CBR but also have a lot of challenges and issues that should be taken into consideration while building systems:

1. How do we design our algorithms to decompose tasks to agents and allocate problems to them?
2. If systems are widely distributed, how are agents going to communicate and what communication protocols will they use?
3. What if we lost the communication between agents?
4. How do we ensure that agents are working properly and every single agent is doing its task in the perfect manner?
5. How do we troubleshoot issues across all agents?

2.3 Case-Based Reasoning & Big Data

CBR and Big Data collaboration is an emerging topic, some researches and efforts have been carried out in this area. Since the growth of digital data is widely heralded. A 2014 article estimates that “Almost 90% of the world’s data was generated during the past two years, with 2.5 quintillion bytes of data added each day” [10]. In [11] Yu-Hui X & Xiao-Yun Tian provided a CBR model *NT-CBR* based on the data mining technology *NT-SMOTE*. They have tried to solve the problems associated with enterprise risk management and compared their results with different methodologies. The NT-CBR model used big internet data to do the forecasting of risks and give smarter and faster solutions to the risk. In [12] Vahid Jalali & David Leake have initially developed *ensembles of adaptation for regression* (EAR), a family of methods for generating and applying ensembles of adaptation rules for case-based regression. That model suffered from high computational complexity and therefore they decided to go to Big Data techniques (Map Reduce) to improve their model performance and they called it BEAR. BEAR uses MapReduce and Locality Sensitive Hashing (LSH) for finding nearest neighbors of the input query. It consists of two main modules: LSH for retrieving similar cases and EAR for rule generation and value estimation. As a conclusion they got very promising results that encourage them to perform bigger experiments to ensure that the model is reacting in the perfect manner.

3 Research Focus

In my research I will focus on integrating several methodologies, CBR, Big Data, Data Streaming, Multi-Agent Systems and SEASALT as a background architecture orchestrating the interaction between different entities. I mainly aim to address the handling and processing of big case bases to avoid cases retrieval latency. I will also extend the SEASALT [13, 15] architecture to be compliant with Big Data systems architectures. It is also important to mention that at **ICCBR 2016** a dedicated workshop will take place to discuss all time problems related to CBR systems like real time data processing and big data (see <http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=55489©ownerid=89330>).

3.1 Motivation

My work is motivated by the study done on the prior work related to CBR, MAS, and Big Data. I aim to build a big data oriented multi-agent CBR architecture. The main challenge is how I will develop a framework for integrating all these methodologies into one robust architecture and orchestrating the interaction between them all, which will allow CBR systems to deal with big case bases.

3.2 Research Problem

As the case base grows, the swamping utility problem could affect the case retrieval times, degrading system performance and credibility [12]. For example, calculating metrics such as number of visitors or page views for a social media or e-commerce web site with hundreds or million users is a common practice in industry, but in current CBR research, experiments with tens or thousands of cases, or even much fewer, are common. The proposed approach is trying to focus on building scalable CBR systems that able to work with the big case bases and being able to deal with online data streams. I will work to integrate the new Big Data frameworks like Spark or Flink with a Multi-Agent CBR Framework (JADE) in order to build an innovative solution based on the proposed architecture.

3.3 The Proposed SEASALT – Big Data Oriented Architecture

One key innovation of the SEASALT – Big Data Oriented architecture is to improve the problem-solving technology of CBR at the level of big data with the help of the modern frameworks that distribute processes for better performance.

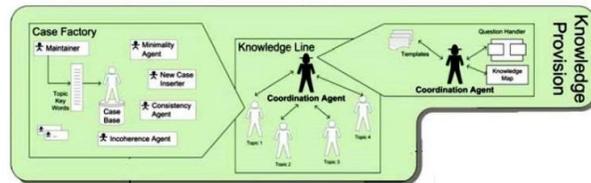


Fig. 1. SEASALT Knowledge Provision Layer

Figure 1 depicts the Knowledge Provision Layer in SEASALT, in which the coordination agent is connected with a number of topic agents. Every topic agent is considered as a standalone CBR system with its own case factory.

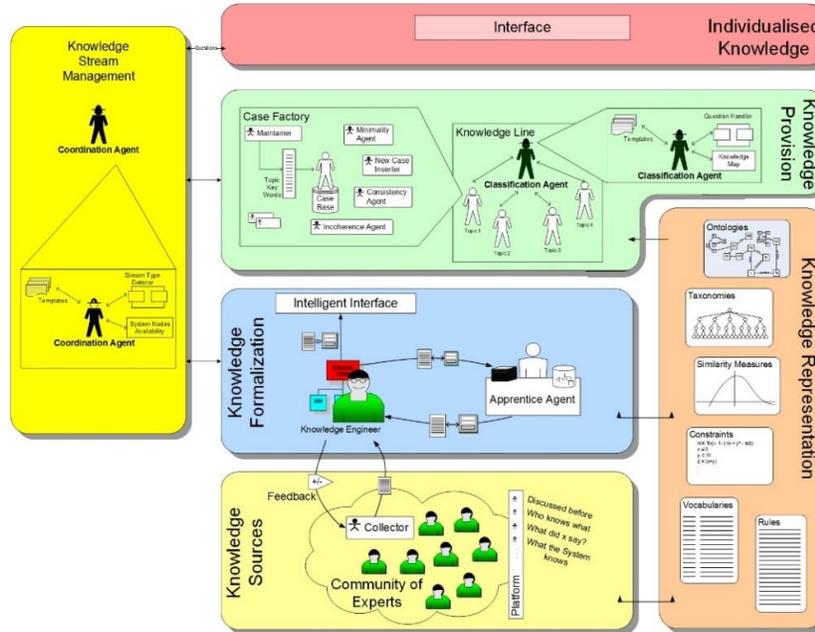


Fig. 2. SEASALT-Big Data Oriented Architecture

Figure 2. Illustrates the changes made to the Knowledge Provision layer and the new Knowledge Stream Management layer. The Knowledge Stream Management layer is getting two types of input, one from Knowledge Formalization layer that is related to new data insertion, and the other input is coming as stream of questions from the Individualized Knowledge layer. According to our architecture, system nodes would be the available processing power – *Node = Single Processing Power*. The Knowledge Provision layer will be distributed across several nodes, and hence each node contains Knowledge Provision agents. The Coordination Agent will act as the system manager who is aware of all the system nodes and responsible for the whole system control. He will be the *data tap* that uses the underlying framework to distribute the incoming requests across the system nodes. Normally, there are two kinds of nodes, one for *Queries* processing to retrieve results and the second for *New Cases* processing. It is possible to have up to N nodes in the system according to the hardware availability. The more nodes, the better performance we get. In every node, there is a Classification Agent to classify the received data and assign it to the intended Topic Agent. Each Classification Agent is aware of the knowledge map gathered from knowledge sources and classify

the incoming requests according to predefined classes. Then, the Classification Agent assigns the request to the intended Topic Agent(s). The Topic Agent is performing queries to retrieve the most similar cases. Since I use distributed nodes in the hardware cluster, the Case Base will be replicated to avoid data integrity using the replication channels to replicate data between the Case Base instances (see Figure 2). Since our Case Base will be distributed among several nodes, the Case Factory agents will be centralized. Therefore the Case Factory will have only one instance that performs case maintenance on a single Case Base. Afterwards, the results will be distributed to the whole system nodes using the replication channels.

This architecture is addressing the domains/tasks that require CBR and are in need for Big Data processing in the same time (e.g., Anomaly Detection, Condition Monitoring, Medical Diagnosis, etc.). A cooperative study with the AGATA (Analyse großer Datenmengen in Verarbeitungsprozessen, engl.: analysis of large amount of data in manufacturing processes) project [16] at DFKI is going to be established to showcase the performance and evaluate the proposed approach.

4 Conclusion and Future Directions

In this PhD proposal I have proposed and presented the SEASALT-Big Data oriented architecture. It is an integrated Big Data-oriented version of the SEASALT architecture that aims to tackle Big Data problems with CBR. In summary,

1. I want to deal with Big Data where processing requires manual knowledge modeling in addition as context/background knowledge.
2. I want to improve CBR to be able to deal with Big Data.
3. I want to develop a methodology for developing CBR-Big Data oriented applications.
4. I want to evaluate the developed methodology and tools based on AGATA and other applications.

As future directions, I would explore more literature and related work for joint researches between CBR and Big Data. I will also be looking for data sets in order to test the implementation and compare it with others.

References

1. A community white paper developed by leading researchers across the United States, "Challenges and Opportunities with Big Data," Purdue University, USA, 2012.
2. J. Singh, "Big Data Analytic and Mining with Machine Learning Algorithms," *International Journal of Information and Computation Technology*, 2014.
3. George Simos, [Online]. Available: <http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/>, [Accessed 19 08 2015]

4. Juan A. Recio-García, Belén Díaz-Agudo and Pedro A. González-Calero, "The COLIBRI Platform: Tools, Features and Working Examples," *Springer-Verlag Berlin Heidelberg*, 2014.
5. Belén. Díaz-Agudo, Juan A. Recio-García and Antonio A. Sánchez-Ruiz-Granados, "Building CBR systems with JCOLIBRI," Special Issue on Experimental Software and Toolkits of the Journal Science of Computer Programming, pp. 68-75, 2007.
6. Juan A. Recio-García, Belén. Díaz-Agudo, Sergio González-Sanz and Lara Q. Sanchez, "Distributed deliberative recommender systems," *T. Computational Collective Intelligence*, pp. 121-142, 2010.
7. Enric Plaza and Lorraine Mcginty, "Distributed case-based reasoning," *Knowledge Engineering Review* 20, p. 261–265, 2006.
8. Aitor Mata, "A Survey of Distributed and Data Intensive CBR Systems," *Springer-Verlag Berlin Heidelberg*, pp. 582-586, 2009.
9. T. Roth-Berghofer, J. Antonio R. García, Christian S. Sauer, Kerstin Bach, KD. Althoff, B. D. Agudo and P. A González Calero, "Building Case-based Reasoning Applications with myCBR and COLIBRI Studio," in *ICCBR*, Lyon, 2012.
10. Gang-Hoon Kim, Silvana Trimi and Ji-Hyong Chung, "Big-data applications in the government sector," *Communications of the ACM*, pp. 78-85, 2014.
11. Yuhui Xu and Xiaoyun Tian, "Internet Big Data Information Analysis and Power Intelligent Automation Risk Prediction Based on Case Based Reasoning," in *3rd International Conference on Machinery, Materials and Information Technology Applications*, Qingdao, 2015.
12. Vahid Jalali and David Leake, "CBR Meets Big Data: A Case Study of Large-Scale Adaptation Rule Generation," *Case-Based Reasoning Research and Development*, pp. 181-196, 2015.
13. K. Bach, M. Reichle and Klaus-Dieter. Althoff, "A Domain Independent System Architecture for Sharing Experience," in *Workshop Wissens- und Erfahrungsmanagement*, 2007.
14. Meike Reichle, Kerstin Bach and Klaus-Dieter. Althoff, "Knowledge engineering within the application-independent architecture SEASALT", *International Journal Knowledge Engineering and Data Mining*, vol. 1.1, no. 3, pp. 202-215, 2011.
15. Albert Pla, B. Lopez, P. Gay and C. Pous, "eXiT*CBR.v2: Distributed case-based reasoning tool for medical prognosis," *Decision Support Systems*, vol. 54, no. 3, p. 1499–1510, February 2013.
16. S. Windmann, A. Maier, O. Niggemann, C. Frey, A. Bernardi, Ying Gu, H. Pfrommer, T. Steckel, M. Kruger and R. Kraus, "Big Data Analysis of Manufacturing Processes," in *12th European Workshop on Advanced Control and Diagnosis*, 2015.
17. Katia P. Sycara, "Multiagent Systems," *AI Magazine*, 1998.
18. "Better business outcomes with IBM Big Data Analytics," 2016. Available http://www935.ibm.com/services/multimedia/59898_Better_Business_Outcomes_White_Paper_Final_NIW03048-USEN-00_Final_Jan21_14.pdf, [Accessed 01 07 2016].

Towards rapidly developing database-supported machine learning applications

Frank Rosner¹ and Alexander Hinneburg²

¹ Global Data and Analytics, Allianz SE, Germany

² Computer Science, Martin-Luther-University Halle-Wittenberg, Germany

Abstract. The development of a big data analytics application benefits from a conceptual model that jointly represents aspects about data management as well as machine learning. We demonstrate a recently proposed method to translate a Bayesian network into a usable entity relationship model using the real world example of the TopicExplorer system. TopicExplorer is an interactive web application for text mining that uses Bayesian topic models as a core component. Further, we sketch a vision of a conceptual framework that eases machine learning specific development tasks during building big data analytics applications.

1 Introduction

The implementation of a big data analytics application requires to join data management software with machine learning tools. However, the fields of data management and machine learning developed quite different models and notations. The former frequently uses entity-relationship models (ERM) [5] while the latter uses probabilistic graphical models in particular Bayesian networks (BN) to communicate key concepts during development. Even while both kinds of graphical notations show many details of the data, information explicit on one side remains implicit on the other one and vice versa — there is no natural understanding of the two worlds. However, a common conceptual description of the contributions from both worlds is crucial for the successes of big data analytics development projects.

Recently, we proposed a translation [13, 14] from a graphical BN model in plate notation into an entity relationship model. Such ERM can be easily integrated into the overall ERM of the whole application. Thus, we gain the advantage of a formal conceptual view of the machine learning part that is integrated into the conceptual view of the data management side. Thereby, developers from the data management side understand the basic in- and outputs of the machine learning part that remains no longer as a black box behind an abstract API in the data management model.

We demonstrate the method in the real world example of the TopicExplorer in Section 2. Based on this, we describe our vision of a conceptual framework that uses pre-translated BNs as a library of ERM snippets in Section 3. Such library could be used by data management developers to conceptually include machine

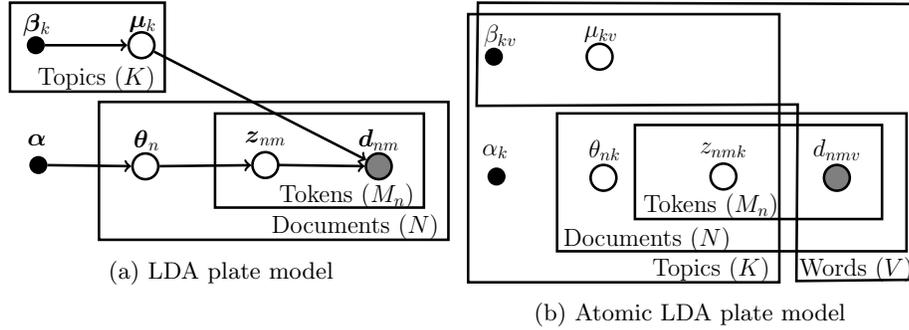


Fig. 1: Transformation of the LDA plate model to an APM.

learning methods into analytics applications. We believe that software development of big data analytics applications could benefit from machine learning implementations that are attached to the pre-translated BNs. Last, we discuss related work in Section 4 and conclude the paper in Section 5.

2 Case Study: Text Topic Modeling

We demonstrate the recent method [13, 14] to translate BN to ERM using the example of the TopicExplorer [10, 11], an application to explore document collections using probabilistic topic models [4]. We explain how the translated LDA topic model is represented as ERM. Furthermore, we show use cases for typical analyses supported by the translated ERM.

2.1 Translation of Latent Dirichlet Allocation to ERM

LDA [4] models a collection of documents that is indexed by the set N . Each document $n \in N$ consists of a set of tokens M_n that represents the words occurring in this document. The document specific token index sets M_n partition the total index set of tokens M . A token $m \in M_n$ corresponds to exactly one word type v from a vocabulary V . In the Bayesian network, Figure 1a, this information is coded as a bit vector $\mathbf{d}_{nm} \in \{0, 1\}^{|V|}$ that has exactly a single 1 at the index associated with the respective word $v \in V$. Each token $m \in M_n$ is also assigned to a topic $k \in K$. This assignment is coded by the bit vector $\mathbf{z}_{nm} \in \{0, 1\}^{|K|}$, which has a single 1 at the respective topic index. Furthermore, each topic has its own word distribution parameterized by a vector of positive real number $\boldsymbol{\mu}_k \in \mathbb{R}^{|V|}$. The topic proportions per document are represented by a similar vector $\boldsymbol{\theta}_n \in \mathbb{R}^{|K|}$. The hyper-parameter vectors $\boldsymbol{\alpha} \in \mathbb{R}^{|K|}$ and $\boldsymbol{\beta}_k \in \mathbb{R}^{|V|}$ regulate the prior distributions for the respective hidden parameters $\boldsymbol{\theta}_n$ and $\boldsymbol{\mu}_k$.

The translation [13, 14] delivers the ERM shown in the right part of Figure 2 for the given LDA plate model (Figure 1a). The translation employs several intermediate steps, one of which is the transformation of the plate model into

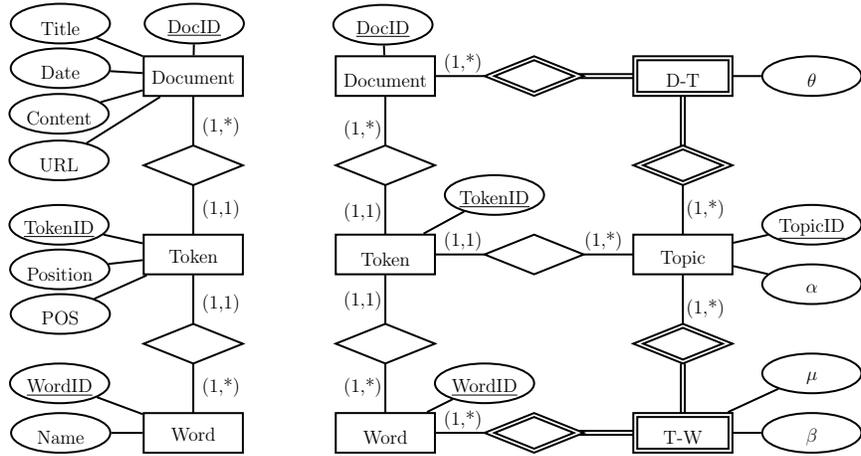


Fig. 2: ERM of given Data (left) and translated ERM for LDA (right).

an atomic plate model (APM), see Figure 1b. The APM represents implicit relational information hidden in the original plate model in an explicit way using the plate notation. By the transformation and reduction rules [13, 14, 8], the APM is translated into a sequence of several intermediary ERMs and then reduced into a usable final ERM.

Such ERM is close to a manually designed ERM for LDA. A document consists of one or more tokens which are of exactly one word type. Each token is assigned to a topic, while one topic can have multiple tokens assigned. The inferred topic mixture for each document is stored in $D-T.\theta$, while $T-W.\mu$ holds the word probabilities for each topic. The hyper-parameter α of the prior for the topic mixture resides as an attribute of the topic entity type. The parameters for the individual priors for the word distributions are stored in $T-W.\beta$.

2.2 TopicExplorer

TopicExplorer is a web application that helps users from the humanities to work with topic models, e.g. in a collaboration with the institute for Japanese studies at Martin-Luther-University, we analyzed blog posts about the Fukushima disaster. After crawling relevant blogs, each blog entry is preprocessed by computer linguistic software to extract tokens from full text and store them in their lemmatized forms together with their part-of-speech tags (e.g. noun, verb or adjective) and their string positions in the text.

TopicExplorer interactively visualizes the topic structure of the documents. The visualizations require to join the data about documents and words together with results from the topic model. An ERM that would integrate the left and the right part of Figure 2 is obtained by merging the matching entities from both sides. It gives the application developer a good idea how to access those data, without needing to understand the machine learning details of a topic model.

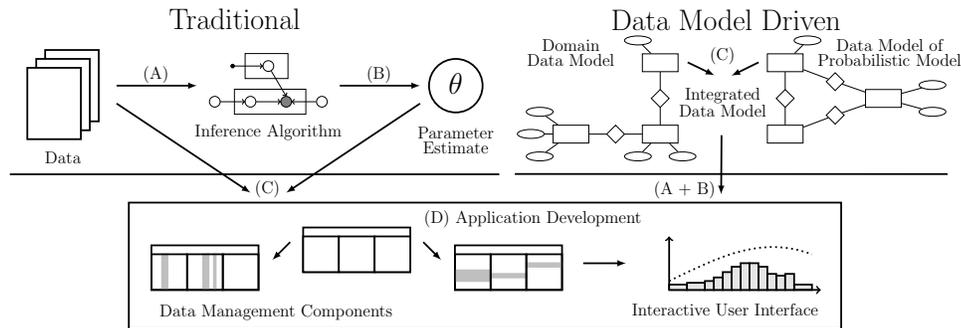


Fig. 3: Comparison of traditional development versus data model driven development of big data analysis applications.

We present how to derive a few visualizations that are part of the current version of TopicExplorer [10].

Document topic mixture. LDA assigns a vector of topic probabilities stored in $D-T.\theta$ for each document, called a topic mixture. This could be presented as a list of topics with decreasing order of probabilities.

Topic Documents. Reversing the idea behind the document topic mixture, one can visualize a topic as a list of the most representative documents for this topic. This is done by joining `Document`, `Token` and `Topic`, grouping by both IDs of documents and topics and counting the number of tokens in each group. For each topic the entries are sorted with decreasing token count, yielding a list of representative documents.

Topic words. As stated above, a topic can be represented as a list of words sorted in decreasing probability ($T-W.\mu$). Furthermore, the topics appear linearly ordered by similarity in the user interface to allow browsing in a semantically uninterrupted way. Computing all pair-wise similarities between topics is well supported by a relational database using table `T-W`.

Topic frames. Another visualization of topics uses the concept of frames. A topic frame consists of a noun and a verb that are assigned to the same topic and appear close together in the same documents. Topic frames can be computed using `Token`, `Word` and `Topic`, grouping by topic ID and word IDs of the frame tokens, and counting the number of frames using the same words.

Topic time. To analyze how discussions in blogs evolve, TopicExplorer allows to visualize the number of tokens assigned to topics over time. This analysis is also directly supported by joining documents with topics, grouping by date and topic ID and then counting the tokens.

3 Conceptual Modeling Framework

Based on the method for translating probabilistic models to ERMs and our experiences with the development of the TopicExplorer system, we propose a

first idea for a new data model driven development approach dedicated to big data analytics applications. Figure 3 visualizes the traditional and our proposed data model driven development approaches. Both address four different tasks, namely (A) gather the data sources and make them available to a probabilistic model, (B) run machine learning components, (C) integrate the data sources with the machine learning output and (D) build the application consisting of data management components and an interactive user interface.

The traditional approach addresses the tasks mainly in sequential order. The first three steps implement data mining process following the CRISP model [7], while the last step addresses standard application development.

The translation method [13, 14] from BN to ERM allows an alternative, data model driven approach. We assume that for a wide spectrum of machine learning problems abstract, readily developed BNs already do exist. Those BNs could be pre-translated to ERMs to build a library. Thus, conceptual information about the machine learning component is already available when integrating the data sources, task (C). The BN could be treated as just another data source. The entities corresponding to observed variables in the BN, including their respective relationships, have to be matched with those from other available data sources. The matching conceptually defines the interface between data sources and machine learning, task (A). Furthermore, translating the integrated ERM into a (relational) model for a big data framework, e.g. Flink [2] or Spark [3], conceptually defines the API between the output of machine learning and the rest of the application, task (B). Depending on the framework, the tasks (A) and (B) could be supported by generation of efficient code for interfaces to access the given data as well as machine learning implementations.

As a consequence, application developers just need knowledge about in- and outputs, and the relationships among the variables in the Bayesian model, but not about probabilistic distributions and dependencies. Thus, our new data model driven approach eliminates unnecessary complexity caused by a lack of compatible conceptual languages on both sides of machine learning and data management. Thereby, it makes the collaboration between both sides more direct and offers potential for optimization.

4 Related Work

There are several approaches that combine data management with machine learning, however, none of them reaches a comparable conceptual level like ERMs. Hazy [12] provides programming, infrastructure and statistical processing abstractions, the latter are based on Markov logic [6]. This requires a deeper understanding of the machine learning algorithms in order to combine them effectively with data management. Several approaches combine machine learning APIs with SQL [1-3, 9] or with their own declarative language [16].

Last, data management is combined with machine learning at the level of user interfaces. An recent example is scikit-learn [15], which enable users to quickly select data sources and try different algorithms. Both do not offer an easy way to

integrate machine learning results with domain specific meta data. Our approach also contrasts with statistical programming languages and software like R and SAS that just offer programming APIs to data sources and machine learning algorithms.

5 Conclusion

Knowledge about the machine learning side of a project helps developers to build a big data analytics application. The subsequently proposed framework gives guidelines how to effectively build an integrated conceptual model that includes details about domain specific aspects as well as the machine learning side of a big data analytics application. Future work includes the implementation of the framework and optimizing efficiency when translating integrated conceptual models to a particular implementation.

References

1. Akdere, Cetintemel, Riondato, et al. The case for predictive database systems: Opportunities and challenges. In *CIDR*, pp. 167–174, 2011.
2. Alexandrov, Bergmann, Ewen, et al. The stratosphere platform for big data analytics. *The VLDB Journal*, 23(6):939–964, 2014.
3. Armbrust, Xin, Lian et. al. Spark sql: Relational data processing in spark. In *SIGMOD*, pp. 1383–1394, 2015.
4. Blei, Ng, and Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
5. Chen. The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS)*, 1(1):9–36, 1976.
6. Domingos and Richardson. Markov logic: A unifying framework for statistical relational learning. *Introduction to statistical relational learning*, pp. 339–371, 2007.
7. Han, Kamber, and Pei. *Data Mining: Concepts and Techniques*. MK Pub., 2011.
8. Heckerman, Meek, and Koller. Probabilistic entity-relationship models, prms, and plate models. *Introduction to statistical relational learning*, pp. 201–238, 2007.
9. Hellerstein, Ré, Schoppmann, et al. The madlib analytics library: or mad skills, the sql. *VLDB*, 5(12):1700–1711, 2012.
10. Hinneburg, Oberländer, Rosner, et al.. Exploring document collections with topic frames. in *CIKM 2014*, pp. 2084–2086, 2014.
11. Hinneburg, Preiss, and Schröder. Topicexplorer: Exploring document collections with topic models. In *PKDD*, Part II, pp. 838–841, 2012.
12. Kumar, Niu, and Ré. Hazy: Making it easier to build and maintain big-data analytics. *Communications of the ACM*, 56(3):40–49, 2013.
13. Rosner and Hinneburg. Translating bayesian networks into entity relationship models. In *35th Int. Conf. on Conceptual Modeling, ER*, 2016. to appear.
14. Rosner and Hinneburg. Translating Bayesian Networks into Entity Relationship Models, Extended Version. *ArXiv e-prints, 1607.02399*, July 2016.
15. Scikit. scikit-learn, 2014. Machine Learning in Python.
16. Sparks, Talwalkar, Smith, et al. Mli: An api for distributed machine learning. In *ICDM*, pp. 1187–1192, 2013.

Vereinheitlichung internationaler Bibliothekskataloge

Christian Scheel, Claudia Schmitz, and
Ernesto William De Luca

Georg-Eckert-Institut — Leibniz-Institut für internationale Schulbuchforschung,
Celler Straße 3, 38114 Braunschweig, Deutschland
{scheel,schmitz,deluca}@gei.de
<http://www.gei.de>

Zusammenfassung Aus dem Bereich der Schulbuchforschung beschreiben wir exemplarisch, welche Herausforderungen, Chancen und Stolpersteine es gab, ein Rechercheinstrument einer deutschen Schulbuchsammlung zu einem internationalem Rechercheinstrument auszubauen. Hierbei standen vor allem das Erstellen und Erweitern einer gemeinsamen Repräsentationsbasis, sowie das Wiederverwenden existierender Lösungen im Mittelpunkt. Diese Ausarbeitung hilft bei der zielorientierten Planung ähnlich ausgerichteter Vorhaben.

Keywords: Cross-Lingual IR, Cross-Cultural IR, Digital Humanities

1 Einleitung

Der Gegenstand der Schulbuchforschung ist die Untersuchung von gesellschaftlich (politisch), pädagogisch und fachwissenschaftlich sanktioniertem Wissen für die Bildung der jungen Generation. Zum Beispiel gewinnen Schulbücher auf der Suche nach „populärem Wissen“ für die historische Forschung zunehmend an Bedeutung, da sich in ihnen Weltanschauungen, Denkströmungen und erwünschte Wissensbestände widerspiegeln. Am Georg-Eckert-Institut¹ wird die Forschung deshalb vor allem von den Erziehungswissenschaften, der Geschichtswissenschaft, der Geographie, der Politikwissenschaft und der Religionswissenschaft bestimmt [5,6,4]. Die Vorteile für einen internationalen Blickwinkel liegen klar auf der Hand, weshalb die Idee entstand ein institutionelles Rechercheinstrument für Schulbücher zu einem internationalen Rechercheinstrument auszubauen.

Ein Rechercheinstrument speziell für Schulbücher ist deshalb notwendig, da Bibliotheken Schulbücher im Allgemeinen nicht forschungsadäquat als Schulbücher nachweisen, sondern schlicht als Bücher. Dem Schulbuchforscher ist es deshalb nicht möglich, explizit nach Schulbüchern zu recherchieren. Hinzu kommt, dass Schulbücher spezielle Eigenschaften wie „Geltungsland“, „Unterrichtsfach“, „Bildungslevel“, „Schulform“, etc. aufweisen, die ein Rechercheinstrument berücksichtigen muss.

¹ <http://www.gei.de/das-institut.html>

Erweitert man ein monolinguales Rechercheinstrument mit Katalogen aus internationalen Beständen, hat man es zuallererst mit zwei Problemen zu tun. Zum einen möchte der internationale Wissenschaftler das Rechercheinstrument in seiner Sprache benutzen, zum anderen nutzen die Bibliothekare natürlich ihre Muttersprache und gegebenenfalls eigene Codes um die Eigenschaften der Bücher zu beschreiben. Man muss also nicht nur jedes Attribut und jede Ausprägung dieser Attribute in jeder Sprache beschreiben können, sondern auch die Beschreibungen in den Katalogen auf eben diese Attribute und deren Ausprägungen abbilden (mappen) können.

Die Überführung vom institutionellem Rechercheinstrument TextbookCat² zum internationalen Rechercheinstrument International TextbookCat dient neben dem Aufbau einer erweiterbaren Architektur unter anderem dazu, Erfahrungen mit Workflows und Arbeitsschritten zu sammeln sowie etwaige Herausforderungen und notwendige Ressourcen besser einschätzen zu können. Um die Chancen und Fallstricke eines solchen Vorhabens offen zu legen, konzentrierte sich der International TextbookCat auf die Zusammenführung verschiedensprachiger Schulbuchdatenbanken aus drei Institutionen: dem Georg-Eckert-Institut, der Università degli Studi di Torino und der Universidad Nacional de Educación a Distancia. Die Zusammenarbeit verfolgte die Festlegung einheitlicher Standards zur Erfassung von Schulbuchdaten und die Anfertigung von Mappings auf ein einheitliches Datenformat.

Das entstandene Rechercheinstrument kann stellvertretend für jedes spezialisierte Rechercheinstrument stehen, bei dem geplant ist, die Datenbasis durch internationale Kataloge zu einem Metakatalog zu erweitern.

2 Die Ausgangslage

Um die Ausgangslage vor der Internationalisierung erfassen zu können, werden im Folgenden die institutionelle Schulbuchsammlung und internationale Schulbuchsammlungen beschrieben. Es folgt eine Beschreibung des Rechercheinstruments und eine Betrachtung der Vorteile eines internationalen Rechercheinstruments.

2.1 Die institutionelle Schulbuchsammlung

Die Schulbuchsammlung der Forschungsbibliothek³ umfasst rund 175.000 Schulbücher, welche auch vor Ort hinterlegt sind. Um diese Schulbücher besser beschreiben zu können erarbeiteten die Bibliothekare ein Klassifikationssystem mit welchem jedes Buch der Sammlung detailliert beschrieben werden konnte (siehe Tabelle 1). Eine Besonderheit beim Geltungsland ist der teilweise

² <http://vufind.gei.de/vufind2/>

³ <http://www.gei.de/bibliothek>

Gebrauch von Jahreszahlen zur genauen Spezifikation ⁴. Die Klasse „Bundesland“ beinhaltet dabei nur Codes für die deutschen Bundesländer. Aufgrund der schulwissenschaftlichen Ausrichtung des Georg-Eckert-Instituts finden sich im Bereich „Unterrichtsfach“ nur die Fächer Geschichte, Sozialkunde, Geographie und Religion, sowie muttersprachlicher Unterricht. Mathematikschulbücher sind zum Beispiel nicht Teil der Sammlung, da man mit ihnen die Fragestellungen der Schulbuchwissenschaft nicht beleuchten kann. Die Klasse „Bildungsgang“ entspricht fast ausschließlich der Klassifikation der International Standard Classification of Education (ISCED) der UNESCO. Die „Schulform“ orientiert sich am deutschen Schulsystem, auch wenn bei circa 100.000 Schulbüchern des Bestandes das „Geltungsland“ nicht Deutschland ist. Die Klasse „Zeitraum“ ist redundant zum Publikationsjahr ermöglicht jedoch eine genaue Klassifikation zu bestimmten historisch relevanten Epochen. Die schulbuchspezifische Klasse „Publikationsform“ unterscheidet zum Beispiel zwischen Schulbuch, Lehrplan, Lehrmittel, Lehrerhandbuch, Aufgabensammlung, etc. Abbildung 1 zeigt die Abdeckung der annotierten Attribute des Klassifikationssystems in der institutionellen Schulbuchsammlung. Da ein großer Teil der Sammlung aus internationalen Schulbüchern besteht, haben die international gültigen Attribute erwartungsgemäß eine höhere Abdeckung.

Tabelle 1. Lokale Notation, wobei _ den Platzhalter für Ziffern oder Buchstaben darstellt.

Code(s)	Klasse	Ausprägungen
l_ _ _	(Geltungs-) Land	181
b_ _ / b_bz	Bundesland / Besatzungszone Deutschland	16 / 4
u_ _ _	Unterrichtsfach/Lernbereich	15
k_ _	Klassenstufe/Bildungsgang	7
s_ _	Schulform Deutschland	11
z_ _ _	Zeitraum	15
d_ _	Publikationsform – Inhaltsform (Dokumenttyp)	12

2.2 Internationale Schulbuchsammlungen

Internationale Sammlungen sind meist aus der Notwendigkeit heraus entstanden, bekannte Schulbücher geordnet zu erfassen, weil es keine explizite Behandlung von Schulbüchern in Bibliotheken gab. Anders als am Georg-Eckert-Institut, bei dem alle Schulbücher der Sammlung Teil der Forschungsbibliothek sind, stellen die Datenbanken internationaler Schulbuchforscher meist Bibliografien dar. Dies ist kein Nachteil, denn andererseits wären Informationen über Schulbücher, die

⁴ Zum Beispiel:

/025: Jugoslawien (-1992)

/125: Jugoslawien, Föderative Republik (1992-2003)

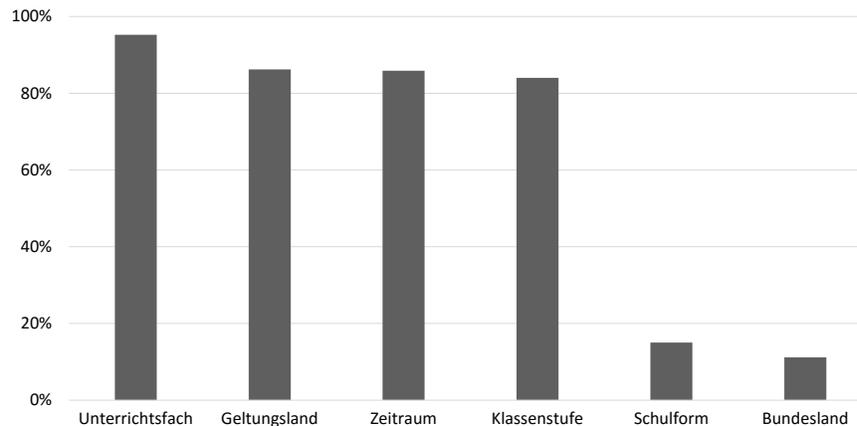


Abbildung 1. Abdeckung der annotierten Attribute des Klassifikationssystems in der institutionellen Schulbuchsammlung.

in Klosterbibliotheken hinterlegt sind schlicht unbekannt. Zusätzlich ist die Qualität der Informationen gerade bei schulbuchspezifischen Eigenschaften als hoch einzuschätzen und ähnlich genug, dass sie sich in ein einheitliches Klassifikationssystem abbilden lassen. Beispielhaft werden im Folgendem drei prominente Sammlungen beschrieben. Weitere Sammlungen können [10] entnommen werden.

EDISCO Das Research Center for digitization and creation of digital libraries for the humanities der Universität Turin hostet die Datenbank EDISCO, die mit etwa 30.000 Datensätzen italienische Schulbücher aus dem 19. und 20. Jahrhundert nachweist.

MANES Das Department of History of Education and Comparative Education hostet die Datenbank MANES, die mit etwa 35.000 Datensätzen spanische, portugiesische und lateinamerikanische Schulbücher aus dem 19. und 20. Jahrhundert nachweist. Die Metadaten in MANES werden nach abgestimmten und festgeschriebenen Katalogisierungsrichtlinien erfasst [9].

Emmanuelle Textbook Project Alain Choppin war ein Pionier der Schulbuchforschung. Unter dem Titel „Emmanuelle Textbook Project“ erschuf er 1979 eine Sammlung französischer Schulbücher ab dem Jahr 1789 [2,3].

2.3 Der TextbookCat

Der TextbookCat ist das Rechercheinstrument des Georg-Eckert-Institut — Leibniz-Institut für internationale Schulbuchforschung. Neben der Recherche

nach Schulbüchern, die bestimmte Attribute aufweisen, wird der TextbookCat von Schulbuchforschern genutzt, um Forschungsfragen zu finden, zu beantworten und zu beleuchten.

Im TextbookCat spiegelt sich das Klassifikationssystem der Forschungsbibliothek als Facetten wider. Durch die Auswahl von Ausprägungen gewünschter Attribute über die Facetten können Benutzer bei der Recherche die Ergebnismenge manipulieren, sodass sie ihrer Forschung dienlich ist [1].

Eine Analyse der Log-Dateien des Rechercheinstruments zeigte, dass der Großteil der Benutzer die Facetten für Recherchen nutzte. Drei Viertel aller Rechercheanfragen basierten ausschließlich auf der Nutzung der Facetten. Dementsprechend wurde bei einem Viertel mit der Hinzunahme von Suchbegriffen, also der klassischen Textsuche, recherchiert.

Bei der Migration des TextbookCat sollten nicht nur die Datenbestände internationalisiert werden, sondern auch die Qualitätseigenschaften des Rechercheinstruments, wie die Abdeckung der schulbuchspezifischen Attribute (vgl. Abbildung 1), bestehen bleiben. So sollte verhindert werden, dass die Schulbuchsammlung des Georg-Eckert-Instituts in den Ergebnismengen überrepräsentiert wird.

2.4 Chancen des International TextbookCat

Die Erweiterung des institutionellen Rechercheinstruments *TextbookCat* zum internationalen Rechercheinstrument *International TextbookCat* erfolgt vor allem im Backend der Architektur. Bevor Sammlungen in sprachunabhängige Suchindexe aufgenommen werden können, müssen sie in eine einheitliche Datenstruktur überführt werden (vgl. [8]).

Die bekannten Schulbuchsammlungen nach und nach in einem Rechercheinstrument zu vereinen, bringt mehrere Vorteile.

1. Man nähert sich immer mehr einer generellen Repräsentationsstruktur von Schulbüchern und somit einem Standard. Wenn die Materie am Anfang zu groß erscheint, benötigt es einen iterativen Ansatz, um sich mit jedem weiteren Datensatz einem Standard zu nähern.
2. An einem Ort in verschiedenen Sammlungen zu recherchieren erspart den Aufwand für das Finden der individuellen Rechercheoberflächen der Sammlungen⁵.
3. Ein einheitliches Klassifikationssystem, welches Eigenschaften und Ausprägungen mit Codes beschreibt, unterstützt Multilingualität ideal, da die Codes in die jeweiligen Sprachen abgebildet werden können.
4. Auch wenn man nie weiß, wie viele Schulbücher man nicht im System hat, liefert ein vereinigt Rechercheinstrument mit der Internationalität eine neue Dimension in der Schulbuchforschung.

⁵ Beispielfhaft kann der Leser versuchen das Rechercheinstrument vom Emmanuelle Textbook Project zu finden.

3 Beobachtungen

Im Idealfall hätten die Sammlungen der internationalen Partner eins zu eins auf die Datenstruktur des Georg-Eckert-Instituts gemappt werden können. Dass dies nicht der Fall sein konnte, war schon vor dem Vorhaben klar, da andere Länder andere Schulsysteme haben, die im erarbeiteten Klassifikationssystem keine Abbildung finden. Weitere Gründe wurden bei der Betrachtung der internationalen Sammlungen offenbart.

3.1 Individuelle Felder in den Sammlungen

Da die Erzeuger der Sammlungen keine Bibliothekare waren, wurden zum Teil nur Eigenschaften erfasst, die für Schulbücher im Speziellen, aber nicht Bücher im Allgemeinen wichtig sind. Beispielhaft trennt ein Bibliothekar Titel, Untertitel und Titelnachsatz, wobei ein „Laie“ diese Informationen als ein Ganzes aufnehmen würde, weil die Trennung im Schulbuchkontext unwichtig erscheint.

3.2 Interpretation individueller Felder

Eine weitere Herausforderung stellten Felder dar, deren Bezeichnung nicht eindeutig waren, sodass sie von verschiedenen Personen unterschiedlich interpretiert wurden. Beispielhaft wurde das Feld „Verwendung“ bei MANES als „Verwendung für“ und als „Verwendung als“ interpretiert, was dazu führte, dass dort zum einen Personen und Gruppen, sowie Dokumenttypen vorzufinden waren.

3.3 Unterschiede im Detaillierungsgrad

Bei der Betrachtung der Sammlungen zeigte sich der Nachteil von Feldern, die mit nicht vordefinierten Einträgen beschrieben werden konnten. Das äußerte sich zum Beispiel in 155 Schulfächern bei EDISCO und 85 Schulfächern bei MANES die den 15 Schulfächern des Klassifikationssystems gegenüber standen. Auch wenn ein großer Teil der Varianten Schreibfehlern geschuldet war, blieben nach der Bereinigung 86 neue Kategorien übrig, die berücksichtigt werden mussten. Dies führte zu grundsätzlichen Entscheidungen, ob man Eigenschaften generalisieren (weniger Optionen), spezialisieren (alle Optionen) oder nicht abbildbare Eigenschaften einfach ignorieren sollte. Das Generalisieren oder Ignorieren würde zum Verlust an Informationen führen, während das Spezialisieren den Nachteil birgt, dass man jede Ausprägung (auch zukünftige) in jede unterstützte Sprache (eindeutig) übersetzen muss.

3.4 Einträge lassen sich kaum mappen

Aufgrund unterschiedlicher Bildungssysteme sind Felder wie die „Schulform“ nicht aufeinander abbildbar. Im Rechercheinstrument würde ein Filtern auf ein solches Feld meist nur zu Ergebnissen aus der speziellen Sammlung führen, was nicht gewollt war. Zudem wurden internationale Schulbücher in der Sammlung des Georg-Eckert-Instituts nicht hinsichtlich der „Schulform“ katalogisiert.

3.5 Unterschiedliches Verständnis der Materie

Der eigentliche Vorteil, dass man reine Schulbuchsammlungen zusammenführt, offenbart die Frage, ab wann ein Buch ein Schulbuch ist. Dass Mathematikbücher Schulbücher sind, auch wenn sie kein Bestandteil traditioneller Schulbuchforschung sind, ist klar ersichtlich. Anders verhält es sich bei Curricula oder wissenschaftlichen Abhandlungen zur Schulbuchforschung, bei denen die Nutzer des Rechercheinstruments entscheiden müssten, ob diese Bücher recherchierbar sein sollten. Sind Kochbücher, die in EDISCO zu finden sind, Schulbücher?

3.6 Generalisierung

Zusammenfassend lassen sich die Beobachtungen wie folgt generalisieren:

1. Datensatzfelder tragen oft Informationen für mehrere Felder.
2. Freitextfelder begünstigen Schreibfehler und mehrere Schreibvarianten für eine Ausprägung.
3. Datensatzfelder sind oft in der Muttersprache formuliert.
4. Ein bestehendes Klassifikationssystem kann zu speziell, aber auch zu generell sein.

Der Idealfall sähe wie folgt aus:

1. Datensatzfelder finden ihre genauen Entsprechungen in bibliographischen Datenformaten wie PICA⁶ oder MARC21⁷.
2. Belegungen der Datensatzfelder werden über ein festes Vokabular bestimmt.
3. Belegungen der Datensatzfelder werden durch Codes beschrieben, die sich nach internationalen Standards richten.
4. Existierende Klassifikationssysteme lassen sich auf Codes internationaler Standards abbilden.

4 Vereinheitlichtes Klassifikationssystem

Das Klassifikationssystem beschreibt Klassen und deren Ausprägungen als Codes. Es soll garantieren, dass die beschriebenen Ausprägungen der relevanten Eigenschaften der Sammlungen im vereinigten Rechercheinstrument nutzbar sind. Auch wenn die Belegungen der Datensatzfelder nicht als Codes vorliegen, so lassen sie sich dennoch in solche überführen [7].

⁶ <https://www.gbv.de/de/katricht/inhalt.shtml>

⁷ <https://www.loc.gov/marc/>

4.1 Erstellen des Klassifikationssystems

Der erste Schritt im Erstellen des Klassifikationssystems kann nur, wie in Abschnitt 3 beschrieben, das Untersuchen der Sammlungen sein. Idealerweise kann man den zu betrachtenden Datensatz in eine csv-Datei exportieren und ein Tabellenkalkulationsprogramm oder OpenRefine⁸ nutzen, um sich einen Überblick zu verschaffen. Erfahrungsgemäß folgt als zweiter Schritt das Aufräumen der Daten, das zwar der Sammlung dient, aber für das Abbilden ins Klassifikationssystem keine Rolle spielt, da auch Schreibfehler und Schreibvarianten auf Codes abgebildet werden können.

Nachdem jede relevante Klasse bestimmt, beziehungsweise im Datensatz identifiziert wurde, müssen alle Ausprägungen notiert werden. Jede Ausprägung, die nicht schon Teil der Klasse ist, wird mit einem neuen Code ins Klassifikationssystem aufgenommen. Dieser Schritt ist nicht trivial, da entsprechende Sprach- und Fachkenntnisse vorausgesetzt werden, um zu erkennen, ob zwei Ausprägungen in zwei Sammlungen die gleiche Eigenschaft darstellen.

In der Praxis sollte man diesen Schritt aufgrund des Aufwands infrage stellen und schauen, ob es nicht ratsamer ist, eine Klasse durch existierende Standards zu beschreiben (siehe Abschnitt 4.2).

Ein intensives Beschäftigen mit einer Klasse führt in einigen Fällen zur Notwendigkeit, Klassen mit Unterklassen erweitern zu müssen, sodass das Klassifikationssystem im Prinzip baumartig repräsentiert werden müsste. Stattdessen sollten Ober- und Unterklassen durch die Codestructur repräsentiert werden (siehe Abbildung 2), bei der Teile des Codes auf die Klasse und Unterklasse(n) einer Ausprägung schließen lassen.

Ein einmal festgelegtes Klassifikationssystem sollte möglichst nicht um neue Klassen erweitert werden, da die bestehenden Sammlungen höchstwahrscheinlich nicht in diese Klasse abgebildet werden können. Anders verhält es sich bei den Ausprägungen bestehender Klassen, denn vorher unbekannte Eigenschaftsausprägungen sollten mit neuen Codes in das Klassifikationssystem aufgenommen werden. Ein Neustrukturieren mit Unterklassen ist möglich, wenn sich die bisher vergebenen Codes diesen Unterklassen zuweisen lassen, aber man muss aufpassen, da sich die Codes aufgrund der Unterklasse gegebenenfalls ändern.

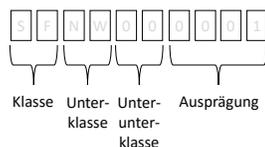


Abbildung 2. Klassen, Unterklassen und potentielle weitere Unterklassen lassen sich direkt im Code abbilden, was eine lineare Repräsentation im Klassifikationssystem ermöglicht. Beispielhaft könnte der Code `SFNW00001` für „Schulfach > Naturwissenschaften > Mathematik“ stehen.

⁸ <http://openrefine.org>

4.2 Klassifikation durch Standards

Es wird dringend dazu geraten Standards zu nutzen oder nach existierenden Klassifikationen Ausschau zu halten, da in deren Erstellung Expertisen und vor allem Zeit geflossen sind. Standards ermöglichen es, die Erstellung des eigenen Klassifikationssystems entscheidend zu beschleunigen.

Der Standard für Länder ist die ISO-3166-1, für subnationale Einheiten (wie Bundesländer) die ISO-3166-2, für Sprachen die ISO-639-2. Für anwendungsspezifische Klassen sollte man eine handvoll Ausprägungen als Suchbegriffe verwenden, um mithilfe einer Suchmaschine festzustellen, ob jemand bereits eine Klassifikation erstellt hat.

Ob im endgültigen System dann alle durch einen Standard definierten Ausprägungen genutzt werden ist zweitrangig, da das Rechercheinstrument nur genutzte Ausprägungen in Facetten anzeigen wird. Zeitgleich sollten keine nicht vergebenen Ausprägungen weggelassen werden, weil unklar ist, welche Datensätze mit welchen Eigenschaften zu einem späteren Zeitpunkt repräsentiert werden müssen.

Es ist ratsam, einen gefundenen Standard nicht mit dessen existierenden Codes zu beschreiben, sondern eigene Codes zu vergeben, die auf diesen Standard abbilden (vgl. Abbildung 2). Der Grund hierfür ist die Erweiterbarkeit, falls ein Standard nicht ganz genau auf die gewünschte Materie abbildet.

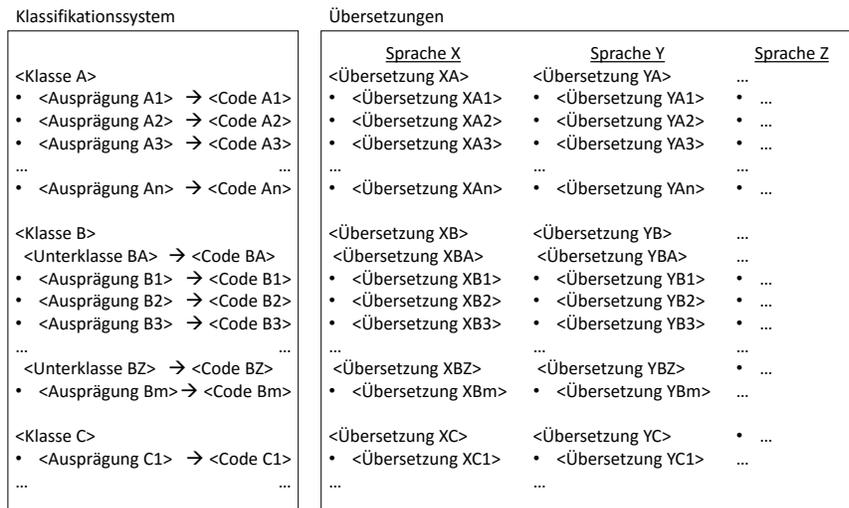


Abbildung 3. Struktur des Klassifikationssystem im Zusammenspiel mit Übersetzungen

4.3 Übersetzungen

Das fertige Klassifikationssystem besteht aus Klassen und Codes. Diese müssen für jede Sprache in entsprechende Begriffe abbildbar sein, was eindeutige Übersetzungen benötigt. Hat man für die Klassifikation auf ISO-Normen zurück gegriffen, sollten entsprechende Übersetzungen verfügbar sein.

Für die Sprachen der Sammlungen, für die die Codes vergeben wurden stehen diese Begriffe ebenfalls zur Verfügung. Für andere Sprachen müssen die Übersetzungen mit hohem Aufwand zusammengetragen werden.

Da das Rechercheinstrument intern nur mit den Codes arbeitet und Übersetzungen nur in der Weboberfläche dargestellt werden, ist eine (vorläufige) Übersetzung durch automatisierte Systeme⁹ besser als keine Übersetzung. Jede Übersetzung kann jederzeit überarbeitet und so Mehrdeutigkeit eliminiert werden.

5 Mapping

Nach dem Festlegen des Klassifikationssystems müssen alle relevanten Ausprägungen der Sammlungen auf deren Entsprechungen im Klassifikationssystem abgebildet werden, bevor sie in einem Suchindex zusammengefasst werden können.

5.1 Dialoggestütztes Mapping

Da die Mapping-Regeln für jede neue Sammlung individuell erstellt und für den Indexierer in digitaler Form vorliegen müssen, haben wir ein System entwickelt, das den Prozess des Integrierens, über das Mapping, Indexieren und Testen unterstützt. Das Erzeugen der Mapping-Regeln wird dabei durch eine Weboberfläche unterstützt und kann jederzeit überarbeitet werden.

Um diesen Prozess zu unterstützen verlangen wir, dass jede Partnerdatenbank ihre Daten über eine standardisierte Schnittstelle (Z39.50, OAI-PMH) anbieten muss. Dies wird üblicherweise durch die Verwendung eines aktuellen Bibliothekssystems gewährleistet. Während der Indexierer über diese Schnittstelle die aktuellen Dokumentinformationen gewinnt und auf ihnen die Mapping-Regeln anwendet, werden für das Erstellen der Mapping-Regeln erst einmal nur alle Attribute und deren Ausprägungen gesammelt und nach Häufigkeit sortiert.

Der Dialog führt durch die domain-spezifischen Klassen (die im Rechercheinstrument als Facetten repräsentiert werden) und fragt, welches Feld einer bestimmten Klasse entspricht. Wie in Abbildung 4 zu sehen ist, werden die ermittelten Ausprägungen dieses Feldes nach der Häufigkeit sortiert präsentiert, um das Mapping festlegen zu können. Ziel ist, dass jedes Attribut eine Entsprechung im Klassifikationssystem findet. Mehrere Ausprägungen dürfen dabei auf das selbe Feld abgebildet werden. Eine Ausprägung kann aber auch auf mehrere Felder im Klassifikationssystem abgebildet werden.

⁹ wie Google Translate <https://translate.google.com>

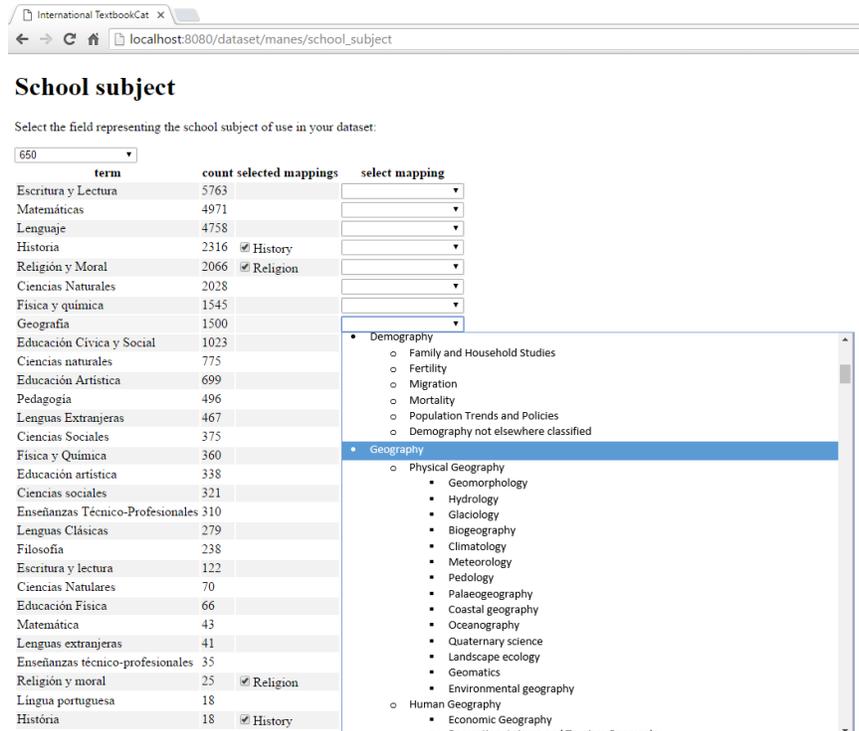


Abbildung 4. Dialoggestütztes Erstellen der Mapping-Regeln von einer Sammlung in das Klassifikationssystem.

5.2 Vom Mapping zum Suchindex

Mit den gegebenen Mapping-Regeln kann ein Indexierer ohne Umwege über eine weitere Repräsentation der Sammlungen einen homogenen und in Teilen sprachunabhängigen Suchindex aufbauen.

6 Zusammenfassung und Ausblick

Bei der Überführung vom institutionellen Rechercheinstrument *TextbookCat* zum internationalen Rechercheinstrument *International TextbookCat* musste eine allgemeingültige Repräsentationsstruktur gefunden werden, die internationale Schulbuchsammlungen nahezu vollständig abbilden kann. Der Schulbuchforschung wurde so ein Instrument zur Hand gegeben, welches die internationale Dimension in Schulbüchern besser beleuchten kann.

Wir haben gezeigt, wie ein Klassifikationssystem aussehen und welche Schritte man bei der Erstellung dieser Repräsentationsstruktur befolgen sollte. Länger-

fristig kann ein solches Klassifikationssystem als Standard verwendet werden, wenn es sich bewähren sollte.

Des Weiteren haben wir gezeigt, wie ein Dialogsystem eingesetzt werden kann, um komplette Sammlungen im erstellten Klassifikationssystem abzubilden.

Durch die Nutzung von Klassifizierungs-Codes als Beschreibung der Eigenschaften der Bücher ist das Rechercheinstrument sprachunabhängig, da diese Codes erst im Web-Browser in Begriffe der jeweiligen Sprache umgewandelt werden. Zur kompletten Multilingualität fehlt jedoch eine entsprechende Behandlung der textuellen Suchanfragen und der Volltexte.

Sobald Daten gänzlich durch Codes beschrieben werden können, ist das Verknüpfen mit Ontologien trivial. In diesem Zustand können dann semantische Verfahren aufsetzen, um einen noch größeren Mehrwert für die Schulbuchforschung zu generieren.

Literatur

1. Calhoun, K., Cellentani, D., et al.: Online Catalogs: What Users and Librarians Want: an OCLC report (2009), <http://www.oclc.org/reports/onlinecatalogs/fullreport.pdf>
2. Choppin, A.: EMMANUELLE: a data base for textbooks' history in Europe. Historical Social Research 14(4), 52-58 (1989), <http://nbn-resolving.de/urn:nbn:de:0168-ssoar-51666>
3. Choppin, A.: The Emmanuelle Textbook Project. Journal of Curriculum Studies 24(4), 345-356 (1992), <http://dx.doi.org/10.1080/0022027920240404>
4. Fiedler, M., Scheel, C., Weiß, A., De Luca, E.: Welt der Kinder. Semantisches Information Retrieval als Zugang zu Wissensbeständen des 19. Jahrhunderts. In: FVWG und GCDH Workshop für Wissenschaftsgeschichte und Digital Humanities in Forschung und Lehre (2016)
5. Fuchs, E., Kahlert, J., Sandfuchs, U.: Schulbuch konkret: Kontexte - Produktion - Unterricht. Klinkhardt (2010)
6. Fuchs, E., Niehaus, I., Stoletzki, A.: Das Schulbuch in der Forschung: Analysen und Empfehlungen für die Bildungspraxis. Eckert. Expertise, V&R Unipress (2014), <http://www.gei.de/publikationen/eckert-expertise/>
7. de Groat, G.: Future Directions in Metadata Remediation for Metadata Aggregators. Tech. rep., Digital Library Federation (2009)
8. Mayr, P., Petras, V.: Cross-concordances: terminology mapping and its effectiveness for information retrieval. CoRR abs/0806.3765 (2008), http://archive.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf
9. Ossenbach, G.: Research about school handbooks in Latin America: The project MANES contribution. Historia de la Educación 19(0) (2013), <http://revistas.usal.es/index.php/0212-0267/article/view/10797>
10. Ossenbach, G.: Textbook databases and their contribution to international research on the history of school culture. History of Education & Children's Literature IX(1), 163-174 (2014)

Sequential Modeling and Structural Anomaly Analytics in Industrial Production Environments

Martin Atzmueller¹ and Andreas Schmidt¹ and David Arnu²

¹ University of Kassel, Research Center for Information System Design
{atzmueller, schmidt}@cs.uni-kassel.de

² RapidMiner GmbH, darnu@rapidminer.com

Abstract. The analysis of sequential data is a prominent research topic, e. g., for investigating actions, events or log entries. This demonstration paper presents an integrated approach for anomaly analytics in an industrial production scenario. Based on first-order Markov chain models, we analyse sequential trails relative to specific hypotheses in an industrial application context. We summarize the applied method and present its implementation in a data analytics process.

1 Introduction

In many industrial areas, production facilities have reached a high level of automation. Here, knowledge about the respective processes is crucial, e. g., targeting the topological structure of a plant, sequences of operator notifications (alarms), and unexpected (critical) situations. Then, the analysis of (exceptional) sequential patterns is an important task for obtaining insights into the process and for modelling predictive applications.

Context. The BMBF funded research project “Early detection and decision support for critical situations in production environments”³ (short FEE) aims at detecting critical situations in production environments as early as possible and to support the facility operator in handling these situations, e. g., [9]. In abnormal situations, typically such a large number of notifications is generated, that it often cannot be physically assessed by the operator [31]. Therefore, appropriate abstractions and analytics methods are necessary in order to adapt and to change from a reactive to a proactive behaviour. The consortium of the FEE project consists of several partners also including application partners from the chemical industry. These partners provide the use cases for the project and background knowledge about the production process which is important for designing suitable analytical methods.

Objectives. This paper presents sequential modelling and anomaly analytics in an industrial application context. We present the implementation of a comprehensive modelling approach for comparing hypotheses with observed “reference” sequential patterns, based on methods for modelling and comparing networks and transition matrices, in particular the HYPGRAPHS [12] and DASHTrails approaches [11]. Then, we aim to identify deviating (abnormal/anomalous) and conforming (normal) hypotheses. Implemented as a new RapidMiner operator and embedded in an analytical process, we demonstrate the application (cf., Section 4) of the proposed approach.

³ <http://www.fee-projekt.de>

2 Related Work

The investigation of sequential patterns and sequential trails are interesting and challenging tasks in data mining and network science, in particular in graph mining and social network analysis, e. g., [4,7]. A general view on modeling and mining of ubiquitous and social multi-relational data is given in [5] focusing on social interaction networks. Here, dynamics and evolution of contacts patterns [8,16,20], for example, and their underlying mechanisms, e. g., [23] are analyzed. However, the analysis in these contexts focuses on aggregated sequential data. Navigational patterns, as sequential (link) patterns in online systems, have been analysed and modelled, e. g., in [25,29]. In contrast to that, our approach focuses on modelling and comparing sequential patterns (hypothesis) in a graph-based network representation.

For comparing hypotheses and sequential trails, the HypTrails [28] algorithm has been proposed. In [11] we have presented the DASHTrails approach that incorporates probability distributions for deriving transitions utilizing HypTrails. Based on that, the HYPGRAPHS framework [12] provides a more general modeling approach. Using general weight-attributed network representations, we can infer transition matrices as *graph interpretations*, while HYPGRAPHS consequently also relies on first-order Markov chain modeling [19,29] and Bayesian inference [29,30].

Sequential pattern analysis has also been performed in the context of alarm management systems, where sequences are represented by the order of alarm notifications. Folmer et al. [14] proposed an algorithm for discovering temporal alarm dependencies based on conditional probabilities in an adjustable time window. To reduce the number of alarms in alarm floods Abele et al. [2] performed root cause analysis with a Bayesian network approach and compared different methods for learning the network probabilities. Vogel-Heuser et al. [31] proposed a pattern-based algorithm for identifying causal dependencies in the alarm logs, which can be used to aggregate alarm information and therefore reduce the load of information for the operator. In contrast to those approaches, the proposed approach is not only about detecting sequential patterns. We provide a systematic approach for the analysis of (derived) sequential transition matrices and its comparison relative to a set of hypotheses. Thus, similar to evidence networks in the context of social networks, e. g., [22], we model transitions assuming a certain interpretation of the data towards a sequential representation. Then, we can identify important influence factors

Process Mining [1] aims at the discovery of business process related events in a sequence log. The assumption is that event logs contain fingerprints of a business process, which can be identified by sequence analysis. One task of process mining is conformance checking [24,27] which has been introduced to check the matching of an existing business process model with the a segmentation of the log entries. Compared to these approaches, we do not use any apriori knowledge about business processes to create our hypothesis. Furthermore, our hypothesis do not necessarily need to conform with an existing business process.

There are different definitions of an anomaly. According to the classical definition of [15], “an outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism”. Then, interesting, important or exceptional groups [3,26] can be identified. In contrast to approaches for

anomaly detection that only provide a classification of anomalous and normal events, we can assess different anomaly hypotheses: Applying the proposed approach, we can then generate an anomaly indicator – as a potential kind of second opinion method for assessing the state of a production plant that can help for indicating explanations and traces of unusual alarm sequences in the plant. Furthermore, using the network representation, we can analyze anomalous episodes relative to structural (plant topology) as well as dynamic (alarm sequence) episodes.

3 Method

Our application context is given by (abstracted) alarm sequences in industrial production plants in an Industry 4.0 context, cf., [31]. Specifically, we consider the analysis of the plant topology and anomaly detection in alarm logs. We formulate the “reference behaviour” collecting normal episodes as sequences of normal situations, which is typically observed for long running processes, and can also be simply validated by a domain expert or notes from the operator journals. Then, we compare episodes of alarm sequences (formulated as hypotheses) in order to detect deviations, i. e., abnormal episodes, and conforming ones corresponding to the “normal behaviour”. We map these sequences to transitions between functional units of an industrial plant, applying the modeling approach described below. The results can both be used for anomaly analytics as well as for diagnostics, by inspecting the transitions in detail.

3.1 Overview

Following [11, 12], we model transition matrices given a probability distribution of certain states. We assume a discrete set of such states Ω corresponding to the nodes of a network (without loss of generality $\Omega = \{1, \dots, n\}$, $n \in \mathbb{N}$, $|\Omega| = n$). For modelling, we consider a sequential interpretation (according to the first-order Markov property) of the original data with respect to the obtained transition probabilities (Markov chain).

As shown in Figure 1, we perform three steps, discussed below in more detail:

1. **Modeling:** Determine a transition model given the respective weighted network using a *transition modeling function* $\tau : \Omega \times \Omega \rightarrow \mathbb{R}$. Transitions between sequential states $i, j \in \Omega$ are captured by the elements m_{ij} of the transition matrix M , i. e., $m_{ij} = \tau(i, j)$. Then, we collect sequential transition matrices for the given network (data) and hypotheses.
2. **Estimation:** Apply HypTrails, cf., [28] on the given data transition matrix and the respective hypotheses, and return the resulting evidence.
3. **Analysis:** Present the results for semi-automatic introspection and analysis, e. g., by visualizing the network as a heatmap or characteristic sequence of nodes.

Thus, using τ , we can model (derived) transition matrices corresponding to the *observed data*, e. g., given frequencies of alarms on measurement points, as well as hypotheses on sequences of alarms. For data transition matrices, we need to map the transitions into derived counts in relation to the data; for hypotheses we provide (normalized) transition probabilities. As a simple transformation for normalization, we can, e. g., directly convert the weighted network using the defined transition modeling function (i. e., we convert the obtained values to probabilities by row-normalization).

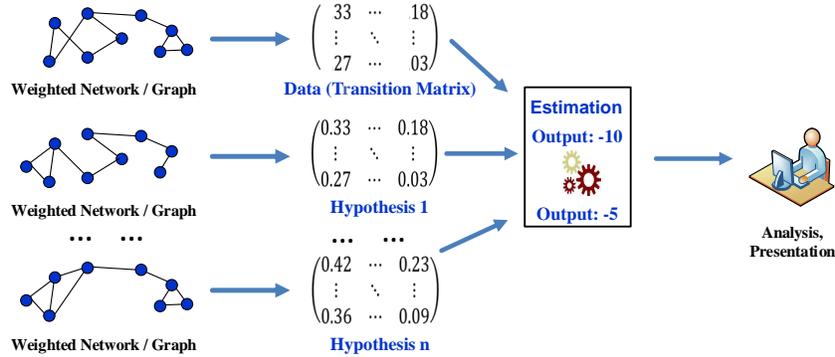


Fig. 1. Overview on the HYPGRAPHS modeling and analysis process, cf., [12] for more details.

3.2 Modeling

For explicitly observed sequences we can simply construct transition matrices counting the transitions between the individual states, e. g., corresponding to the set of alarms (or according abstractions for aggregating sets of alarms). Then, $\tau(i, j) = |suc(i, j)|$, where $suc(i, j)$ denotes the successive sequences from state i to state j contained in the sequence. For a derived data matrix, e. g., given by calculating similarities between log entries, we typically normalize the obtained transition values. In the (general) continuous case, for example, this can be achieved by a suitable transformation on the values, e. g., simply dividing by the minimum value, or by interpreting the obtained values according to some prior distribution, e. g., to the static distribution on the (static) network (for more complex processing see [11]).

For assessing a set of hypotheses that consider different transition probabilities between the respective states, we apply the core Bayesian estimation step of Hyp-Trails [28] for comparing a set of hypotheses representing beliefs about transitions between states. In summary, we utilize Bayesian inference on a first-order Markov chain model. As an input, we provide a (data) matrix, containing the transitional information (frequencies) of transition between the respective states, according to the (observed) data. In addition, we utilize a set of hypotheses given by (row-normalized) stochastic matrices, modelling the given hypotheses. The estimation method outputs an evidence value, for each hypothesis, that can be used for ranking. Also, using the evidence values, we can compare the hypotheses in terms of their significance. We refer to [11, 12, 28] for more details on modelling and inference, respectively.

As an alternative, we can apply the quadratic assignment procedure [18] (QAP) as a frequentist approach for comparing network structures. For comparing two graphs G_1 and G_2 , it estimates the correlation of the respective adjacency matrices [18] and tests a given graph level statistic, e. g., the graph covariance, against a QAP null hypothesis. QAP compares the observed graph correlation of (G_1, G_2) to the distribution of the respective resulting correlation scores obtained on repeated random row and column permutations of the adjacency matrix of G_2 . As a result, we obtain a correlation value and a statistical significance level according to the randomized distribution scores.

4 Process Model & Implementation

In the case of the FEE project and large-scale application in production, a distributed storage and computation system can handle the requirements of evaluating several years of production data. The RapidMiner [21] platform, e. g., offers with Radoop a simple integration to Hadoop systems and is able to build preprocessing and analytical processes on a local machine and transfers them to a big data environment. It is also Open Source and its functionality can be extended with self written code.

Overall, in the context of the FEE project our goal is to build a two layered computation architecture. Long running and computational expensive processes will run in the Hadoop infrastructure and either the prepared data or the final models, in this case the set of transition matrices, can be applied on a local machine. By running the computation on Spark/MapReduce [13], for example, and orchestrating the data with RapidMiner, the deployment process is speed up even further. Also the results can easily be visualized for the process engineer, for example as a heatmap of anomaly scores and be embedded in a dashboard, cf., Figure 2.

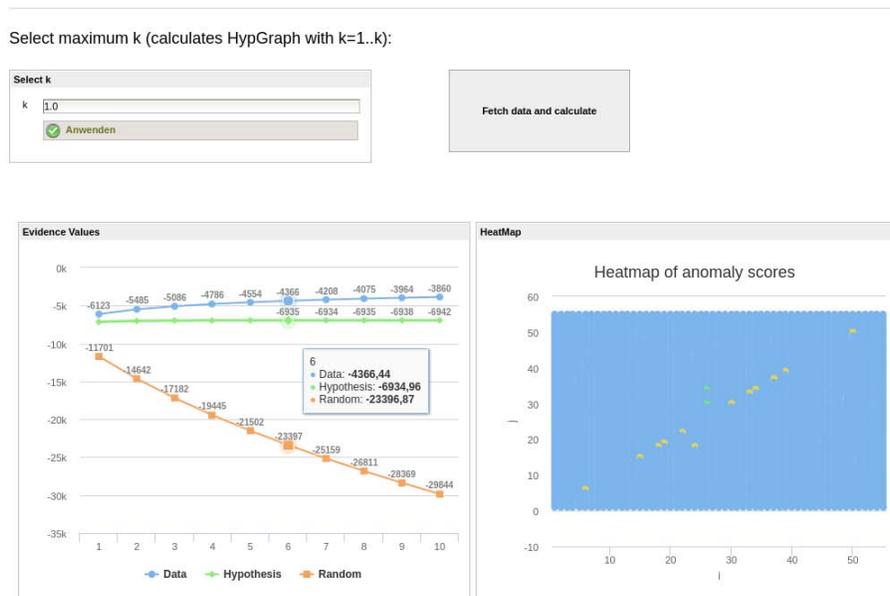


Fig. 2. RapidMiner Server Dashboard: By adjusting the parameter k we can tune our belief in the hypotheses, cf., [11]. Then, the resulting *evidence* values (depending on the parameter k , x-axis) allows for a ranking of these compared to the *data* and to a randomized hypothesis (null model).

For the process we can load the data from local storage, but chunks of the live data can easily be accessed with a database query. Figure 3 shows examples of sequential (abnormal and normal) transitions in a network visualization. Such a visualization can be used for data exploration or be embedded into the dashboard discussed above.

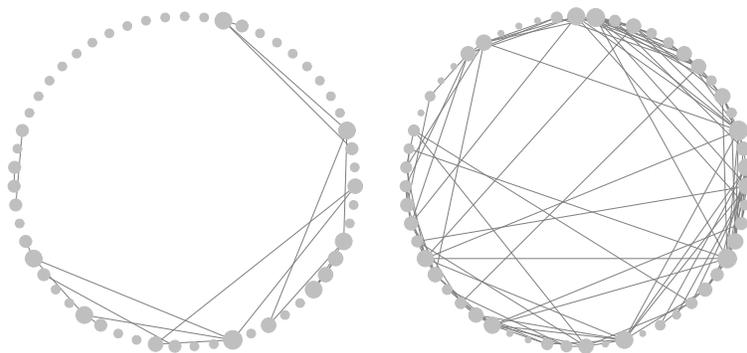


Fig. 3. Examples of transition network visualizations: Anomaly (left) vs. normal state (right). The nodes of the network denote aggregations of different alarm sources, where the size of a node denotes the shares of outgoing alarm notifications, and an edges denotes at least one transition (without self loops). These can also be filtered according to subsets of functional plant units. The figure shows such a case, where normal and abnormal situations show different characteristics and can be clearly distinguished.

Starting with process data from a streaming data source (or also historic data, e. g., as flat files) Figure 4 illustrates the inner loop of the data flow in the RapidMiner process. The calculated evidence values are collected in a table for further processing, for example in an interactive dashboard as shown in Figure 2. In the industrial context such a dashboard can be used as a standalone tool for analysing the production process online, or for evaluating historic data in order to optimize the process offline.

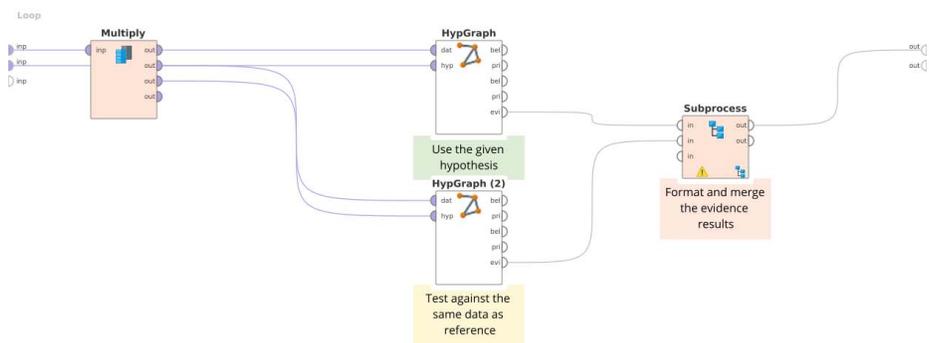


Fig. 4. Exemplary RapidMiner Process

Thus, a predefined RapidMiner process can simply be started by an engineer in order to get feedback of the current production state that can then be easily interpreted by the dashboard visualization.

5 Conclusions

This paper presented a sequential modelling and anomaly analytics approach in an industrial application context. Based on first order Markov chain models and methods for modelling and comparing networks and transition matrices [11, 12, 28], we sketched an approach for comparing hypotheses with observed “reference” sequential patterns. In that way, we can identify deviating (abnormal) and conforming (normal) hypotheses, in order to support anomaly analytics and diagnostics. Finally, we demonstrated the application of the proposed approach implemented as a RapidMiner operator.

For future work, we aim at extending the visualization options in order to support further introspective analytics options. For enabling (semi-automatic) techniques for detecting exceptional sequential trails and patterns, e. g., [6, 10], the according adaptation and extension of the presented approach in order to enable integrated anomaly detection and analysis seems promising. Furthermore, the development of comprehensive Big Data software architectures for plant operator support, e. g., [17] is another interesting direction for future research.

Acknowledgements

This work was funded by the BMBF project FEE under grant number 01IS14006E. We wish to thank Florian Lemmerich (GESIS, Cologne) for discussions on HypTrails [28].

References

1. Aalst, W.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin (2011)
2. Abele, L., Anic, M., Gutmann, T., Folmer, J., Kleinstauber, M., Vogel-Heuser, B.: Combining knowledge modeling and machine learning for alarm root cause analysis. In: MIM. pp. 1843–1848. International Federation of Automatic Control (2013)
3. Akoglu, L., Tong, H., Koutra, D.: Graph Based Anomaly Detection and Description. *Data Min Knowl Disc* 29(3), 626–688 (May 2015)
4. Atzmueller, M.: Analyzing and Grounding Social Interaction in Online and Offline Networks. In: Proc. ECML-PKDD. LNCS, vol. 8726, pp. 485–488. Springer, Heidelberg, Germany (2014)
5. Atzmueller, M.: Data Mining on Social Interaction Networks. *Journal of Data Mining and Digital Humanities* 1 (June 2014)
6. Atzmueller, M.: Detecting Community Patterns Capturing Exceptional Link Trails. In: Proc. IEEE/ACM ASONAM. IEEE Press, Boston, MA, USA (2016)
7. Atzmueller, M.: Local Exceptionality Detection on Social Interaction Networks. In: Proc. ECML-PKDD 2016: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Springer, Heidelberg, Germany (2016)
8. Atzmueller, M., Doerfel, S., Hotho, A., Mitzlaff, F., Stumme, G.: Face-to-Face Contacts at a Conference: Dynamics of Communities and Roles. In: Modeling and Mining Ubiquitous Social Media, LNAI, vol. 7472. Springer, Heidelberg, Germany (2012)
9. Atzmueller, M., Kloepper, B., Mawla, H.A., Jäschke, B., Hollender, M., Graube, M., Arnu, D., Schmidt, A., Heinze, S., Schorer, L., Kroll, A., Stumme, G., Urbas, L.: Big Data Analytics for Proactive Industrial Decision Support: Approaches First Experiences in the Context of the FEE Project. *atp edition* 58(9) (2016)

10. Atzmueller, M., Mollenhauer, D., Schmidt, A.: Big Data Analytics Using Local Exceptionality Detection. In: Enterprise Big Data Engineering, Analytics, and Management. IGI Global, Hershey, PA, USA (2016)
11. Atzmueller, M., Schmidt, A., Kibanov, M.: DASHTrails: An Approach for Modeling and Analysis of Distribution-Adapted Sequential Hypotheses and Trails. In: Proc. WWW 2016 (Companion). IW3C2 / ACM, New York, NY, USA (2016)
12. Atzmueller, M., Schmidt, A., Kloepper, B., Arnu, D.: HypGraphs: An Approach for Modeling and Comparing Graph-Based and Sequential Hypotheses. In: Proc. ECML-PKDD Workshop on New Frontiers in Mining Complex Patterns (NFMCP). Riva del Garda, Italy (2016)
13. Becker, M., Mewes, H., Hotho, A., Dimitrov, D., Lemmerich, F., Strohmaier, M.: SparkTrails: A MapReduce Implementation of HypTrails for Comparing Hypotheses About Human Trails. In: Proc. WWW (Companion). ACM Press, New York, NY, USA (2016)
14. Folmer, J., Schuricht, F., Vogel-Heuser, B.: Detection of temporal dependencies in alarm time series of industrial plants. Proc. 19th IFAC World Congr pp. 24–29 (2014)
15. Hawkins, D.: Identification of Outliers. Chapman and Hall, London, UK (1980)
16. Kibanov, M., Atzmueller, M., Scholz, C., Stumme, G.: Temporal Evolution of Contacts and Communities in Networks of Face-to-Face Human Interactions. Science China Information Sciences 57 (March 2014)
17. Klöpper, B., Dix, M., Schorer, L., Ampofo, A., Atzmueller, M., Arnu, D., Klinkenberg, R.: Defining Software Architectures for Big Data Enabled Operator Support Systems. In: Proc. INDIN. IEEE Press, Boston, MA, USA (2016)
18. Krackhardt, D.: QAP Partialling as a Test of Spuriousness. Social Networks 9, 171–186 (1987)
19. Lempel, R., Moran, S.: The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. Computer Networks 33(1), 387–401 (2000)
20. Macek, B.E., Scholz, C., Atzmueller, M., Stumme, G.: Anatomy of a Conference. In: Proc. ACM Hypertext. pp. 245–254. ACM Press, New York, NY, USA (2012)
21. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: Yale: Rapid prototyping for complex data mining tasks. In: Proc. KDD. pp. 935–940. ACM, New York, NY, USA (2006)
22. Mitzlaff, F., Atzmueller, M., Benz, D., Hotho, A., Stumme, G.: Community Assessment using Evidence Networks. In: Analysis of Social Media and Ubiquitous Data. LNAI, vol. 6904 (2011)
23. Mitzlaff, F., Atzmueller, M., Hotho, A., Stumme, G.: The Social Distributional Hypothesis. Journal of Social Network Analysis and Mining 4(216) (2014)
24. Munoz-Gama, J., Carmona, J., van der Aalst, W.M.P.: Single-entry single-exit decomposed conformance checking. Inf. Syst. 46, 102–122 (2014)
25. Pirolli, P.L., Pitkow, J.E.: Distributions of Surfers’ Paths Through the World Wide Web: Empirical Characterizations. World Wide Web 2(1-2) (1999)
26. Ranshous, S., Shen, S., Koutra, D., Harenberg, S., Faloutsos, C., Samatova, N.F.: Anomaly Detection in Dynamic Networks: A Survey. WIREs: Comput. Statistics 7(3), 223–247 (2015)
27. Rozinat, A., Aalst, W.: Conformance Checking of Processes Based on Monitoring Real Behavior. Information Systems 33(1), 64–95 (2008)
28. Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Hyptrails: A Bayesian Approach for Comparing Hypotheses about Human Trails. In: Proc. WWW. ACM, New York, NY, USA (2015)
29. Singer, P., Helic, D., Taraghi, B., Strohmaier, M.: Memory and Structure in Human Navigation Patterns. PLoS ONE 9(7) (2014)
30. Strelhoff, C.C., Crutchfield, J.P., Hübler, A.W.: Inferring Markov Chains: Bayesian Estimation, Model Comparison, Entropy Rate, and Out-of-Class Modeling. Physical Review E 76(1), 011106 (2007)
31. Vogel-Heuser, B., Schütz, D., Folmer, J.: Criteria-based alarm flood pattern recognition using historical data from automated production systems (aps). Mechatronics 31, 89–100 (2015)

Ontology-based Communication Architecture Within a Distributed Case-Based Retrieval System for Architectural Designs

Viktor Ayzenshtadt^{1,3}, Ada Mikyas¹, Klaus-Dieter Althoff^{1,3},
Saqib Bukhari³, Andreas Dengel^{2,3}

¹University of Hildesheim, Institute of Computer Science
Samelsonplatz 1, 31141 Hildesheim, Germany

²Kaiserslautern University
P.O. Box 3049, 67663 Kaiserslautern, Germany

³German Research Center for Artificial Intelligence
Trippstadter Strasse 122, 67663 Kaiserslautern, Germany
{firstname.lastname}@dfki.de

Abstract The communication and cooperation of agents is one of the key features of the multi-agent systems theory. In this work we discuss how the agents can communicate by means of applying a domain-specific ontology for the purpose of case-based retrieval of similar architectural designs. The domain ontology and the corresponding communication patterns are parts of the communication architecture of the distributed case-based retrieval system MetisCBR. We also present a vision of the results explanation component that enhances the existing architecture with own patterns and concepts and is able to recognize the corresponding contexts in search results returned by the system.

Keywords: case-based design, multi-agent systems, ontology, communication

1 Introduction

In multi-agent systems, the communicative interconnection of agents available in the system is established by providing a communication module that is able to transport messages from one agent to another. Modern FIPA-compliant¹ multi-agent frameworks like JADE [5] support the ontology-based communication of agents [7]. This allows for a convenient way of implementing a communication and cooperation component that is based on a domain-specific ontology where concepts and relations can be appropriately selected for the given task.

In this work we present the communication architecture of MetisCBR [3], the distributed case-based retrieval system for search of architectural designs,

¹ The Foundation for Intelligent Physical Agents, <http://fipa.org>

developed in context of the Metis project (*Metis – Knowledge-based search and query methods for the development of semantic information models for use in early design phases*).² This interdisciplinary project was initiated by the DFKI (German Research Center for Artificial Intelligence) and the TUM (Technical University of Munich) and unites the research areas of computer-aided architectural design (CAAD), case-based reasoning (CBR), and multi-agent systems (MAS). The project is funded by the German Research Foundation (*Deutsche Forschungsgemeinschaft, DFG*).

This paper is structured as follows: first we present the related work in the area of ontology-based agent communication. In the next section we describe the current communication architecture of the system that consists of the communication ontology and the corresponding communication patterns. After that we present a vision of results explanation module and the corresponding modification of the communication architecture. Discussion and conclusion close this paper.

2 Related Work

To date much important work has been done in the domain of ontology-based multi-agent communication. In this section we shortly describe some of the papers that we consider inspirational and helpful for conceptualization and development of our communication architecture.

The work of Steels [13] discusses the creation mechanisms of ontologies in multi-agent systems by using a number of conventions adapted from biology (self-organisation, selectionism, and co-evolution). These mechanisms are applied to the agents domain in this paper. Steels' general conclusion for *co-evolution* is especially important for the purposes of our approach: a shared ontology emerges during the communication and possesses abilities of dynamism and incompleteness – dynamism allows for the extension of the ontology by new concepts, incompleteness implies the possibility of communication with different, yet compatible definitions.

Brena und Ceballos propose in [6] a hybrid approach that combines the centralization and distribution of ontologies within a multi-agent system. In this approach, a special ontology agent plays a role of a carrier of the complete ontology and delivers the needed parts of it to other agents of the system that implement only basic parts of the ontology and make requests for needed parts when required. This approach gave us an inspiration to keep the main (meta) information of the ontology centralized, and to distribute the parts for communication and explanation into two separate ontology modules (see Section 4).

For the structure of the communication ontology (see Section 3.1), we took the work of Zhan [14] as source of inspiration. In this work the *layered* ontology is applied to the product design and analysis domain, the advantages of such a structure (extensibility and domain-oriented efficiency) are described in [14] as well. We adapted the main idea of this approach for the purposes of our domain.

² *Metis – Wissensbasierte Such- und Abfragemethoden für die Erschließung von Informationen in semantischen Modellen für die Recherche in frühen Entwurfsphasen.*

3 MetisCBR Agents Communication

MetisCBR is a case-based search engine for retrieval of architectural building designs that uses the application of *Semantic Fingerprint* [9] patterns to the design instances during the search. The retrieval with fingerprint patterns (fingerprints) is related to the concept of similarity footprints described in [11], the fingerprints themselves are structured by means of applying the AGraphML specification [8] to the designs. The cases (semantically transformed building designs) in the case bases of MetisCBR are built with the specific domain model described in [2].

Being a distributed system, MetisCBR contains a number of (case-based) agents that are able to communicate with each other in order to coordinate their tasks, as well as to cooperate in order to achieve their common goal (find the most similar cases for a given design query). To establish the communication between the agents and to standardize the cooperation processes inside the system, a special *communication architecture* was developed that governs the normalization of the communication process. It currently consists of the specific *communication ontology* created for the domain of retrieval of architectural building designs, and the corresponding specific *communication patterns* that are based on the retrieval tasks of the system and its agents. In the following sections we present the structure of the communication ontology and the structure of the basic retrieval communication pattern.

3.1 The General Structure of the Communication Ontology

The communication ontology is based on the concepts of MetisCBR's domain model, but contains some additional features (for example, a number of concepts that are specific only for some particular agents or agent groups). The communication ontology is divided in three different layers (see Figure 1), where each of them is used during the corresponding step of the retrieval process:

- *Object Layer* – This layer represents the general concepts of the *query* and *result* objects that are being *received from* or *sent to* the user as the object that is created or parsed by the user interface that is connected to MetisCBR. Thus, this layer is used in the first and the last step of the retrieval.
- *Data Layer* – In this layer the query and result objects are decomposed into the data representations according to the CBR domain model [2] of the retrieval system. The architectural design concepts FLOORPLAN, ROOM and EDGE from the model will be represented with their corresponding ontological equivalents (*metadata, rooms and edges data*) and will be used during the actual retrieval steps of the complete retrieval process.
- *Action Layer* – This layer is responsible for representation of categories of actions that agents of the system are able to execute. For example, to parse and transform the query object into an ontological representation, resolve the query using the given retrieval strategy, forward the query (or its parts) to other agents, or to construct and save a concept instance that will be used to represent a case for the retention component of the coordination agent.

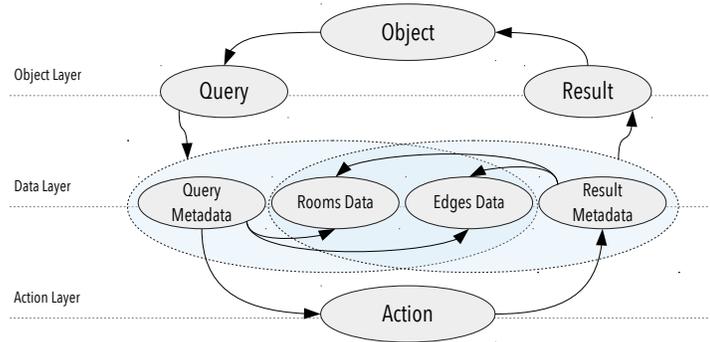


Figure 1. The current general structure of the communication ontology.

To utilize the ontology in order to communicate with each other, the agents of the system use communication patterns that are based on the concepts of the ontology. Communication patterns consist of steps that are named after the *action class* that contains the action the agent is requested to execute. The patterns can contain further sub-patterns. Following components are required for construction of a communication (sub-)pattern:

- *Action Class* – The category of the action to be assigned. Strategical system restrictions specify which actions an agent is free to execute when requested.
- *Actor* – The local identification address of the agent that is requested to perform the selected action.
- *Purpose* – The goal(s) of the action the agent is requested to accomplish, if it has committed or was assigned to this task.
- *Content* – An ontological object (for example the rooms data or a list of result floorplan IDs) that the agent uses as information source to accomplish its task. Can also contain further objects or references to objects.

3.2 Retrieval Communication Pattern

In this section we demonstrate how the agents communicate with each other using the communication patterns of the system. We show it by providing a detailed description of the steps of the general *retrieval communication pattern*. The communication flow of this pattern involves almost every agent type available in the system (except the case base maintainer agent). This pattern is the basic pattern of the communication and cooperation and uses almost all of the available action classes and ontology concepts to establish the undisturbed communication process during the retrieval. Figure 2 shows the graph-based representation of the structure of the pattern that consists of the following steps:

- *XHR* – The purpose of this action class is the transmission of the user query in XML format for later parsing and resolving. The actor (receiver of the query) in this case is the coordination agent (denoted as *Coord.* in Figure 2).

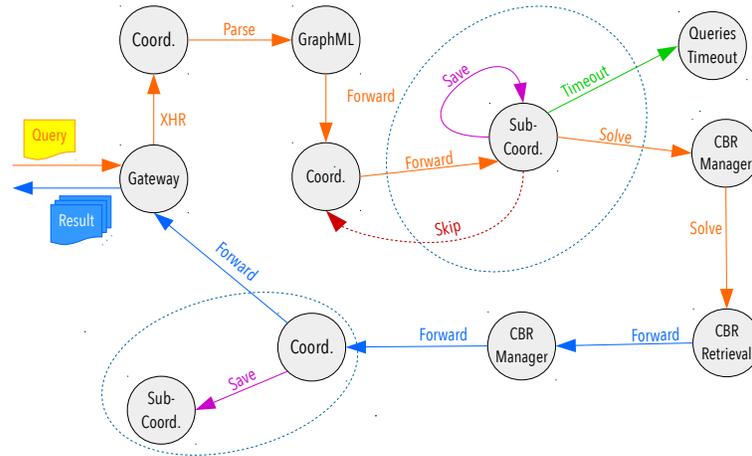


Figure 2. The graph-based representation of the *retrieval communication pattern*. The node labels denote the agents, the edge labels denote the action classes (steps).

- *Parse* – The receiver (actor) of the parsing request is the corresponding parsing agent (GraphML) that transforms the XML-formatted query into the ontological query in *SL* language format.
- *Forward* – This action class is intended to be used for forwarding of the content (ontological query or result) to other agents. Any agent can be a sender or receiver of this kind of task.
- *Solve* – The purpose of this action category is to request an appropriate agent to resolve the ontological building design query (or its parts) and to find results in a case base of such designs. A manager agent or a retrieval agent can be requested to accomplish this task.
- *Timeout* – The *Timeout* agent that receives the request containing this action class is asked to add the current retrieval task to the list of *active* tasks and to *check periodically* if the query is expired (not resolved during the given amount of time).
- *Skip* and *Save* – These two classes are related to each other and are intended to be used in the *case learning* sub-pattern (enclosed in the dashed areas in Figure 2), which is a sub-pattern of the *retrieval communication pattern* and is used during the retention step of the Coordinator’s internal CBR cycle, which is the step that helps to find the most similar previous query to the current one (this task is delegated to the *SubCoordinator*). The internal case base of Coordinator consists of the previous queries and is filled by means of applying the IB2 algorithm [1], in the way that only unique queries (and corresponding results) will be saved there. If the current query is identical to a previous one, the *Skip* action will be sent to the Coordinator to indicate this fact. The *Save* action class will only be used if the current query is *not identical* to a previous one. In this case the query will be saved before the actual retrieval starts, the results after the retrieval has been finished.

4 Vision of Results Explanation Module

Explanations are one of the core elements in user-centered CBR applications. Foundations, perspectives, and goals of explanations in CBR are described in [10] and [12]. In this section of the paper we present our vision of the extension of the MetisCBR system with an explanation module (see Figure 3) that contains its own explanation ontology. This module is currently being conceptualized in a bachelor thesis. Our general idea is to *combine two separate ontology modules* (the communication ontology and the new *explanation ontology*) into a *system ontology* (where the main meta information about these modules is kept permanently), but to use their concepts separately for the corresponding communication and explanation tasks. The explanation ontology will be used for the corresponding *explanation patterns* (that will have a structure similar to the communication patterns described in Section 3.2) and connected to a specific *explanation engine* that can use this patterns to work with different contexts (i.e., recognize if some results have one or more contexts as common criteria) and return an explanation of the retrieval results (based on these recognized contexts) to the user (an architect). The contexts can represent different semantic fingerprints or other criteria (for example, some of the floor plan results can belong to the same building, i.e., have the common building ID). It should be possible to have the permanent contexts (saved in the explanation ontology) as well as the temporary contexts that are specific only for the current search process.

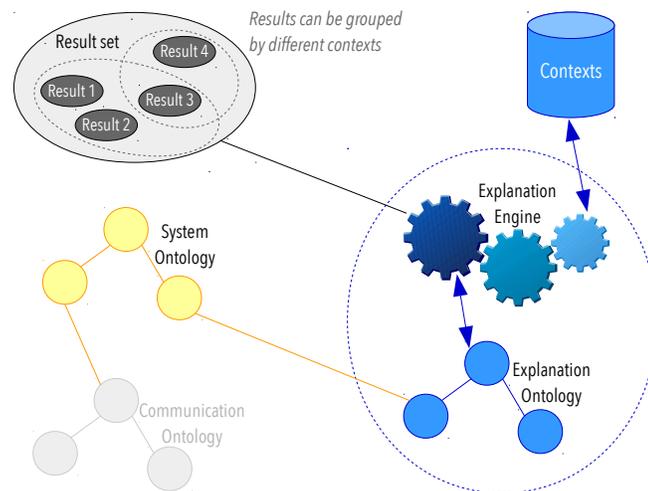


Figure 3. The current vision of the results explanation component for MetisCBR.

5 Discussion

The whole potential of the ontology-based agent communication architectures is not fully explored, but is used often to provide a basis for template- or pattern-based communication and cooperation among the agents contained in the system. In our retrieval system the ontology plays a role of the *layer-structured* relational vocabulary of objects and corresponding action classes that can be used by appropriate agents to request an action that needs to be executed for the current retrieval task.

In the evaluations of MetisCBR conducted to date (for example in [2] and [4]), and also during the development process, the ontology-based communication architecture showed a good performance (currently the size of the communication ontology does not allow for conducting of the performance test for the ontology only, so that the performance could only be estimated in context of the complete retrieval process, but no technical ontology-related issues worthy of mention were detected during the evaluations). The clearest advantage of such an architecture is the possibility to extend and restructure the underlying structure of concepts and actions by adding the new ones and/or deleting/editing the currently available ones. Though extensible, a certain technical limitation of the ontology scope exists as well, characterized by non-extensibility of the number of actions available for each of the agents at the runtime of the system.

6 Conclusion and Future Work

In this paper we presented the current communication architecture of MetisCBR, a distributed case-based retrieval system for search of semantically represented architectural designs. In this architecture the communication ontology plays a key role by providing a communication and cooperation basis for the agents of the system. We showed that the communication relies on the special communication patterns and provided and explained in detail an example of such a pattern (the basic *retrieval communication pattern*). We also provided our idea of how the concept of such a communication architecture can be used and adapted by an explanation module that is able to detect certain contexts in the result sets returned by the retrieval system, and how we can combine the ontologies of both communication and explanation.

In our future work on our case-based retrieval system we will concentrate on finalizing of the conceptualization of the above named explanation component and include it as a permanent part of the retrieval system and the corresponding retrieval process. Elaboration and extension of the available contexts, in order to improve the context recognition, will also be part of our future work in this area. The further development of other parts of the retrieval system, for example, implementation of new retrieval methods, will also be continued.

References

1. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. *Machine learning* 6(1), 37–66 (1991)
2. Ayzenshtadt, V., Langenhan, C., Bukhari, S.S., Althoff, K.D., Petzold, F., Dengel, A.: Distributed domain model for the case-based retrieval of architectural building designs. In: Petridis, M., Roth-Berghofer, T., Wiratunga, N. (eds.) *Proceedings of the 20th UK Workshop on Case-Based Reasoning*. UK Workshop on Case-Based Reasoning (UKCBR-2015), located at SGAI International Conference on Artificial Intelligence, December 15-17, Cambridge, United Kingdom. School of Computing, Engineering and Mathematics, University of Brighton, UK (2015)
3. Ayzenshtadt, V., Langenhan, C., Bukhari, S.S., Althoff, K.D., Petzold, F., Dengel, A.: Thinking with containers: A multi-agent retrieval approach for the case-based semantic search of architectural designs. In: Filipe, J., van den Herik, J. (eds.) *Proceedings of the 8th International Conference on Agents and Artificial Intelligence*. International Conference on Agents and Artificial Intelligence (ICAART-2016), February 24-26, Rome, Italy. SCITEPRESS (2016)
4. Ayzenshtadt, V., Langenhan, C., Roth, J., Bukhari, S.S., Althoff, K.D., Petzold, F., Dengel, A.: Comparative evaluation of rule-based and case-based retrieval coordination for search of architectural building designs. In: Goel, A., Roth-Berghofer, T., Diaz-Agudo, B. (eds.) *Case-based Reasoning in Research and Development*. International Conference on Case-Based Reasoning (ICCB-16), 24th International Conference on Case Based Reasoning, October 31 - November 2, Atlanta, Georgia, USA. Springer, Berlin, Heidelberg (2016)
5. Bellifemine, F.L., Caire, G., Greenwood, D.: *Developing multi-agent systems with JADE*, vol. 7. John Wiley & Sons (2007)
6. Brena, R., Ceballos, H.: A hybrid local-global approach for handling ontologies in a multiagent system. In: *Intelligent Systems, 2004. Proceedings. 2004 2nd International IEEE Conference*. vol. 1, pp. 261–266. IEEE (2004)
7. Caire, G., Cabanillas, D.: *Jade tutorial: application-defined content languages and ontologies*. TILab SpA (2002)
8. Langenhan, C.: A federated information system for the support of topological bim-based approaches. *Forum Bauinformatik Aachen* (2015)
9. Langenhan, C., Petzold, F.: The fingerprint of architecture-sketch-based design methods for researching building layouts through the semantic fingerprinting of floor plans. *International electronic scientific-educational journal: Architecture and Modern Information Technologies* 4, 13 (2010)
10. Roth-Berghofer, T.R.: Explanations and case-based reasoning: Foundational issues. In: *European Conference on Case-Based Reasoning*. pp. 389–403. Springer (2004)
11. Smyth, B., McKenna, E.: Footprint-based retrieval. In: *Case-Based Reasoning Research and Development*, pp. 343–357. Springer (1999)
12. Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning—perspectives and goals. *Artificial Intelligence Review* 24(2), 109–143 (2005)
13. Steels, L.: The origins of ontologies and communication conventions in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 1(2), 169–194 (1998)
14. Zhan, P.: An ontology-based approach for semantic level information exchange and integration in applications for product lifecycle management. Ph.D. thesis, Citeseer (2007)

Classification of German Newspaper Comments

Christian Godde and Konstantina Lazaridou and Ralf Krestel

Hasso-Plattner-Institut, Potsdam, Germany
Christian.Godde@student.hpi.uni-potsdam.de,
konstantina.lazaridou@hpi.de,
ralf.krestel@hpi.de

Abstract. Online news has gradually become an inherent part of many people's every day life, with the media enabling a social and interactive consumption of news as well. Readers openly express their perspectives and emotions for a current event by commenting news articles. They also form online communities and interact with each other by replying to other users' comments. Due to their active and significant role in the diffusion of information, automatically gaining insights of these comments' content is an interesting task. We are especially interested in finding systematic differences among the user comments from different newspapers. To this end, we propose the following classification task: Given a news comment thread of a particular article, identify the newspaper it comes from. Our corpus consists of six well-known German newspapers and their comments. We propose two experimental settings using SVM classifiers build on comment- and article-based features. We achieve precision of up to 90% for individual newspapers.

Keywords: media analysis, news comment analysis, comment classification

1 Introduction

Many online news sites offer their readers the possibility to comment on news articles either directly below the article in a forum-style way, or via Twitter or Facebook. While the latter is more suitable for sharing news, the former is more appropriate for discussion of the articles' contents. These online comments are huge reservoirs of user generated content with readers expressing opinions on various news-related topics. These range from comments on the article's style, specific arguments of the article, to general opinions about greater questions.

Not only does the discussion in these sections often reflect the readers' opinions about the article itself, but also about the overall topic and beyond, with readers referring to each other or introducing new arguments. Figure 1 shows excerpts of an article together with a comment and a reply to this comment. In general, the discussions are not limited to the specific article's topic and often introduce new arguments and opinions. Sentiments are expressed as well, towards either the content of the article or statements of other users. The content of one individual comment is not easily machine-understandable. It needs to be

evaluated in the context of the surrounding thread and associated article. Nevertheless, we argue that discussion style and topics may differ between various news providers, depending on their respective audience and possibly bias in the article’s coverage. For example, German newspapers and the majority of their readers are traditionally associated with a certain political alignment. If this is true, the political leaning should be reflected in the comment sections of the respective news sites as well. Even if the bias in the articles themselves is minimal, the reaction of the readers to the covered event may be much more diverse, which in return could be used to infer arguments for the political alignment of the news sites.

The image shows a screenshot of a news article and its comment section from ZEIT ONLINE. The article is titled "Assad würdigt deutsche Flüchtlingshilfe" (Assad honors German refugee aid) and is dated 1. März 2016. The article text states that the ceasefire in Syria is holding and that President Bashar al-Assad has promised an interview with the ARD. Below the article, there are two comment excerpts. Excerpt (a) is the article text. Excerpt (b) is a comment by "gutoderböse" which says: "Echt klasse, ein Diktator bekommt im ARD eine Plattform für seine Propaganda. An Zynismus kaum zu überbieten. Ich zähle mal einige Punkte auf: 1. Assad ist weder gewählt, noch regierte er demokratisch." Excerpt (c) is a reply by "Nuncio" which says: "Laut Seymour Hersh lieferte die Türkei mit Hilfe der USA Waffen aus Libyen nach Syrien und sogar das Sarin was in Ghuta zum Einsatz kam. http://www.lrb.co.uk/v36/..."

Fig. 1: Example of an article, comment, and reply from “Zeit”

In this paper we analyze the user comments on six major German news sites regarding their differences in discussion focus, language and sentiment. Based on the assumption that user comments on various news sites differ in these characteristics, we propose a classifier to predict the source of specific comments, that is, the news site on which the comments have been posted. To analyze this, a prediction method is developed and evaluated, which, given a set of user comments, predicts the originating news site.

2 Related Work

User comments can be found in different online platforms and communities. Social media platforms, such as Twitter, Facebook, and Youtube, are the most popular environment for users to generate personal content, share pieces of news, build social relations etc. Recent research focuses on analyzing comments’ content on these platforms, as well as analyzing the commenters. An extensive analysis [13] of comments in social media communities investigates comments’

sentiment, rating and popularity in Youtube videos and Yahoo! News posts. Momeni and Sageder [9] perform a comparative analysis of comments in Flickr and Youtube. The authors point out different textual, semantic, and topical features of the comments, which are later used to predict the comment’s usefulness. Towards identifying the characteristics of influential users, Martin et al. [8] introduce an emotion lexicon-based technique that predicts the helpfulness of reviews posted on Trip advisor and Yelp.

In addition to social media, related research focuses on news media as well. Here, understanding and potentially predicting the user characteristics and preferences is the main goal. The problem of user profiling in media is tackled in [1], where the authors introduce the notion of *comment-worthy* news articles. They predict the comments’ interestingness in blogs and news sites using an adapted topic model aiming at personalized recommendation of news articles to users. Similarly, Shmueli et al. [12] address the problem of ranking news comments according to the reader’s personal interests in Yahoo! News using a factor model. Instead of analyzing existing comments, Cao et al. [2] extract relevant microblog posts to news articles and use them to automatically generate user comments for these news articles.

Moreover, since users shape the general public’s opinion with their comments by often supplementing the news stories with new facts and expertise, approaches that automatically evaluate the comments’ quality have received high interest in the literature. To this end, tools distinguishing the (in)appropriate and (ir)relevant comments could assist media to improve the news quality they offer. Related work includes the analysis of the quality of comments [4], and the measurement of the comment sentiment in order to conclude about the media’s political leaning [10]. Additionally, the problem of comment relevance is addressed by [9], [3] and [5], with the latter assessing the degree of pertinence of comments by comparing their tf-idf vectors to the articles’ in News York Times. Detecting the comments that shift the main article topic and change the article’s focus at Digg.com is tackled by Wang et al. [15], while Zhang and Setty [16] identify sets of topic-wise diverse user comments in Reddit news articles.

Finally, multiple interesting prediction tasks emerge from news comments analysis. Among others, the volume of news comments is predicted with a random forest classifier by Tsagkias et al. [14] using a variety of comment and article metadata, as well as textual and semantic features derived from the comments. Rizos et al. predict news stories popularity based on users’ comments and the properties of the social graph they form [11]. Since users abuse the commenting mechanism frequently by stating offensive or hate comments, Kant et al. [7] compare an SVM classifier to a pattern mining approach in order to detect spam comments in Yahoo! News articles.

In contrast to the above works, we analyze comments to investigate differences in readership and bias among different German newspapers. Automatically gaining insights in the huge amount of user-generated content in media will help us discover people’s opinion over several issues. More specifically, the way readers perceive reality regularly depends on the different writing styles of different

news outlets and their respective journalists. For instance, it would be interesting to discover that users tend to leave more informative or insightful comments, when a newspaper is being brief and doesn't discuss thoroughly certain topics. Alternatively, a user may post funny or hate comments, when an article criticizes openly a person or an event.

Furthermore, the ability to identify a comment's origin is a step towards detecting correlations between the news providers and the news consumers. We share the intuition of [10] regarding media bias detection in news articles, that is, users tend to leave negative comments to articles that oppose their perspective and positive otherwise. Additionally, as introduced in [6], readers often choose to be informed by the sources that share their beliefs. Namely, one is more likely to perceive bias the further the slant of the news is from their own political position.

3 Predicting comments' original source

Our motivation stems from the idea that readers from different newspapers might use unique language and present different commenting patterns. There are indeed differences among users in news media in general: some users tend to be objective and include new facts to the articles, others leave subjective messages (e.g. supporting a party, an opinion), others may attack the journalist or comment writers with hate comments, etc. We are interested in whether the above styles are indicative of the comments' source or not. Hence, we aim at identifying the comment features that distinguish the users of different news outlets. This will allow us to classify comment threads belonging to certain newspapers. To this end, all the direct comments and comment replies in a given article are considered as a single document in our prediction task. That is, one document is the complete news comment thread of a given news article. We then use an SVM classifier to classify each instance to its respective newspaper. The feature selection and the parameter setting are described below.

3.1 Datasets

We analyzed six popular German newspapers, namely *Bild*, *Focus*, *Welt*, *Spiegel*, *Zeit* and *Faz*. The dataset characteristics are shown in Table 1. Our crawled data span from March 2016 until June 2016. The fifth column depicts the average comment length for each source after removing [stop words](#). It appears that *Spiegel* and *Faz* readers tend to leave longer comments than users from other sources. Additionally, we also observed that *Bild* commenters could be characterized as more active in comparison to the rest of the outlets, as the average number of comments per article in *Bild* is higher than in the rest of the newspapers.

Although the number of articles does not vary significantly among the newspapers, we can observe that *Welt* is the outlet with the most comments and commented articles in total. In our experiments, after considering all articles having at least 1, 5 or 10 comments in separate configurations, we concluded

that the threshold (H) of 5 yields the best precision results and thus we only report on results using this threshold. The last column in Table 1 represents the number of articles with at least 5 comments for each source.

Table 1: Dataset Characteristics

Source	Articles	Articles with ≥ 1 Comments	Comments	Average Comment Length	Articles with ≥ 5 Comments
Bild	1,358	316	11,332	21.6	186
Focus	1,764	965	2,651	58	80
Welt	1,852	1782	31,125	31.7	830
Spiegel	1,654	664	5,771	61.8	188
Zeit	1,045	1032	8,553	46.1	642
Faz	1,656	458	1,329	71.3	61

3.2 Features

This subsection describes the comment-based and article-based features that we use for the SVM classifier.

Number of comments and average comment length. The number of direct comments and comment replies are summed up representing the first dimension of the feature vector. In addition, the average comment length is calculated for each article after filtering out the terms that appear in our stop word list. As shown in Table 1, there are significant differences among the outlets regarding the volume of comments and their length. Hence, our intuition is that the above-mentioned features will constitute an important indicator for the respective news source.

Direct comment/reply ratio and distinct authors. The next two features refer to the users, regarding their activity and commenting behavior. The ratio between the direct comments and the nested ones is a numerical indicator of how interactive the commenters are and whether discussions are initiated by them or not. For instance, as illustrated in Figure 2, *Zeit* and *Bild* appear to have a higher number of user discussions than the other sources.

Moreover, the distinct number of authors per article is interesting as well, as it informs us about the comment availability and potential diversity. Articles with multiple commenters should contain a variety of opinions and statements, in comparison to stories that don't attract high user interest. Figure 3 presents the news articles that are covered by certain numbers of commenters. That is, e.g. around 90% of *Bild* and *Faz* news articles would be covered, if the top-30 commenters were considered. It should be also noted that for this plot we only

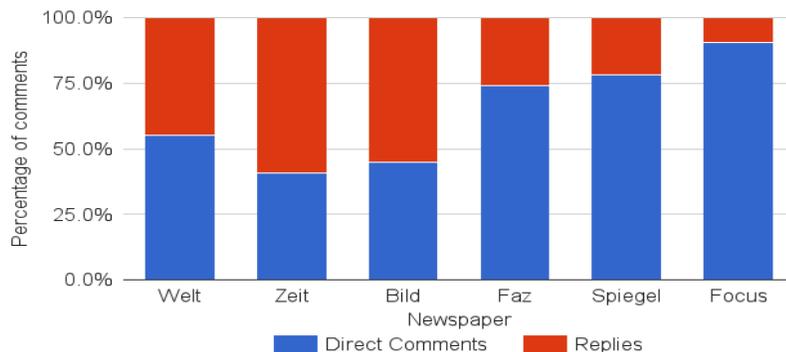


Fig. 2: Direct comments and nested replies for all news sources

use articles with H equal to 5. Our findings are in line with the work of Park et al. [10], where 50 commenters appear to cover around 80% of the overall dataset (when also considering solely articles with more than 5 comments).

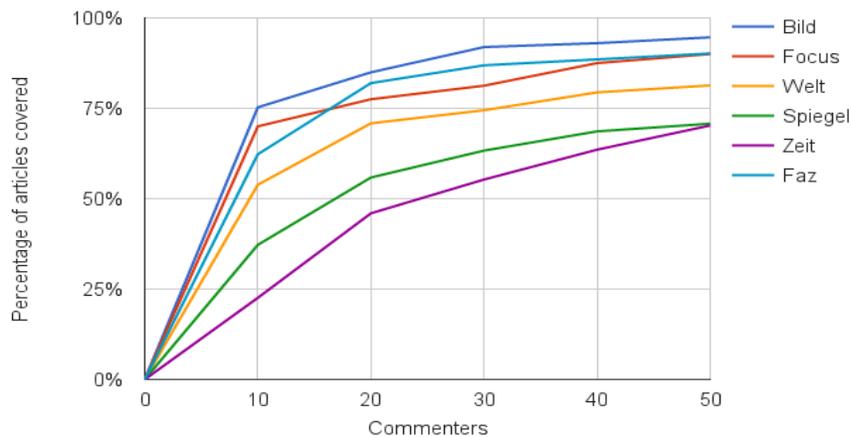


Fig. 3: Percentage of articles covered by k commenters, where $k = 10, 20, 30, 40$ and 50 .

Comment terms. The current feature targets the comments’ content and possibly opinionated language. We argue that the choice of language in the comments is the most representative feature of the users’ perspective. Some comments aim at pointing out neglected facts from the articles and others might criticize the article’s position or a politician’s behavior, etc. Figure 2 illustrates

an example of a comment in *Zeit* and one of its replies, where the two users express two different sides of the same story. It is notable that we only consider the terms' *tf-idf* scores that are not stopwords, since only these provide semantic and meaningful information about the users' interests.

Newspaper uniqueness metric. Apart from user features, newspapers' characteristics play a key-role to our prediction task as well. Towards discovering representative and specific language used by different newspapers, we measure the similarity between comments and news articles of all sources, in terms of their common words. We compare the comments' terms with the articles' terms from all sources and measure their *overlap coefficient*. That is, for each comment thread to be classified, we compute the *overlap* (or also known as Szymkiewicz-Simpson) *coefficient* between its terms and the overall vocabulary from the articles of each newspaper, which results in six separate numeric counts as individual features. Our intuition is that this metric indicates whether the journalists and the readers from a given newspaper mention the same words.

Since commenters are often subjective and emotional, the current feature might also extract words that are not expected to be found in news media. This word set is a possible bias indicator, considering that news articles are expected to publish objective and well-rounded news pieces, so that readers are adequately informed.

4 Topic analysis

To ensure that all articles/comments are comparable across media outlets, we analyze the topics discussed in each news outlet.

As a first step towards understanding the discussions in our data, we are interested in detecting the topics mentioned in the newspapers' articles during our given time frame. For this purpose we use the latent Dirichlet allocation (LDA) implementation in [Mallet](#), a Machine Learning Java Toolkit. We experiment with different values for the number of topics, namely 10, 20 and 40, but report only our findings for 20 topics, since the results are rather stable with varying topic numbers.

As shown in [Table 2](#), the most discussed topics (15, 0) among all newspapers are focused on local affairs, with $topic_0$ touching upon financial issues. The least mentioned topics (9, 11, 1, 10, 17, 16) concentrate more on foreign politics, especially U.S. politics, which is an emerging topic as the general elections are approaching in the U.S.

In addition, [Figure 4](#) presents the topic distributions across all newspapers. The x-axis represents the topics and the y-axis the volume of the discussion. One could infer that there are no extreme differences in the topic distributions among the outlets, that is, the same events/issues are covered by all newspapers. However, one notable exception are the comments in *Welt*, where the U.S. election topics (9,16,17) are clearly over represented.

Table 2: Top Terms for Each Topic (Ordered by Descending Popularity)

Topic Id	Frequent Terms
15	leben, politik, land, frage, deutschland, sagen, steht, kinder, sogar
0	prozent, deutschland, regierung, deutschen, zahl, land, praesident, frankreich, millionen
19	polizei, polizisten, frauen, demonstranten, koelner, maenner, verletzt, silvesternacht, koeln
7	euro, milliarden, deutschland, schaeuble, griechenland, geld, spd, gesetz, integration
3	spd, cdu, merkel, prozent, gabriel, afd, csu, seehofer, partei
5	russland, putin, usa, russischen, russische, praesident, obama, ukraine, nato
12	syrien, getoetet, stadt, waffenruhe, syrischen, terrormiliz, staat, aleppo, syrische
14	hofer, oesterreich, prozent, stimmen, fpoe, partei, wahl, parlament, van
4	afd, partei, deutschland, petry, islam, gruenen, cdu, kretschmann, npd
8	tuerkei, erdogan, boehmermann, tuerkischen, merkel, tuerkische, ankara, tayyip, recep
18	nordkorea, kim, journalisten, regierung, gericht, duendar, verurteilt, urteil, land
2	bruessel, anschlaegen, paris, anschlaege, flughafen, bruesseler, polizei, abdeslam, terroristen
13	panama, rousseff, papers, bundeswehr, zeitung, briefkastenfirmen, leyen, praesidentin, temer
6	cameron, khan, buergermeister, honecker, duterte, grossbritannien, london, johnson, britischen
16	trump, clinton, donald, sanders, republikaner, demokraten, hillary, cruz, vorwahlen
17	trump, clinton, sanders, donald, obama, hillary, prozent, cruz, trumps
10	the, waehler, and, twitter, primaries, staat, you, com, pic
1	trump, trumps, kasich, cruz, republikaner, senator, new, york, partei
11	fluechtlinge, tuerkei, griechenland, deutschland, grenze, fluechtlingskrise, migranten, fluechtlingen, europa
9	trump, sanders, clinton, cruz, rubio, donald, prozent, hillary, ted

Future work would be to incorporate this topical information in the classification task and discover whether it can improve our results, i.e. the users' commenting behavior differs for different combinations of topics and newspapers.

5 Classification results

The main goal of our work is to identify the newspaper that a certain comment thread comes from. Due to the small length of a single comment and the absence of rich content, we classify all the comments for a given article at once, instead

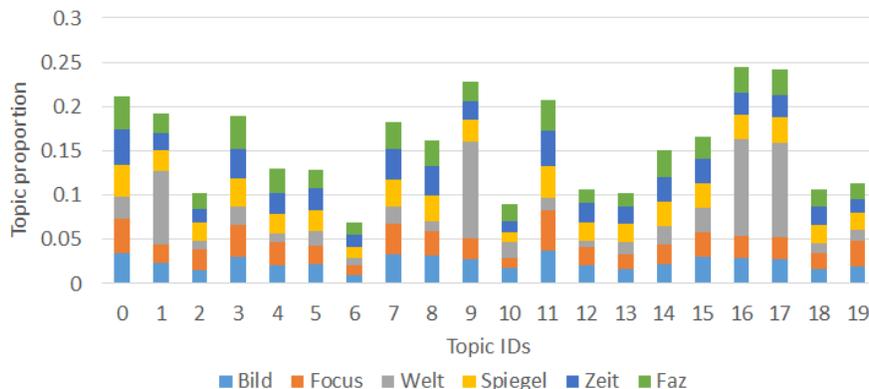


Fig. 4: Topic distributions for all sources using 20 topics

of considering them separately. For this purpose, we use the implementation of SVM classifier in [Weka](#) with the default parameter settings.

Regarding the training phase, we initially perform one-versus-one classification, training $m=k*(k-1)/2$ classifiers (one for each pair of newspapers) and output the majority vote among all classifiers for each input instance. Namely, we train the model with 40 documents per source and tested it on 20 documents per source — all randomly selected from our original dataset. Our second experiment is a one-versus-all classifier that is trained and tested on articles from all outlets, but it performs binary classification for a single given source. In particular, six different classifiers are built (one for each outlet) using 40 articles from the target source and 40 random articles from the remaining sources. The test set consists of 20 articles from the target news outlet and 20 arbitrary ones from the other outlets.

The above numbers of articles are set after examining the last column of Table 1. The maximum possible numbers are considered, in order to obtain a sufficient and equal amount of comments per source that will result to balanced training and test sets. Our future work includes obtaining more articles and subsequently more comments, fairly distributed to all six outlets, to achieve a higher comment quantity and diversity.

One-versus-one classification. The results of our first experiment are depicted in Table 3 and Table 4a. We can observe that the classifier performs best for *Bild* and yields inadequate results for *Focus* and *Zeit* with low recall or precision values respectively. The confusion matrix illustrated in Table 3 reveals that there is at least one comment from each source that is incorrectly classified as originating by *Zeit*. Considering that *Zeit* is the top-2 news outlet regarding the published number of articles with more than 5 comments, one might argue that highly popular and centrist newspapers, such as *Zeit*, contain a variety of

Table 3: One-Versus-One Classification Confusion Matrix

classified as →	a	b	c	d	e	f
a = Bild	19	0	0	0	1	0
b = Focus	0	8	0	0	5	7
c = Welt	2	0	12	0	5	1
d = Spiegel	1	2	2	11	3	1
e = Zeit	1	1	2	1	13	2
f = Faz	0	1	0	4	2	13

comments and commenter behaviors. This makes such news sources a good candidate for an unseen comment, as they could contain a wide range of different commenting styles.

Additionally, *Bild* articles are largely classified successfully. According to Table 1, *Bild* is also one of the sources with the most overall comments, whereas the average comment length is relatively very low. Observing Table 1 and Table 3 concurrently, one can distinguish that when taking into account the most right-wing sources, namely *Bild*, *Welt* and *Focus*, the lower the average comment length is the higher our precision result becomes. Since short user comments can often be sharp or pithy, this is an interesting observation for readers of right-wing newspapers.

The average achieved precision is 65% and average recall 63%. Although the average performance score is a promising start, there is significant room for improvement, which we will further discuss in the following paragraph.

Table 4: Classification Results

(a) One-Versus-One			(b) One-Versus-All		
Newspaper	Precision	Recall	Newspaper	Precision	Recall
Bild	0.82	0.95	Bild	0.85	0.80
Focus	0.66	0.40	Focus	0.83	0.80
Welt	0.75	0.60	Welt	0.73	0.72
Spiegel	0.68	0.55	Spiegel	0.74	0.70
Zeit	0.44	0.65	Zeit	0.80	0.75
Faz	0.54	0.65	Faz	0.90	0.90
Aaverage	0.65	0.63	Average	0.80	0.77

One-versus-all classification. Our next experiment is a one-versus-all classification. As previously mentioned, we build six different classifiers considering 40 articles from the target source and 40 random articles from the rest for the training set. The results are shown in Figure 4b. Surprisingly, although for the *Faz* articles the previous classifier achieved the worst results regarding precision, the

current classifier performs best for this particular outlet. The overall results vary from 73% (*Welt*) to 90% (*Faz*) precision. Moreover, recall is significantly higher, ranging from 70% (*Spiegel*) to 90% (*Faz*). This leads to an average precision of 80% and an average recall of 77%.

6 Conclusion and Future work

In this paper, we address the problem of automatically identifying the original newspaper that a comment thread of a given article belongs to. We analyze six well-known German newspapers, namely *Bild*, *Focus*, *Welt*, *Spiegel*, *Zeit* and *Faz*. To this end, we use an SVM classifier with different comment- and article-based features. For instance, the comment terms' *tf-idf* values and the number of available comments for an article are considered. The best results are accomplished by six one-versus-one classifiers (one for each newspaper pair), where our precision scores range from 70% to 90%. In order to reduce the variance of our results between the two classifiers and also among all news sources, we will perform the experimental evaluation on multiple random training and test sets.

Towards improving our current work, we would like to experiment with different feature combinations and evaluate their impact to our classifier. Apart from our version of measuring the unique characteristics of the newspapers, one could also attempt to take into account the levels of subjectivity in the news text, as an indication of the writing style. Moreover, the polarity (positive, negative, neutral) of each comment is a valuable information as well. It might hold that users in certain newspapers express their emotions more than in others or that users from specific outlets tend to express more their disapproval and criticism to certain issues than in other sources.

Instead of using all comment terms, we also experimented with using only the named entities found in the comments. The results were slightly worse than the reported ones, therefore we will continue to use all terms of the comments, as presented in this work. Although the named entities along with different feature combinations might work in the future, it is interesting to note that not only named entities are crucial for this problem, but verbs, adjectives and adverbs as well. Named entities mainly depict a text's topic, whereas adjectives and adverbs represent the author's perspective and discussion style. Finally, as previously mentioned, a direction we would like to follow is the incorporation of topical information in our classifier, which could potentially lead us to identify the original source of a comment thread more reliably.

References

1. Bansal, T., Das, M., Bhattacharyya, C.: Content Driven User Profiling for Comment-Worthy Recommendations of News and Blog Articles. RecSys pp. 195–202 (2015)
2. Cao, X., Chen, K., Long, R., Zheng, G., Yu, Y.: News comments generation via mining microblogs. In: WWW. pp. 471–472 (2012)

3. Das, M.K., Bansal, T., Bhattacharyya, C.: Going beyond Corr-LDA for detecting specific comments on news & blogs. In: WSDM. pp. 483–492 (2014)
4. Diakopoulos, N., Naaman, M.: Towards quality discourse in online news comments. In: CSCW. pp. 133–142 (2011)
5. Diakopoulos, N.A.: The Editor’s Eye. In: CSCW. pp. 1153–1157 (2015)
6. Groseclose, T., Milyo, J.: A measure of media bias. In: The Quarterly Journal of Economics. pp. 1191–1237 (2005)
7. Kant, R., Sengamedu, S.H., Kumar, K.S.: Comment spam detection by sequence mining. In: WSDM. pp. 183–192 (2012)
8. Martin, L., Sintsova, V., Pu, P.: Are influential writers more objective? In: WWW. pp. 799–804 (2014)
9. Momeni, E., Sageder, G.: An empirical analysis of characteristics of useful comments in social media. In: WebSci. pp. 258–261 (2013)
10. Park, S., Ko, M., Kim, J., Liu, Y., Song, J.: The Politics of Comments: Predicting Political Orientation of News Stories with Commenters’ Sentiment Patterns. CSCW (2011)
11. Rizos, G., Papadopoulos, S., Kompatsiaris, Y.: Predicting News Popularity by Mining Online Discussions. WWW pp. 737–742 (2016)
12. Shmueli, E., Kagian, A., Koren, Y., Lempel, R.: Care to comment? Recommendations for commenting on news stories. In: WWW. p. 429 (2012)
13. Siersdorfer, S., Chelaru, S., Pedro, J.S., Altingovde, I.S., Nejd, W.: Analyzing and Mining Comments and Comment Ratings on the Social Web. TWeb 8, 1–39 (2014)
14. Tsagkias, M., Weerkamp, W., de Rijke, M.: Predicting the volume of comments on online news stories. In: CIKM. pp. 1765–1768 (2009)
15. Wang, J., Yu, C.T., Yu, P.S., Liu, B., Meng, W.: Diversionary comments under political blog posts. In: CIKM. pp. 1789–1793 (2012)
16. Zhang, H., Setty, V.: Finding diverse needles in a haystack of comments. In: WebSci. pp. 286–290 (2016)

k-Means Clustering via the Frank-Wolfe Algorithm

Christian Bauckhage

B-IT, University of Bonn, Bonn, Germany
Fraunhofer IAIS, Sankt Augustin, Germany
<http://multimedia-pattern-recognition.info>

Abstract. We show that *k*-means clustering is a matrix factorization problem. Seen from this point of view, *k*-means clustering can be computed using alternating least squares techniques and we show how the constrained optimization steps involved in this procedure can be solved efficiently using the Frank-Wolfe algorithm.

1 Introduction

In this paper, we are concerned with theoretical aspects of machine learning. In particular, we revisit *k*-means clustering and investigate it from the point of view of data matrix factorization.

The *k*-means procedure is a popular technique to cluster a data set X of numerical data into subsets C_1, \dots, C_k . The underlying ideas are intuitive and simple and most properties of *k*-means clustering are text book material [1, 2]. Adding to this material, several authors have recently argued that *k*-means clustering can be understood as a constrained matrix factorization problem [3–7]. However, reading the related literature, one cannot but notice that most authors consider this fact self explanatory and appeal to intuition.

Our goals with this paper are therefore as follows: i) we provide a rigorous proof for the equivalence of *k*-means clustering and constrained data matrix factorization. ii) we show that the matrix factorization point of view immediately reveals several properties of *k*-means clustering which are usually difficult to work out. In particular, we show that *k*-means clustering is an integer programming problem and therefore NP-hard, that *k*-means clustering allows for invoking the kernel trick, and that *k*-means clustering is closely related to archetypal analysis [8, 9] and non-negative matrix factorization [10, 11]. iii) we show that the matrix factorization perspective leads to yet another algorithm for computing *k*-means clustering and we discuss how to efficiently implement it using the Frank-Wolfe optimization scheme [12–14].

We begin our presentation with a brief summery of the traditional view on *k*-means clustering and then move on to the matrix factorization perspective. Our discussion assumes that readers are familiar with theory and practice of matrix factorization for data mining and machine learning. Those interested in a gentle introduction into the underlying mathematical ideas are referred to [11].

2 k -Means Clustering: Known Properties and Algorithms

Given a set $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of data points $\mathbf{x}_j \in \mathbb{R}^m$, *hard k -means clustering* attempts to partition the data into k clusters C_1, \dots, C_k such that $C_i \subset X$, $C_i \cap C_l = \emptyset$, and $C_1 \cup C_2 \cup \dots \cup C_k = X$. In particular, hard k -means clustering is a prototype-based clustering technique because it understands clusters to be defined in terms of prototypes or *cluster centroids* $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k \in \mathbb{R}^m$, namely

$$C_i = \left\{ \mathbf{x}_j \in X \mid \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \leq \|\mathbf{x}_j - \boldsymbol{\mu}_l\|^2 \forall l \neq i \right\} \quad (1)$$

The problem at the heart of hard k -means clustering is therefore to search for k appropriate cluster centroids which are typically determined as the minimizers of the following objective function

$$E(k) = \sum_{i=1}^k \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 = \sum_{i=1}^k \sum_{j=1}^n z_{ij} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \quad (2)$$

where the

$$z_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \in C_i \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

are binary indicator variables which indicate whether or not data point \mathbf{x}_j belongs to cluster C_i .

Since hard k -means clustering aims at disjoint clusters where each \mathbf{x}_j is assigned to one and only one C_i , we point out the following important properties of the $z_{ij} \in \{0, 1\}$. If we fix the data index j and sum over the cluster index i , we obtain the number of clusters data point \mathbf{x}_j is assigned to, namely

$$\sum_{i=1}^k z_{ij} = 1. \quad (4)$$

Also, by fixing the cluster index i and summing over the data index j , we find

$$\sum_{j=1}^n z_{ij} = |C_i| = n_i \quad (5)$$

where n_i indicates the number of data points assigned to cluster C_i .

Although the objective in (2) looks rather innocent, it is actually NP-hard [15] and has to be approached using heuristics for which there is no guarantee to find the optimal solution. Indeed, there are various k -means heuristics or algorithms of which well known examples include Lloyd's algorithm (aka "the" k -means algorithm) [16], Hartigan's algorithm [17–19], MacQueen's algorithm [20], or gradient descend methods [21].

3 k -Means Clustering Is Data Matrix Factorization

Having recalled properties of hard k -means clustering in the previous section, our goal is now to rigorously establish that k -means clustering is matrix factorization. In other words, given the objective in (2), we will prove the following identity

$$\sum_{i=1}^k \sum_{j=1}^n z_{ij} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 = \|\mathbf{X} - \mathbf{MZ}\|^2 \quad (6)$$

where

$$\mathbf{X} \in \mathbb{R}^{m \times n} \text{ is a column matrix of } n \text{ data vectors } \mathbf{x}_j \in \mathbb{R}^m \quad (7)$$

$$\mathbf{M} \in \mathbb{R}^{m \times k} \text{ is a column matrix of } k \text{ cluster centroids } \boldsymbol{\mu}_i \in \mathbb{R}^m \quad (8)$$

$\mathbf{Z} \in \mathbb{R}^{k \times n}$ is a matrix of binary indicator variables such that

$$z_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \in C_i \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

3.1 Notation and Preliminaries

Throughout, we write \mathbf{x}_j to denote j -th column vector of a matrix \mathbf{X} . To refer to the (l, j) element of a matrix \mathbf{X} , we either write x_{lj} or $(\mathbf{X})_{lj}$. Moreover, subscripts or summation indices i will be understood to range from 1 to k (the number of clusters), subscripts or summation indices j will range from 1 up to n (the number of data), and subscripts or summation indices l will be used to expand inner products between vectors or rows and columns of matrices.

Finally, regarding the squared Frobenius norm of a matrix, we recall that

$$\|\mathbf{X}\|^2 = \sum_{l,j} x_{lj}^2 = \sum_j \|\mathbf{x}_j\|^2 = \sum_j \mathbf{x}_j^T \mathbf{x}_j = \sum_j (\mathbf{X}^T \mathbf{X})_{jj} = \text{tr}[\mathbf{X}^T \mathbf{X}] \quad (10)$$

3.2 Step by Step Derivation of (6)

To substantiate the claim in (6), we first point out a crucial property of the binary indicator matrix \mathbf{Z} in (9). The property of the $z_{ij} \in \{0, 1\}$ we worked out in (4) translates to the statement that each column of \mathbf{Z} contains a single entry of 1 and $k - 1$ entries of 0. This immediately establishes that the rows of \mathbf{Z} are pairwise perpendicular because

$$z_{ij} z_{i'j} = \begin{cases} 1, & \text{if } i = i' \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

which is then to say that the matrix $\mathbf{Z}\mathbf{Z}^T$ is a diagonal matrix where

$$(\mathbf{Z}\mathbf{Z}^T)_{ii'} = \sum_j (\mathbf{Z})_{ij} (\mathbf{Z}^T)_{ji'} = \sum_j z_{ij} z_{i'j} = \begin{cases} n_i, & \text{if } i = i' \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

Having discussed this property of \mathbf{Z} , we are now positioned to establish the equality in (6) and we will do this in a step by step manner.

Step 1: Expanding the expression on the left of (6) First, we expand the conventional k -means objective function and find

$$\begin{aligned} \sum_{i,j} z_{ij} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 &= \sum_{i,j} z_{ij} (\mathbf{x}_j^T \mathbf{x}_j - 2\mathbf{x}_j^T \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i) \\ &= \underbrace{\sum_{i,j} z_{ij} \mathbf{x}_j^T \mathbf{x}_j}_{T_1} - 2 \underbrace{\sum_{i,j} z_{ij} \mathbf{x}_j^T \boldsymbol{\mu}_i}_{T_2} + \underbrace{\sum_{i,j} z_{ij} \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i}_{T_3}. \end{aligned} \quad (13)$$

This expansion leads to further insights, if we examine the three terms T_1 , T_2 , and T_3 one by one. First of all, we find

$$T_1 = \sum_{i,j} z_{ij} \mathbf{x}_j^T \mathbf{x}_j = \sum_{i,j} z_{ij} \|\mathbf{x}_j\|^2 = \sum_j \|\mathbf{x}_j\|^2 \sum_i z_{ij} = \sum_j \|\mathbf{x}_j\|^2 \quad (14)$$

$$= \text{tr}[\mathbf{X}^T \mathbf{X}] \quad (15)$$

where we made use of (4) and (10). Second of all, we observe

$$T_2 = \sum_{i,j} z_{ij} \mathbf{x}_j^T \boldsymbol{\mu}_i = \sum_{i,j} z_{ij} \sum_l x_{lj} \mu_{li} \quad (16)$$

$$= \sum_{j,l} x_{lj} \sum_i \mu_{li} z_{ij} \quad (17)$$

$$= \sum_{j,l} x_{lj} (\mathbf{M}\mathbf{Z})_{lj} \quad (18)$$

$$= \sum_{j,l} (\mathbf{X}^T)_{jl} (\mathbf{M}\mathbf{Z})_{lj} \quad (19)$$

$$= \sum_j (\mathbf{X}^T \mathbf{M}\mathbf{Z})_{jj} \quad (20)$$

$$= \text{tr}[\mathbf{X}^T \mathbf{M}\mathbf{Z}] \quad (21)$$

Third of all, we note that

$$T_3 = \sum_{i,j} z_{ij} \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i = \sum_{i,j} z_{ij} \|\boldsymbol{\mu}_i\|^2 = \sum_i \|\boldsymbol{\mu}_i\|^2 \sum_j z_{ij} = \sum_i \|\boldsymbol{\mu}_i\|^2 n_i \quad (22)$$

where we applied (5).

Step 2: Expanding the expression on the right of (6) Next, we look at the squared Frobenius norm on the right hand side of (6); it can be written as

$$\begin{aligned} \|\mathbf{X} - \mathbf{M}\mathbf{Z}\|^2 &= \text{tr}[(\mathbf{X} - \mathbf{M}\mathbf{Z})^T (\mathbf{X} - \mathbf{M}\mathbf{Z})] \\ &= \underbrace{\text{tr}[\mathbf{X}^T \mathbf{X}]}_{T_4} - 2 \underbrace{\text{tr}[\mathbf{X}^T \mathbf{M}\mathbf{Z}]}_{T_5} + \underbrace{\text{tr}[\mathbf{Z}^T \mathbf{M}^T \mathbf{M}\mathbf{Z}]}_{T_6} \end{aligned} \quad (23)$$

Given our earlier results, we immediately recognize that $T_1 = T_4$ and $T_2 = T_5$. Thus, to establish that (13) and (23) are indeed equivalent, it remains to verify whether $T_3 = T_6$?

Regarding T_6 , we note that, because of the cyclic permutation invariance of the trace operator, we have

$$\text{tr}[\mathbf{Z}^T \mathbf{M}^T \mathbf{M} \mathbf{Z}] = \text{tr}[\mathbf{M}^T \mathbf{M} \mathbf{Z} \mathbf{Z}^T]. \quad (24)$$

We also note that

$$\text{tr}[\mathbf{M}^T \mathbf{M} \mathbf{Z} \mathbf{Z}^T] = \sum_i (\mathbf{M}^T \mathbf{M} \mathbf{Z} \mathbf{Z}^T)_{ii} \quad (25)$$

$$= \sum_i \sum_l (\mathbf{M}^T \mathbf{M})_{il} (\mathbf{Z} \mathbf{Z}^T)_{li} \quad (26)$$

$$= \sum_i (\mathbf{M}^T \mathbf{M})_{ii} (\mathbf{Z} \mathbf{Z}^T)_{ii} \quad (27)$$

$$= \sum_i \|\boldsymbol{\mu}_i\|^2 n_i \quad (28)$$

where we used the fact that $\mathbf{Z} \mathbf{Z}^T$ is diagonal. This result, however, shows that $T_3 = T_6$ and, consequently, that (13) and (23) are equivalent. That is, we have proven that k -means clustering can indeed be cast as a matrix factorization problem.

4 Consequences

Having proved our central claim in (6), we will next discuss several consequences we can obtain from the matrix factorization formulation of k -means clustering.

4.1 k -Means Clustering Is NP Hard

Given the above result, further insights into the nature of k -means clustering result from eliminating matrix \mathbf{M} from the right hand side of (6). That is, we next ask for the matrix \mathbf{M} that, for a given \mathbf{Z} , would minimize $\|\mathbf{X} - \mathbf{M} \mathbf{Z}\|^2$. To this end, we consider

$$\begin{aligned} \frac{\partial}{\partial \mathbf{M}} \|\mathbf{X} - \mathbf{M} \mathbf{Z}\|^2 &= \frac{\partial}{\partial \mathbf{M}} \left[\text{tr}[\mathbf{X}^T \mathbf{X}] - 2 \text{tr}[\mathbf{X}^T \mathbf{M} \mathbf{Z}] + \text{tr}[\mathbf{Z}^T \mathbf{M}^T \mathbf{M} \mathbf{Z}] \right] \\ &= 2(\mathbf{M} \mathbf{Z} \mathbf{Z}^T - \mathbf{X} \mathbf{Z}^T) \end{aligned} \quad (29)$$

which, upon equating to $\mathbf{0}$, leads to

$$\mathbf{M} = \mathbf{X} \mathbf{Z}^T (\mathbf{Z} \mathbf{Z}^T)^{-1} \quad (30)$$

which beautifully reflects the fact that each of the k -means cluster centroids $\boldsymbol{\mu}_i$ coincides with the mean of the corresponding cluster C_i , namely

$$\boldsymbol{\mu}_i = \frac{\sum_j z_{ij} \mathbf{x}_j}{\sum_j z_{ij}} = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j. \quad (31)$$

Given the result in (30), we therefore find that the k -means objective can be cast solely in terms of the data matrix \mathbf{X} and the indicator matrix \mathbf{Z}

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \left\| \mathbf{X} - \mathbf{X} \mathbf{Z}^T (\mathbf{Z} \mathbf{Z}^T)^{-1} \mathbf{Z} \right\|^2 \\ \text{s.t.} \quad & z_{ij} \in \{0, 1\} \\ & \sum_i z_{ij} = 1 \end{aligned} \quad (32)$$

Looking at (32), we recognize k -means clustering as an integer programming problem, since it corresponds to the discrete optimization problem of finding a column stochastic binary matrix \mathbf{Z} that minimizes the objective in (32). Integer programming problems are NP hard and we can actually read this off (32). \mathbf{Z} is an $k \times n$ binary matrix such that each column contains a single 1; thus, for each column there are k ways to place that 1 and since there are n columns, there are $O(k^n)$ matrices among which we have to determine the optimal one.

4.2 Kernel k -Means Clustering

Observe that the squared Frobenius norm in (32) can also be expanded in terms of trace operators. If we substitute $\boldsymbol{\Xi} = \mathbf{Z}^T (\mathbf{Z} \mathbf{Z}^T)^{-1} \mathbf{Z}$ for brevity, we find

$$\left\| \mathbf{X} - \mathbf{X} \boldsymbol{\Xi} \right\|^2 = \text{tr} \left[\mathbf{X}^T \mathbf{X} \right] - 2 \text{tr} \left[\mathbf{X}^T \mathbf{X} \boldsymbol{\Xi} \right] + \text{tr} \left[\boldsymbol{\Xi} \mathbf{X}^T \mathbf{X} \boldsymbol{\Xi} \right] \quad (33)$$

and recognize that each occurrence of the data vectors \mathbf{x}_i is in form of an inner product, because $(\mathbf{X}^T \mathbf{X})_{ij} = \mathbf{x}_i^T \mathbf{x}_j$.

This, however, shows that k -means allows for invoking the kernel trick as we may replace the $n \times n$ Gramian $\mathbf{X}^T \mathbf{X}$ by an $n \times n$ kernel matrix \mathbf{K} whose entries correspond to kernel evaluations $k(\mathbf{x}_i, \mathbf{x}_j)$.

4.3 k -Means Clustering, Archetypal Analysis and Non-Negative Matrix Factorization

Even further insights into the nature of k -means clustering arise, if we substitute $\mathbf{Y} = \mathbf{Z}^T (\mathbf{Z} \mathbf{Z}^T)^{-1}$ so that we can write $\mathbf{M} = \mathbf{X} \mathbf{Y}$. This again reveals that the centroid vectors $\boldsymbol{\mu}_i$, i.e. the columns of \mathbf{M} , are convex combinations of data points \mathbf{x}_j . That is, $\boldsymbol{\mu}_i = \mathbf{X} \mathbf{y}_i$ where \mathbf{y}_i is an n dimensional vector with n_i entries equal to $1/n_i$ and $n - n_i$ entries equal to 0.

We note that, by definition, \mathbf{Y} is a column stochastic matrix. It is non-negative, i.e. $\mathbf{Y} \succeq \mathbf{0}$, and its columns sum to one, i.e. $\sum_j y_{ji} = \mathbf{1}^T \mathbf{y}_i = 1$. Moreover, each column will have high entropy $H(\mathbf{y}_i) = -\sum_j y_{ji} \log y_{ji} \gg 0$.

We also recall that, as a binary matrix, matrix \mathbf{Z} will obey $\mathbf{Z} \succeq \mathbf{0}$, $\mathbf{1}^T \mathbf{z}_j = 1$, and $H(\mathbf{z}_j) = -\sum_i z_{ij} \log z_{ij} = 0$.

Hard k -Means Clustering Given these prerequisites, we can therefore express hard k -means clustering as a constrained quadratic optimization problem

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{Z}} \quad & \left\| \mathbf{X} - \mathbf{X} \mathbf{Y} \mathbf{Z} \right\|^2 \\ \text{s.t.} \quad & \mathbf{Y} \succeq \mathbf{0}, \mathbf{1}^T \mathbf{y}_i = 1, H(\mathbf{y}_i) \gg 0 \\ & \mathbf{Z} \succeq \mathbf{0}, \mathbf{1}^T \mathbf{z}_j = 1, H(\mathbf{z}_j) = 0. \end{aligned} \tag{34}$$

Archetypal Analysis If we drop the entropy constraints in (34), we obtain

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{Z}} \quad & \left\| \mathbf{X} - \mathbf{X} \mathbf{Y} \mathbf{Z} \right\|^2 \\ \text{s.t.} \quad & \mathbf{Y} \succeq \mathbf{0}, \mathbf{1}^T \mathbf{y}_i = 1 \\ & \mathbf{Z} \succeq \mathbf{0}, \mathbf{1}^T \mathbf{z}_j = 1 \end{aligned} \tag{35}$$

and recover a problem known as *archetypal analysis* (AA) [8, 9]. Dropping the entropy constraints has an interesting effect. Instead of computing basis vectors $\mathbf{M} = \mathbf{X} \mathbf{Y}$ that correspond to local means, AA determines basis vectors that are extreme points of the data in \mathbf{X} . In fact, the *archetypes* in matrix \mathbf{M} reside on the data convex hull [22].

Non-Negative Matrix Factorization If we further drop the sum-to-one constraints in (35), we obtain

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{Z}} \quad & \left\| \mathbf{X} - \mathbf{X} \mathbf{Y} \mathbf{Z} \right\|^2 \\ \text{s.t.} \quad & \mathbf{Y} \succeq \mathbf{0} \\ & \mathbf{Z} \succeq \mathbf{0} \end{aligned} \tag{36}$$

a problem known as *non-negative matrix factorization* (NMF) [10]. Dropping the stochastic constraints has the effect that NMF computes basis vectors $\mathbf{M} = \mathbf{X} \mathbf{Y}$ that are conic combinations of the data in \mathbf{X} .

The expressions in (34), (35), and (36) therefore reveal k -means clustering to be a particularly severely constrained quadratic optimization problem. This is interesting in so far as algorithms for computing k -means clustering conceptually much simpler than AA or NMF algorithms. Nevertheless, it now appears as if methods that have been developed for AA and NMF might also apply to k -means clustering. In the next section, we show that this is indeed the case.

5 Yet Another Algorithm for k -Means Clustering

Since we just found that hard k -means clustering is indeed a constrained form of archetypal analysis, the question is if algorithms that have been developed for archetypal analysis can be used to compute k -means clustering.

In order to see how this can be accomplished, we note that the objective function in (34) is convex in either \mathbf{Y} or \mathbf{Z} but not in their product \mathbf{YZ} . An idea for solving the problem in (34) could therefore be to apply the following constrained alternating least squares procedure

- 1) randomly initialize \mathbf{Y} and \mathbf{Z} under the appropriate constraints
- 2) fix matrix \mathbf{Z} and update \mathbf{Y} to the solution of

$$\begin{aligned} \min_{\mathbf{Y}} \quad & \left\| \mathbf{X} - \mathbf{X}\mathbf{Y}\mathbf{Z} \right\|^2 \\ \text{s.t.} \quad & \mathbf{y}_i \succeq \mathbf{0}, \mathbf{1}^T \mathbf{y}_i = 1, H(\mathbf{y}_i) \gg 0 \end{aligned}$$

- 3) fix matrix \mathbf{Y} and update \mathbf{Z} to the solution of

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \left\| \mathbf{X} - \mathbf{X}\mathbf{Y}\mathbf{Z} \right\|^2 \\ \text{s.t.} \quad & \mathbf{z}_j \succeq \mathbf{0}, \mathbf{1}^T \mathbf{z}_j = 1, H(\mathbf{z}_j) = 0 \end{aligned}$$

- 4) if not converged, continue at 2)

for which we note that the seemingly difficult constrained quadratic optimization problems it involves can actually be solved rather easily.

First of all, we observe that the stochastic constraints $\mathbf{y}_i \succeq \mathbf{0}$ and $\mathbf{1}^T \mathbf{y}_i = 1$ require the columns of matrix \mathbf{Y} to reside in the standard n -simplex

$$\Delta^{n-1} = \{ \mathbf{y} \in \mathbb{R}^n \mid \mathbf{y} \succeq \mathbf{0} \wedge \mathbf{1}^T \mathbf{y} = 1 \}. \quad (37)$$

Second of all, the column entropy $H(\mathbf{y}_i) = -\sum_j y_{ji} \log y_{ji}$ is a concave function so that $-H(\mathbf{y}_i)$ is convex and we are interested in solutions for \mathbf{Y} that maximize $H(\mathbf{y}_i)$. We can thus rewrite the first problem in the above procedure as

$$\begin{aligned} \min_{\mathbf{Y}} \quad & \left\| \mathbf{X} - \mathbf{X}\mathbf{Y}\mathbf{Z} \right\|^2 - \sum_i H(\mathbf{y}_i) \\ \text{s.t.} \quad & \mathbf{y}_i \in \Delta^{n-1}. \end{aligned} \quad (38)$$

Written in this form, we recognize the problem as a convex minimization problem over a compact convex set. This, however, is to say that it can be tackled using the efficient Frank-Wolfe procedure [12]. The Frank-Wolfe algorithm shown in Alg. 1 solves problems of the form

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in S \end{aligned} \quad (40)$$

Algorithm 1 Frank-Wolfe algorithm to solve problem such as in (40)

guess a feasible point \mathbf{x}_0
for $t = 0, \dots, t_{\max}$ **do**
 determine \mathbf{s}_t by solving

$$\min_{\mathbf{s} \in S} \mathbf{s}^T \nabla f(\mathbf{x}_t) \quad (39)$$
 update the learning rate $\gamma_t = \frac{2}{t+2}$
 update the current estimate $\mathbf{x}_{t+1} = \mathbf{x}_t + \gamma_t (\mathbf{s}_t - \mathbf{x}_t)$

where $S \subset \mathbb{R}^m$ is a compact, convex set and $f : S \rightarrow \mathbb{R}$ is a convex function. The key idea is to compute $\mathbf{s} \in S$ that minimizes $\mathbf{s}^T \nabla f(\mathbf{x}_t)$ and to use sub-gradient updates $\mathbf{s} - \mathbf{x}_t$ which guarantee that the updates will never leave the feasible set. The efficiency of the algorithm stems from the fact that it turns a quadratic optimization problem into a series of linear optimization problems and we point out that the minimum of a linear function $\mathbf{s}^T \nabla f(\mathbf{x})$ over a compact convex set will be attained at a vertex of that set.

With respect to our problem of a matrix minimization problem, we note that the gradient ∇f we are concerned with is given by

$$\nabla_{\mathbf{Y}} \left(\left\| \mathbf{X} - \mathbf{X}\mathbf{Y}\mathbf{Z} \right\|^2 - \sum_i H(\mathbf{y}_i) \right) = 2 \left[\mathbf{X}^T \mathbf{X}\mathbf{Y}\mathbf{Z}\mathbf{Z}^T - \mathbf{X}^T \mathbf{X}\mathbf{Z}^T \right] - \mathbf{L} \quad (41)$$

where the components of matrix \mathbf{L} amount to

$$L_{ji} = \frac{\partial}{\partial y_{ji}} \sum_j y_{ji} \log y_{ji} = \log y_{ji} + 1 \quad (42)$$

We can use the columns of this gradient matrix \mathbf{G} to update the columns \mathbf{y}_i of \mathbf{Y} according to the Frank-Wolfe algorithm. We also point out, the compact convex set we are dealing with is the standard simplex Δ^{n-1} whose vertices are given by the standard basis vectors $\mathbf{e}_l \in \mathbb{R}^n$. The problem of finding the minimizer $\mathbf{s} \in \Delta^{n-1}$ thus simplifies to finding the basis vector \mathbf{e}_l that minimizes $\mathbf{e}_l^T \mathbf{g}_i$ which is simply to determine the minimal entry g_{li} in each column \mathbf{g}_i of \mathbf{G} . All in all, these considerations then lead to Alg. 2 for computing updates of matrix \mathbf{Y} .

Similar considerations apply to the problem of computing the updates of the indicator matrix \mathbf{Z} and correspondingly lead to Alg. 3.

To conclude this discussion, we point out that the Frank-Wolfe procedure quickly achieves ϵ -approximations of the optimal solution that are provably sparse [13]. In fact, one can show that after t iterations the current estimate is $O(1/t)$ from the optimal solution [13] which provides a convenient criterion for choosing the number t_{\max} of iterations to be performed. For further details on the Frank-Wolfe algorithms as well as for a recent excellent survey of projection-free convex optimization over compact convex sets, we refer to [14].

Algorithm 2 Frank-Wolfe procedure to compute \mathbf{Y} whose columns are in Δ^{n-1}

Require: data matrix \mathbf{X} , indicator matrix \mathbf{Z} , and parameter $t_{\max} \in \mathbb{N}$
 $\mathbf{Y} \leftarrow [\mathbf{e}_1, \mathbf{e}_1, \dots, \mathbf{e}_1]$ where $\mathbf{e}_1 = [1, 0, \dots, 0]^T \in \mathbb{R}^n$ // initialize $n \times k$ matrix \mathbf{Y}
 $t \leftarrow 0$
repeat
 $\mathbf{G} = 2 [\mathbf{X}^T \mathbf{X} \mathbf{Y} \mathbf{Z} \mathbf{Z}^T - \mathbf{X}^T \mathbf{X} \mathbf{Z}^T] - \mathbf{L}$ // compute gradient matrix
 for $i = 1, \dots, k$ **do** // update columns \mathbf{y}_i of \mathbf{Y}
 $i' = \operatorname{argmin}_l G_{li}$
 $\mathbf{z}_i \leftarrow \mathbf{z}_i + 2/(t+2) \cdot (\mathbf{e}_{i'} - \mathbf{z}_i)$
 $t \leftarrow t + 1$
until updates “become small” or $t = t_{\max}$

Algorithm 3 Frank-Wolfe procedure to compute \mathbf{Z} whose columns are in Δ^{k-1}

Require: data matrix \mathbf{X} , coefficient matrix \mathbf{Y} , and parameter $t_{\max} \in \mathbb{N}$
 $\mathbf{Z} \leftarrow [\mathbf{e}_1, \mathbf{e}_1, \dots, \mathbf{e}_1]$ where $\mathbf{e}_1 = [1, 0, \dots, 0]^T \in \mathbb{R}^k$ // initialize $k \times n$ matrix \mathbf{Z}
 $t \leftarrow 0$
repeat
 $\mathbf{G} = 2 [\mathbf{Y}^T \mathbf{X}^T \mathbf{X} \mathbf{Y} \mathbf{Z} - \mathbf{Y}^T \mathbf{X}^T \mathbf{X}]$ // compute gradient matrix
 for $j = 1, \dots, n$ **do** // update columns \mathbf{z}_j of \mathbf{Z}
 $j' = \operatorname{argmin}_l G_{lj}$
 $\mathbf{z}_j \leftarrow \mathbf{z}_j + 2/(t+2) \cdot (\mathbf{e}_{j'} - \mathbf{z}_j)$
 $t \leftarrow t + 1$
until updates “become small” or $t = t_{\max}$

6 Conclusion

In this paper, we were concerned with machine learning theory. In particular, we were concerned with theoretical aspects of k -means clustering.

First of all, we rigorously established that k -means clustering is a constrained data matrix factorization problem. Second of all, this insight allowed us to easily uncover several properties of k -means clustering that are otherwise more difficult to show. Third of all, the matrix factorization point of view on k -means clustering allowed us to reveal its connections to archetypal analysis and non-negative matrix factorization. Finally, given that k -means clustering can be understood as a constrained version of archetypal analysis, we discussed yet another algorithm for k -means clustering. Archetypal analysis is often computed using alternating least squares optimization and we showed how to adapt this idea to k -means clustering. In particular, we discussed that the seemingly difficult constrained optimization problems involved in this procedure can be solved using the efficient Frank-Wolfe procedure for convex optimization over compact convex sets.

Again, the work reported here is of mainly theoretical interest. What is particularly striking is that it established k -means clustering as a more constrained and thus more difficult problem than archetypal analysis or non-negative matrix factorization. Yet, at the same time, traditional algorithms for k -means clustering are considerably simpler than those for the latter problems. This can be

seen as a call to arms for it suggests that there may be simpler algorithms for these kind of problems as well. Indeed, techniques such as k -maxoids clustering [23] which were derived from k -means clustering indicate that, say, archetypal analysis should be solvable by simple algorithms, too.

References

1. MacKay, D.: Information Theory, Inference, & Learning Algorithms. Cambridge University Press (2003)
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2001)
3. Ding, C., He, X., Simon, H.: On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In: Proc. SDM, SIAM (2005)
4. Gaussier, E., Goutte, C.: Relations between PLSA and NMF and Implications. In: Proc. SIGIR, ACM (2005)
5. Kim, J., Park, H.: Sparse Nonnegative Matrix Factorization for Clustering. Technical Report GT-CSE-08-01, Georgia Institute of Technology (2008)
6. Arora, R., Gupta, M., Kapila, A., Fazel, M.: Similarity-based Clustering by Left-Stochastic Matrix Factorization. J. of Machine Learning Research **14**(Jul.) (2013)
7. Bauckhage, C., Drachen, A., Sifa, R.: Clustering Game Behavior Data. IEEE Trans. on Computational Intelligence and AI in Games **7**(3) (2015)
8. Cutler, A., Breiman, L.: Archetypal Analysis. Technometrics **36**(4) (1994)
9. Bauckhage, C., Thureau, C.: Making Archetypal Analysis Practical. In: Pattern Recognition. Volume 5748 of LNCS., Springer (2009)
10. Cichocki, A., Zdunek, R., Phan, A., Amari, S.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. Wiley (2009)
11. Bauckhage, C.: A Purely Geometric Approach to Non-Negative Matrix Factorization. In: Proc. KDML-LWA. (2014)
12. Frank, M., Wolfe, P.: An Algorithm for Quadratic Programming. Naval Research Logistics Quarterly **3**(1-2) (1956)
13. Clarkson, K.: Coresets, Sparse Greedy Approximation, and the Frank-Wolfe Algorithm. ACM Trans. on Algorithms **6**(4) (2010)
14. Jaggi, M.: Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. J. of Machine Learning Research **28**(1) (2013)
15. Aloise, D., Deshapande, A., Hansen, P., Popat, P.: NP-Hardness of Euclidean Sum-of-Squares Clustering. Machine Learning **75**(2) (2009)
16. Lloyd, S.: Least Squares Quantization in PCM. IEEE Trans. on Information Theory **28**(2) (1982)
17. Hartigan, J., Wong, M.: Algorithm AS 136: A k -Means Clustering Algorithm. J. of the Royal Statistical Society C **28**(1) (1979)
18. Slonim, N., Aharoni, E., Cramer, K.: Hartigan's k -Means Versus Lloyd's k -Means – Is It Time for a Change? In: Proc. IJCAI. (2013)
19. Telgarsky, M., Vattani, A.: Hartigan's Method: k -means Clustering without Voronoi. In: Proc. AISTATS. (2010)
20. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In: Proc. Berkeley Symp. on Mathematical Statistics and Probability. (1967)

21. Bottou, L., Bengio, Y.: Convergence Properties of the K-Means Algorithms. In: Proc. NIPS. (1995)
22. Bauckhage, C.: A Note on Archetypal Analysis and the Approximation of Convex Hulls. arXiv:1410.0642 [cs.NA] (2014)
23. Bauckhage, C., Sifa, R.: k-Maxoids Clustering. In: Proc. KDML-LWA. (2015)

Plackett-Luce Networks for Dyad Ranking

Dirk Schäfer¹ and Eyke Hüllermeier²

¹ University of Marburg, Germany
dirk.schaefer@uni-marburg.de

² Department of Computer Science
Paderborn University, Germany
eyke@upb.de

Abstract. We propose a new method for dyad ranking, a problem that was recently introduced in the realm of preference learning. Our method, called PLNet, combines a statistical model for rank data, namely the Plackett-Luce model, with neural networks (feed-forward multi-layer perceptrons) in order to learn joint-feature representations for dyads, which are pairs of objects from two domains. The efficacy of PLNet is shown by comparing it experimentally with state-of-the-art methods for dyad and label ranking.

Keywords: Preference learning, dyad ranking, Plackett-Luce model, multi-layer perceptrons.

1 Introduction

A specific problem in the realm of preference learning [7] is the problem of *label ranking*, which consists of learning a model that maps instances to rankings (total orders) over a finite set of predefined alternatives [20]. An instance, which defines the context of the preference relation, is typically characterized in terms of a set of attributes or features; for example, an instance could be a person described by properties such as sex, age, income, etc. As opposed to this, the alternatives to be ranked, e.g., the political parties of a country, are only identified by their name (label), while not being characterized in terms of any properties or features.

In [17], we introduced *dyad ranking* as a practically motivated generalization of the label ranking problem. In dyad ranking, not only the instances but also the alternatives are represented in terms of attributes—a dyad is a pair consisting of an instance and an alternative. Moreover, for learning in the setting of dyad ranking, we proposed an extension of an existing label ranking method based on the Plackett-Luce model, a statistical model for rank data. This approach is based on modeling latent utility scores of dyads in the form of a Kronecker product of the feature representations of the instance and alternative, respectively, which is why we speak of a *bilinear* Plackett-Luce model.

In this paper, we propose a variant of this approach, called *PLNet*, that allows for modeling latent utilities in a more flexible way. Instead of assuming a

bilinear structure of the joint-feature representation, we model utilities as (feed-forward) neural networks; thanks to the hidden layer of such networks, important non-linear dependencies can thus be captured [16].

The rest of the paper is organized as follows. We provide a formal description of the dyad ranking problem in Section 2 and an overview of related methods in Section 3. Our new method PLNet is described in Section 4. Experimental results are presented in Section 5, prior to concluding in Section 6.

2 Problem Setting

Formally, a dyad is a pair of feature vectors $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathbb{Z} = \mathbb{X} \times \mathbb{Y}$, where the feature vectors are from two (not necessarily different) domains \mathbb{X} and \mathbb{Y} . A single training observation ρ_n ($1 \leq n \leq N$) takes the form of a dyad ranking

$$\rho_n : \mathbf{z}^{(1)} \succ \mathbf{z}^{(2)} \succ \dots \succ \mathbf{z}^{(M_n)}, \quad M_n \geq 2, \quad (1)$$

the length M_n of which can vary between observations in the data set $\mathcal{D} = \{\rho_n\}_{n=1}^N$. An equivalent notation for a single training example that will be used later on is a set of dyads

$$\varrho_n = \left\{ \mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(M_n)} \right\}$$

together with a permutation $\pi_n : \{1, \dots, M_n\} \rightarrow \{1, \dots, M_n\}$ indicating how these dyads are ranked.

The task of a dyad ranking method is to learn a ranking function that accepts as input any set of (new) dyads and produces as output a ranking of these dyads.

An important special case, called *contextual dyad ranking*, is closely related to label ranking [17]. As already mentioned, the label ranking problem is about learning a model that maps instances to rankings over a finite set of predefined choice alternatives $\mathcal{Y} = \{y_1, \dots, y_K\}$. In terms of dyad ranking this means that all dyads in an observation share the same context \mathbf{x} , i.e., they are all of the form $\mathbf{z}^{(j)} = (\mathbf{x}, \mathbf{y}^{(j)})$; in this case, (1) can also be written as

$$\rho_n : \left(\mathbf{x}, \mathbf{y}^{(1)} \right) \succ \left(\mathbf{x}, \mathbf{y}^{(2)} \right) \succ \dots \succ \left(\mathbf{x}, \mathbf{y}^{(M_n)} \right) . \quad (2)$$

Likewise, a prediction problem will typically consist of ranking a subset

$$\left\{ \mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(M)} \right\} \subseteq \mathbb{Y}$$

in a given context \mathbf{x} . Contextual dyad ranking generalizes label ranking by considering feature vectors instead of labels. This includes vector space embeddings of labels and additional descriptions (side-information) about labels. Contextual dyadic preferences are equivalent to training triplets of the form $\mathbf{y}^{(1)} \succ_{\mathbf{x}} \mathbf{y}^{(2)}$, which are encountered frequently in applications such as similarity learning.

3 Related Work

The Bilinear Plackett-Luce model (BilinPL) for dyad ranking was introduced in [17]. It builds on a statistical ranking model introduced by Plackett and Luce [14] and represents the parameters of this model as a log-bilinear function $f(\mathbf{x}, \mathbf{y}) = \exp g(\mathbf{x}, \mathbf{y})$, which takes as input the dyad member feature vectors \mathbf{x} and \mathbf{y} . The output of the function is a real positive value, which can be interpreted as a utility score. The bilinear function $g(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{W} \mathbf{y}$ can equivalently be expressed as the function $g(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \mathbf{x} \otimes \mathbf{y} \rangle$, i.e., the dot product between a weight vector and a joint-feature vector consisting of cross-products (also known as the Kronecker product). This kind of joint-feature vector formulation strongly depends on the dyad feature vectors, and the representation bias it imposes on the model can be suboptimal.

Comparison training refers to a framework introduced for learning from pairwise preferences with a neural network [19]. The network architecture consists of two subnetworks which are connected to a single output node and indicates which of the two inputs is preferred over the other. The weights associated with the last hidden layer of one subnetwork is the mirrored version of the other subnetwork's weights. This setup solves two principal problems, namely efficiency and consistency. In the evaluation phase it is sufficient to use just one subnetwork and evaluate n alternatives instead of $\mathcal{O}(n^2)$ pairs. Furthermore, this kind of network architecture also enforces transitivity of the predicted preferences. Although many modifications of the original comparison training have been proposed in the past (see e.g. the survey [6]), its inputs however are essentially restricted to pairwise training signals.

The aforementioned label ranking problem has been tackled with neural network approaches previously [15, 11]. A multi-layer perceptron (MLP) has been utilized to produce recommendations on meta-heuristics (labels) on different meta-features (instances) within the realm of meta-learning [11]. This kind of neural network exhibits an output layer with as many nodes as there are labels. The error signal used to modify the network's weights is formed by using the mean squared error on the target rank positions of the labels. In [15] more effort has been spent to incorporate label ranking loss information into the back-propagation procedure. To this end some variations for this procedure have been investigated. Both architectures are similar to each other and have two essential limitations: first, they depend on a fixed number of labels, and second, they cannot cope with incomplete ranking observations. In addition, they lack the ability to provide probabilistic information on their predictions.

In the domain of information retrieval, the neural network-based approaches RankNet and ListNet have a probabilistic foundation [2, 3]. RankNet [2] uses pairwise inputs to learn a utility scoring function with the cross entropy loss. To this end, the training data consists of sample pairs together with target probabilities. These quantify the probability to rank the first sample higher than the second. With the introduction of target probabilities, this approach enables the possibility of modeling ties between samples. ListNet [3, 13] similarly uses the cross entropy as a metric, but in contrast to RankNet it processes lists of

samples instead of pairwise preferences as basic observation. There are, however, some important differences between ListNet and our approach:

- The learning approach in ListNet addresses only a special case of the Plackett-Luce distribution, namely the case of Top-k data with $k = 1$.
- In ListNet, a *linear* neural network is used. This is in contrast to our approach, in which non-linear relationships between inputs and outputs are learned. Linearity in the ListNet approach implies that much emphasize must be put on engineering joint feature input vectors.
- In ListNet, the query-document features are associated with absolute scores (relevance degrees) as training information, i.e., quantitative data, whereas PLNet deals with rankings, i.e., data of qualitative nature.³

ListNet as well as RankNet expect single instance joint-feature vectors $\mathbf{x} = \Psi(q, d)$ as inputs that are created from query-document pairs (q, d) via feature-engineering. These approaches are closely related to our method, though with the major difference of ranked dyads instead of absolute ratings as training input; and perhaps more importantly, our goal is to *learn* joint-feature representations instead of engineering them.

4 Plackett-Luce Network

Throughout this section we use the following notation. A ranking of dyads is represented in terms of a permutation π . More specifically, given a numbering of dyads from 1 to M , let $\pi(i)$ be the number j of the dyad put on the i -th position in the ranking; the inverse $\pi^{-1}(j)$ denotes the rank position of the dyad specified by index j .

4.1 Plackett-Luce Model

The Plackett-Luce (PL) model is a parameterized probability distribution on the set of all rankings over a set of alternatives y_1, \dots, y_K . It is specified by a parameter vector $\mathbf{v} = (v_1, v_2, \dots, v_K) \in \mathbb{R}_+^K$, in which v_i accounts for the (latent) utility or “skill” of the option y_i . The probability assigned by the PL model to a ranking with a permutation π is given by

$$\mathbf{P}(\pi | \mathbf{v}) = \prod_{i=1}^K \frac{v_{\pi(i)}}{v_{\pi(i)} + v_{\pi(i+1)} + \dots + v_{\pi(K)}} = \prod_{i=1}^{K-1} \frac{v_{\pi(i)}}{\sum_{j=i}^K v_{\pi(j)}} \quad (3)$$

Obviously, the Plackett-Luce model is only determined up to a positive multiplicative constant, i.e., $\mathbf{P}(\pi | \mathbf{v}) \equiv \mathbf{P}(\pi | s \cdot \mathbf{v})$ for all $s > 0$.

As an appealing property of the PL model, we also note that its marginal probabilities (probabilities of rankings on subsets of alternatives) are easy to compute and can be expressed in closed form. More specifically, the marginal of a PL model on $M < K$ alternatives $y_{i(1)}, \dots, y_{i(M)}$ is again a PL model with parameters $v_{i(1)}, \dots, v_{i(M)}$.

³ The use of query-document-associated scores as PL model parameters is arguable.

4.2 Architecture

The core idea of the Plackett-Luce Network (PLNet) is to learn (latent) utility functions $u = g(\mathbf{x}, \mathbf{y})$, where g is a single multi-layer feed-forward neural network. Since we are interested in utility scores that reflect rankings, we utilize the Plackett-Luce model and express its real-valued skill parameters as functions of the form $\log(v) = u = g(\mathbf{x}, \mathbf{y})$. Thus, probabilities of rankings are given by

$$P(\pi | \varrho, \mathbf{u}) = \prod_{k=1}^{K-1} \frac{\exp(u_{\pi(k)})}{\sum_{l=k}^K \exp(u_{\pi(l)})} . \quad (4)$$

The PLNet is a multi-layer perceptron (MLP), which is constructed as shown in Figure 1. It consists of multiple layers and takes as input two vectors corresponding to the members of a dyad. There is at least one hidden layer with nodes using a sigmoidal activation function to learn non-linear mappings. The output layer produces a scalar value of the form $u = \langle \mathbf{w}^L, \mathbf{a}^{L-1} \rangle + b^L$. Technically, a bias term b^L is actually not needed, as it has no effect on the PLNet model (4): With the choice of the exponential function (to ensure positivity of the parameters v) and the independence of b^L of the inputs \mathbf{x} and \mathbf{y} , we have

$$v = \exp(\langle \mathbf{w}^L, \mathbf{a}^{L-1} \rangle + b^L) = b^L \cdot \exp(\langle \mathbf{w}^L, \mathbf{a}^{L-1} \rangle) ,$$

and as noted before, the PL model is invariant toward multiplication of the weights with a positive scalar.

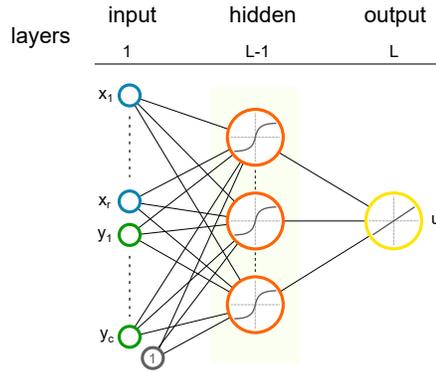


Fig. 1. PLNet architecture. This kind of feed-forward neural network is composed of several layers. Dyad members $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ are entered at the input layer, whose nodes are fully connected with the nodes of the next layer. Inner (hidden) layers have nodes endowed with a non-linear activation function. The output layer consists of a single node with linear activation function.

4.3 Training

The training procedure uses *back-propagation*, which is a gradient technique to find optimal weights [16]. Following [1], a feed-forward network can be seen as a framework for modeling the conditional probability distribution. For a set of training data $\{\varrho_n, \pi_n\}$ with

$$\varrho_n = \left\{ (\mathbf{x}_n^{(1)}, \mathbf{y}_n^{(1)}), \dots, (\mathbf{x}_n^{(M_n)}, \mathbf{y}_n^{(M_n)}) \right\} ,$$

the likelihood can be written as

$$\mathcal{L} = \prod_n \mathbf{P}(\varrho_n, \pi_n) = \prod_n \mathbf{P}(\pi_n | \varrho_n) \mathbf{P}(\varrho_n)$$

if we assume the observations $\{\varrho_n, \pi_n\}$ to be independent and identically distributed. It is generally more convenient to minimize the logarithm of the likelihood, hence we aim to minimize the following error function:

$$E = -\ln \mathcal{L} = -\sum_n \ln \mathbf{P}(\pi_n | \varrho_n) - \sum_n \mathbf{P}(\varrho_n) \quad (5)$$

The second term in (5) can be dropped, as it does not depend on the network parameters (and only represents an additive constant). More specifically, we denote the negative log-likelihood (NLL) by $E = \sum_n E_n$, where

$$E_n = -\log P(\pi_n | \varrho_n, \mathbf{u}) = \sum_{k=1}^{M_n-1} \log \sum_{l=k}^{M_n} \exp(u_{\pi(l)}) - \sum_{k=1}^{M_n-1} u_{\pi(k)} . \quad (6)$$

The errors propagated back to the network can thus be expressed as the (partial) derivatives of the NLL:

$$\delta_i^L \equiv \frac{\partial E_n}{\partial u_i} = \sum_{k=1}^{M_n-1} \frac{\mathbb{1}_{\{\pi^{-1}(i) \geq k\}} \exp(u_i)}{\sum_{l=k}^{M_n} \exp(u_{\pi(l)})} - \mathbb{1}_{\{i \leq M_n-1\}} \quad (7)$$

Note that the calculation of the error in (7) depends on all the M_n utilities of a training sample. To this end, we propose a training procedure as depicted in Figure 2, where we run M_n feed-forward calculations for a sample (consisting of M_n dyads) in parallel to obtain utility values. The δ -values obtained according to (7) can then be used for the standard back-propagation procedure on each of the parallel M_n copies of the master network. After performing the weight updates individually, the master's weights can be updated by aggregating the individual weight changes. In the implementation we applied the following simple strategy: $w_{ji} = w_{ji} - \eta \cdot \Delta w_{ji}^k$, $1 \leq k \leq M_n$, where η denotes the learning rate. The procedure is then repeated several times with the training samples and stopped when the error has diminished sufficiently.

Another point to be mentioned is regularization. Here, we suggest to use *early stopping* by tracking the NLL values of the training and validation data during the learning process. A good point to stop the training and to prevent over-fitting is when the validation NLL values begin to rise again.

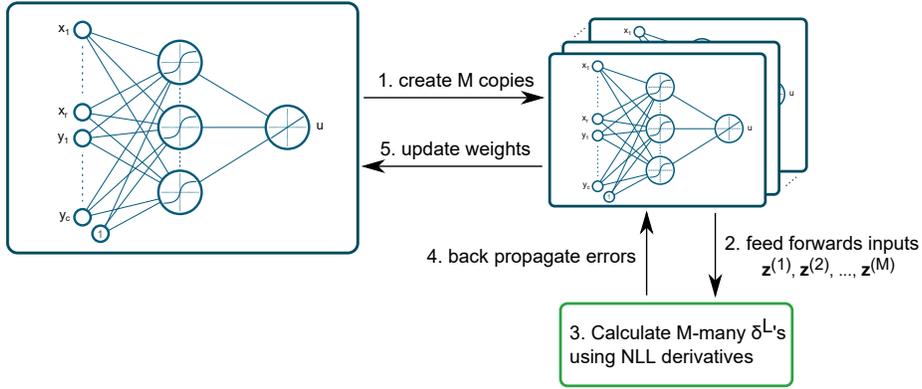


Fig. 2. PLNet training steps. An observation consists of M ranked dyads, which are fed into M copies of the current PLNet model. The output of this composite feed-forward procedure (step 2) is used to calculate the ranking cost (step 3). Finally, the weights are adjusted via back-propagation (step 4), and the changes in the weights are merged into a single network (step 5).

4.4 PLNet for Label Ranking

In order to apply PLNet on the problem of label ranking, one needs a vector representation for labels. Unless other label features are available, a natural approach is to use a 1-of- K encoding. Thus, given K labels, the i -th label is represented by the vector $e_i \in \{0, 1\}^K$ with entry 1 at position i and 0 otherwise.

5 Experiments

In the following experiments, the predictive performance is measured in terms of the Kendall's tau coefficient [12], a rank correlation measure commonly used in the statistical and preference learning literature [20, 21]. It is defined as

$$\tau = \frac{C(\pi, \hat{\pi}) - D(\pi, \hat{\pi})}{K(K-1)/2}, \quad (8)$$

with C and D the number of concordant (put in the same order) and discordant (put in the reverse order) label pairs, respectively, and K the length of the rankings π and $\hat{\pi}$ (number of alternatives). Kendall's tau assumes values in $[-1, +1]$, with $\tau = +1$ for the perfect prediction $\hat{\pi} = \pi$ and $\tau = -1$ if $\hat{\pi}$ is the exact reversal of π .

5.1 Synthetic Data

Here, we compare the dyad ranking methods PLNet and BilinPL. To this end, we sample from a PLNet with three layers. Its inputs are dyads composed of

one-dimensional vectors, i.e., $\mathbf{x}, \mathbf{y} \in [-1, 1]$. The hidden layer has 25 nodes and all weights are initialized at random by sampling from the uniform distribution in $[-15, 15]$. From a total of 400 dyads, we sample 500 training and 50 test rankings consisting of 5 dyads each.

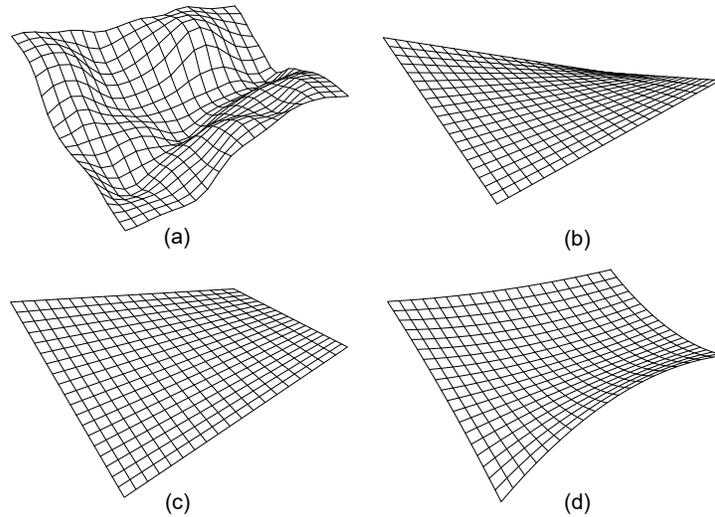


Fig. 3. Log-skill landscapes of the methods PLNet and BilinPL. (a) is produced by PLNet after learning from object pair rankings. (b)-(d) are produced by BilinPL using different feature specifications.

For the BilinPL, we choose from three different variants of input features. They are based on the Kronecker product between object pair features to form joint-feature vectors (resulting in first and second order models):

\mathbf{x}_f	\mathbf{y}_f	$\Rightarrow \mathbf{z}_f$	Identifier
x	y	$[x \cdot y]$	BilinPL-1
$[x, 1]$	$[y, 1]$	$[x, y, xy, 1]$	BilinPL-2
$[x, x^2, 1]$	$[y, y^2, 1]$	$[x, x^2, y, y^2, xy, xy^2, x^2y, x^2y^2, 1]$	BilinPL-3

Table 1 and the corresponding Figure 3 underpin two key aspects. Firstly, the expressiveness of PLNet can be much larger compared to the BilinPL versions. Secondly, the predictive quality varies strongly for BilinPL depending on the choice of the dyad features.

Table 1. Results of the synthetic pair ranking data experiment.

	PLNet	BilinPL-1	BilinPL-2	BilinPL-3
Kendall’s τ	0.944	0.380	0.644	0.852
Figure 3	(a)	(b)	(c)	(d)

5.2 UCI Label Ranking Datasets

A suite of benchmark data sets have been established for the label ranking setting [4].⁴ They are based on 16 well known multi-class and regression data sets from the UCI repository and processed in two ways to address the ranking problem (see Table 2 for their properties). For data sets of type A (multi-class problems), rankings were generated by training a naive Bayes classifier on the complete data set and ordering the class labels according to the predicted class probabilities for each example. For data sets of type B (regression problems), a subset of instance attributes are removed from the data sets and were interpreted as labels. Rankings were then obtained by standardizing the attributes and then ordering them by size. This approach is justified by assuming that the original attributes are correlated and the remaining features contain information about the rankings of the removed attributes.

We compare the performance of PLNet to other state-of-the-art label ranking methods on these data sets using 10-fold cross-validation. For PLNet, we use three layers with 10 neurons for the hidden layer. In addition to BilinPL, we include Ranking by Pairwise Comparison (RPC, [10]), Constrained Classification (CC, [8, 9]), and the log-linear model for label ranking (LL, [5]) as additional baselines.⁵ For BilinPL, we chose as dyad features $x_f = [x, 1]$ and $y_f^{(i)} = e_i$, which by means of the Kronecker product results in a joint-feature vector representation called multi-vector [18].

The results (see Table 3) indicate that PLNet is competitive to the other approaches and most of the time even superior. The performances on data sets where PLNet is less powerful also show its weakness. For data sets consisting of only a few instances but many attributes, PLNet is likely to over-fit. Of course, another issue is the choice of the architecture itself (how many layers, how many nodes, ...). Linear models are advantageous in comparison to PLNet if their inductive bias is correct, which is the case, for example, for the *fried* problem.

6 Conclusion

We introduced a new method for the problem of dyad ranking, called PLNet. The method exhibits some interesting properties, notably the following: it is

⁴ Available online at <https://www-old.cs.uni-paderborn.de/fachgebiete/intelligente-systeme/software/label-ranking-datasets.html>.

⁵ CC was used in its online variant as described in [10].

Table 2. Semi-synthetic label ranking data sets and their properties.

Type A				Type B			
data set	# inst.(N)	# attr.(d)	# labels(M)	data set	# inst.(N)	# attr.(d)	# labels(M)
authorship	841	70	4	bodyfat	252	7	7
glass	214	9	6	calhousing	20640	4	4
iris	150	4	3	cpu-small	8192	6	5
pendigits	10992	16	10	elevators	16599	9	9
segment	2310	18	7	fried	40769	9	5
vehicle	846	18	4	housing	506	6	6
vowel	528	10	11	stock	950	5	5
wine	178	13	3	wisconsin	194	16	16

Table 3. Results on the UCI label ranking data sets.

data set	BilinPL	CC	LL	PLNet	RPC-LR
authorship	0.931±0.013	0.916±0.015	0.935±0.013	0.908±0.025	0.917±0.020
bodyfat	0.268±0.059	0.245±0.052	0.287±0.060	0.251±0.040	0.285±0.061
calhousing	0.220±0.011	0.254±0.009	0.235±0.010	0.272±0.014	0.243±0.010
cpu-small	0.445±0.016	0.468±0.017	0.431±0.017	0.500±0.019	0.449±0.016
elevators	0.730±0.007	0.770±0.009	0.725±0.006	0.788±0.009	0.749±0.008
fried	0.999±0.000	0.999±0.000	0.997±0.001	0.951±0.010	1.000±0.000
glass	0.835±0.072	0.830±0.079	0.825±0.074	0.846±0.080	0.889±0.057
housing	0.655±0.040	0.639±0.044	0.646±0.034	0.703±0.033	0.672±0.041
iris	0.813±0.112	0.800±0.109	0.804±0.101	0.960±0.049	0.911±0.047
pendigits	0.892±0.003	0.896±0.002	0.879±0.002	0.905±0.005	0.932±0.002
segment	0.903±0.008	0.910±0.008	0.879±0.009	0.939±0.008	0.929±0.009
stock	0.704±0.016	0.714±0.016	0.702±0.018	0.882±0.020	0.774±0.024
vehicle	0.855±0.020	0.850±0.025	0.790±0.023	0.872±0.025	0.855±0.015
vowel	0.581±0.026	0.577±0.046	0.608±0.023	0.805±0.016	0.644±0.021
wine	0.929±0.052	0.914±0.069	0.954±0.041	0.942±0.034	0.925±0.054
wisconsin	0.629±0.028	0.612±0.030	0.605±0.027	0.514±0.028	0.632±0.027
average rank	3.250	3.625	3.812	2.125	2.188

probabilistic in nature, can handle incomplete rankings, and builds on standard neural network components. Thus, it is possible to improve the method further with techniques developed in the neural networks literature during the last years,

especially with recent advances in deep learning. This point is an important aspect of future research. The study of other neural network architectures such as (restricted) Boltzman machines for dyad ranking are also conceivable.

References

1. Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA, 1995.
2. Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings ICML, 22nd International Conference on Machine Learning*, pages 89–96, New York, NY, USA, 2005. ACM.
3. Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings ICML, 24th International Conference on Machine Learning*, pages 129–136, New York, NY, USA, 2007. ACM.
4. Weiwei Cheng, Jens Hühn, and Eyke Hüllermeier. Decision tree and instance-based learning for label ranking. In *Proceedings ICML, 26th International Conference on Machine Learning*, pages 161–168, Montreal, Canada, June 2009. Omnipress.
5. Ofer Dekel, Yoram Singer, and Christopher D Manning. Log-linear models for label ranking. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16 (NIPS-2003)*, pages 497–504. MIT Press, 2003.
6. Johannes Fürnkranz. Machine learning in games: A survey. In Johannes Fürnkranz and M. Kubat, editors, *Machines that Learn to Play Games*, chapter 2, pages 11–59. Nova Science Publishers, Huntington, NY, 2001.
7. Johannes Fürnkranz and Eyke Hüllermeier. *Preference Learning*. Springer, New York, NY, USA, 2011.
8. Sarel Har-Peled, Dan Roth, and Dav Zimak. Constraint classification: A new approach to multiclass classification. In *Proceedings ALT, 13th International Conference on Algorithmic Learning Theory*.
9. Sarel Har-Peled, Dan Roth, and Dav Zimak. Constraint classification for multiclass classification and ranking. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 809–816. MIT Press, 2002.
10. Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16):1897–1916, 2008.
11. Jorge Kanda, Carlos Soares, Eduardo R. Hruschka, and André Carlos Ponce Leon Ferreira de Carvalho. A meta-learning approach to select meta-heuristics for the traveling salesman problem using MLP-based label ranking. In *Proceedings ICONIP, 19th International Conference on Neural Information Processing*, pages 488–495, Doha, Qatar, 2012. Springer.
12. M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
13. Tianyi Luo, Dong Wang, Rong Liu, and Yiqiao Pan. Stochastic top-k listnet. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 676–684, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
14. John I. Marden. *Analyzing and Modeling Rank Data*. Chapman & Hall, 1995.

15. Geraldina Ribeiro, Wouter Duivesteijn, Carlos Soares, and Arno J. Knobbe. Multilayer perceptron for label ranking. In *Proceedings ICANN, 22nd International Conference on Artificial Neural Networks*, pages 25–32, Lausanne, Switzerland, 2012. Springer.
16. David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:9, 1986.
17. Dirk Schäfer and Eyke Hüllermeier. Dyad ranking using a bilinear Plackett-Luce model. In *Proceedings ECML/PKDD-2015, European Conference on Machine Learning and Knowledge Discovery in Databases*, Porto, Portugal. Springer.
18. Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.
19. Gerald Tesauro. Connectionist learning of expert preferences by comparison training. In D. Touretzky, editor, *Advances in Neural Information Processing Systems 1 (NIPS-1988)*, pages 99–106. Morgan Kaufmann, 1989.
20. Shankar Vembu and Thomas Gärtner. Label ranking algorithms: A survey. In J. Fürnkranz and E. Hüllermeier, editors, *Preference Learning*, pages 45–64. Springer-Verlag, 2011.
21. Yangming Zhou, Yangguang Liu, Jiangang Yang, Xiaoqi He, and Liangliang Liu. A taxonomy of label ranking algorithms. *Journal of Computers*, 9(3):557–565, 2014.

The Partial Weighted Set Cover Problem with Applications to Outlier Detection and Clustering

Sebastian Bothe¹ and Tamás Horváth^{2,1}

¹Fraunhofer IAIS, Schloss Birlinghoven, 53754 St. Augustin, Germany

²Dept. of Computer Science, University of Bonn, Germany

{firstname.lastname}@iais.fraunhofer.de

Abstract. We define the partial weighted set cover problem, a generic combinatorial optimization problem, that includes some classical data mining problems as special cases. We prove that it is computationally intractable and give a local search algorithm for this problem. As application examples, we then show how to translate clustering and outlier detection problems into this generic problem. Our experiments on synthetic and real-world datasets indicate that the quality of the solution produced by the generic local search algorithm is comparable to that obtained by state-of-the-art clustering and outlier detection algorithms.

1 Introduction

Let \mathcal{S} be a set system over a finite ground set such that for all $X \in \mathcal{S}$ and for all $x \in X$, x is associated with a real-valued *relative weight* with respect to X . That is, x can have as many different relative weights as many sets it belongs to. In addition to the relative weights, \mathcal{S} is equipped with two real-valued set functions measuring the *weight* and the *generality* of the elements of \mathcal{S} . The weight of a set in \mathcal{S} is defined as the sum of the relative weights of its elements. In this paper we consider the following *partial weighted set cover problem*: Given a weighted set system \mathcal{S} over some finite ground set as described above, a positive integer k , and a generality function on \mathcal{S} , find k elements of \mathcal{S} maximizing a utility function composed of a reward and two penalty terms. While the reward term gives preference to sets with the highest weights, the penalty terms discourage the selection of overlapping sets, as well as sets with high generality. Intuitively, our aim is to select k sets that cover as many elements of the ground set as possible, subject to the constraints that the sets have as small pairwise overlap as possible and are as specific as possible.

We show that there is a polynomial reduction from the decision version of the classical set cover problem, implying that the partial set cover problem is NP-hard. To overcome the computational limitation of finding the optimal solution, we resort to a generic local search algorithm. In each iteration of the algorithm, the current k sets are updated by applying the two steps below:

- (i) In the first step we proceed as follows: For each of the k sets, we select one of its supersets from \mathcal{S} that, together with the remaining $k - 1$ sets, maximizes

- the utility function and fulfills the following conditions: The new utility value is greater than the old one and the new candidate set does not cover any new element already contained by any of the other $k - 1$ sets. If all these conditions are satisfied, we replace the set by this maximal superset selected.
- (ii) In the second step we then select $O(\log k)$ sets from the k sets obtained after step (i) uniformly and independently at random. For each set selected, we replace it either with one of its direct subsets (i.e., maximal proper subsets) or direct supersets (i.e., minimal proper supersets) in \mathcal{S} selected uniformly at random. If the new k sets obtained in this way have a better utility or they have the same utility and the outcome of a biased coin flip is head, we keep the new k sets; otherwise we use the ones obtained after step (i).

Regarding (i), note that by construction, if at least one of the k sets will be changed after this step, we have a strict increase in the utility. Furthermore, this step is greedy, as the k sets are processed separately. Regarding (ii), this step may not result in strict increase in the utility with a certain probability specified by the user.

In the second part of the paper, we consider set systems \mathcal{S} over some finite set $S \subseteq \mathbb{R}^d$. More precisely, a subset of S belongs to \mathcal{S} if and only if it can be realized by a d -dimensional ball of radius r around a point $p \in S$, where r is the element of a finite geometric progression for some scale factor specified by the user. The *relative weights* of the elements in a ball are defined by a function monotonically decreasing in their distances to the center. If a set can be realized by more than one ball, we take the ball with the highest weight. Using these weights, the *reward* term for a family of k sets in \mathcal{S} is defined as the sum of their weights. Defining the generality of a set by the radii of the ball representing it, the *penalty* terms are given by the sum of the radii of the k balls and by the sum of all relative weights of the elements except for the highest one.

Using the set system with the utility function as well as the generic algorithm sketched above, we arrive at a combinatorial optimization problem and at an algorithm solving it that are highly relevant e.g. for (soft) clustering and outlier detection problems. In fact, as we experimentally demonstrate on synthetic and real-world data, our empirical results on these two particular data mining problems are comparable to those obtained by state-of-the-art algorithms. This is especially remarkable because, as we will discuss in detail, the balls inducing the set system are split into concentric annuli and two points belonging to the same ball have the same relative weight if and only if they belong to the same annulus. According to our experimental results, this type of *distance discretization* does not seem to have a strong impact on the quality of the output.

One of the main advantages of our approach is its *generality*: After the domain *dependent* step of specifying the set system and the utility function for a broad class of problems, we can solve the problem with a domain *independent* algorithm. Further potential advantages will be discussed in the last section.

The rest of the paper is organized as follows. In Section 2 we formally define the partial weighted set cover problem, show its computational intractability, and give a local search algorithm for approximating its solution. In Section 3

we adapt our generic approach to set systems defined by d -balls and empirically demonstrate the usefulness of our method on clustering and outlier detection problems. Finally, in Section 4 we discuss our results and present some interesting problems for future work.

2 The Partial Weighted Set Cover Problem

In this section we define the partial weighted set cover problem, show that it is computationally intractable, and present a generic local search algorithm for this problem.

Let S be some finite set and $\mathcal{S} \subseteq 2^S$ be a set system over S . We assume that there is a function $w_X : X \rightarrow R_{\geq 0}$ for all $X \in \mathcal{S}$, where $R_{\geq 0}$ denotes the set of non-negative real numbers. Thus, for all sets X in \mathcal{S} and for all $x \in X$, $w_X(x)$ defines the *relative weight* of x with respect to X . In addition to the relative weights on the elements of X , we assume that there are two set functions W and G , both mapping \mathcal{S} to $\mathbb{R}_{\geq 0}$. While

$$W(X) = \sum_{x \in X} w_X(x)$$

defines the *weight* for all $X \in \mathcal{S}$, $G(X)$ specifies the *generality* of X . In the applications we consider, $G(X)$ will be defined by the *size* of X , for some appropriate notion of size depending on the particular problem at hand. Using the definitions above we are ready to define the following combinatorial optimization problem considered in this work:

The Partial Weighted Set Cover (PWSC) Problem: Given a weighted set system \mathcal{S} over some finite set as defined above, a positive integer k , and a function $P_O : \mathcal{S}^k \rightarrow R_{\geq 0}$, *find*

$$\operatorname{argmax}_{\mathcal{S}_k \subseteq \mathcal{S}, |\mathcal{S}_k|=k} \mathcal{U}(\mathcal{S}_k) ,$$

with

$$\mathcal{U}(\mathcal{S}_k) = \sum_{X \in \mathcal{S}_k} W(X) - \sum_{X \in \mathcal{S}_k} G(X) - P_O(\mathcal{S}_k) \quad (1)$$

In the definition above, P_O is used to penalize the elements of S that are covered by more than one set from \mathcal{S}_k . Thus, our goal is to select k sets from \mathcal{S} with maximum weights subject to the constraints that the sets must be as specific as possible and the overlap amongst the k sets must be as small as possible. There are many problems that can be regarded as special cases of the PWSC problem. As an example, in the next section we show that soft clustering and outlier detection problems can be viewed as such special cases, allowing these classical data mining problems to be translated into combinatorial optimization problems.

Before giving our algorithm for the PWSC problem, we first discuss its complexity. Not surprisingly, the PWSC problem is computationally intractable. This is stated in the proposition below.

Proposition 1. *The PWSC problem is NP-hard.*

Proof. We prove the claim by using the following polynomial reduction from the decision version of the set cover problem¹. Let \mathcal{S} be a set system over some finite ground set S and define $w_X(x) = 1$, $W(X) = |X|$, $G(X) = 0$, and

$$P_O(\mathcal{S}_k) = \sum_{Y \in \mathcal{S}_k} |Y| - \left| \bigcup_{Y \in \mathcal{S}_k} Y \right|$$

for all $X \in \mathcal{S}$, for all $x \in X$, and for all $\mathcal{S}_k \subseteq \mathcal{S}$ with $|\mathcal{S}_k| = k$. For the output \mathcal{S}_k of the PWSC problem for the weighted set system constructed we have that $\mathcal{U}(\mathcal{S}_k) = |S|$ if and only if there exist k sets in \mathcal{S}_k that cover S . \square

To overcome the computational limitation stated in Proposition 1 above, we resort to a local search algorithm for finding some approximate solution of the PWSC problem (see Alg. 1). The parameters of the algorithm are a weighted set system \mathcal{S} over some finite set S , a positive integer k , and a set function P_O on \mathcal{S}_k . In lines 1 and 2 of the main algorithm, we greedily select k sets maximizing the utility function given in (1). Then, until some termination condition holds, we iteratively call two update functions (lines 3–6).

The input to the first update function (UPDATE_A) is a family $\mathcal{S}_k \subseteq \mathcal{S}$ with $\mathcal{S}_k = \{S_1, \dots, S_k\}$. For every $i = 1, \dots, k$, it selects a proper superset S'_i of S_i from \mathcal{S} such that the replacement of S_i by S'_i in \mathcal{S}_k maximizes the utility function. If the utility of the k sets after the replacement is greater than that of \mathcal{S}_k , we take the new configuration and continue the process with the next set (lines 2–4 of UPDATE_A). Note that this function is greedy, as it updates the k sets in \mathcal{S}_k separately.

In function UPDATE_B we select $O(\log k)$ sets uniformly at random from the input family $\mathcal{S}_k = \{S_1, \dots, S_k\}$ (line 3 of UPDATE_B) and try to shrink/enlarge them without any decrease in the utility. More precisely, for each set $S_i \in \mathcal{S}_k$ selected, we first calculate the family \mathcal{F}_1 of maximal proper subsets (resp. the family \mathcal{F}_2 of minimal proper supersets) of S_i in \mathcal{S} that minimize the L_1 -distance of the relative weights of the elements belonging to the intersection (line 5 resp. line 6). We then select a set from $\mathcal{F}_1 \cup \mathcal{F}_2$ uniformly at random (line 7) and replace S_i in \mathcal{S}_k by the set selected. After having processed all $O(\log k)$ sets in this way, we compare the utility of the new k sets with that of the input ones. We keep the new configuration if its utility is greater or it has the same utility and the outcome of a biased coin flip is HEAD (lines 9–11). The rationale behind the definition of \mathcal{F}_1 and \mathcal{F}_2 is that we would like to avoid big steps during local search.

¹ The decision version of the set cover problem is defined as follows: *Given* a set system \mathcal{S} over some finite set S and a positive integer k , *decide* whether there exist k sets in \mathcal{S} that cover S . This problem is known to be NP-complete.

Algorithm 1 MAIN

Parameters: weighted set system \mathcal{S} over some finite set S , $k \in \mathbb{N}$, and $P_O : \mathcal{S}^k \rightarrow R_{\geq 0}$

- 1: set $\mathcal{S}_0 = \emptyset$
- 2: **for** $i \in [k]$ **do** $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{ \operatorname{argmax}_{X \in \mathcal{S} \setminus \mathcal{S}_{i-1}} \mathcal{U}(\mathcal{S}_{i-1} \cup \{X\}) \}$
- 3: **repeat**
- 4: $\mathcal{S}_k = \text{Update_A}(\mathcal{S}_k)$
- 5: $\mathcal{S}_k = \text{Update_B}(\mathcal{S}_k)$
- 6: **until** some termination condition holds
- 7: **return** \mathcal{S}_k

UPDATE_A(\mathcal{S}_k) with $\mathcal{S}_k = \{S_1, \dots, S_k\}$:

- 1: $U_{\max} = \mathcal{U}(\mathcal{S}_k)$
- 2: **for** $i \in [k]$ **do**
- 3: $S'_i = \operatorname{argmax}_{X \in \mathcal{S}, X \supseteq S_i} \mathcal{U}(\mathcal{S}_k \setminus \{S_i\} \cup \{X\})$
- 4: **if** $\mathcal{U}(\mathcal{S}_k \setminus \{S_i\} \cup \{S'_i\}) > U_{\max}$ **then** $\mathcal{S}_k = \mathcal{S}_k \setminus \{S_i\} \cup \{S'_i\}$, $U_{\max} = \mathcal{U}(\mathcal{S}_k)$
- 5: **return** \mathcal{S}_k

UPDATE_B(\mathcal{S}_k) with $\mathcal{S}_k = \{S_1, \dots, S_k\}$:

- 1: $\mathcal{S}'_k = \mathcal{S}_k$
 - 2: **for all** $i \in [k]$ **do**
 - 3: flip a biased coin with $\Pr(\text{HEAD}) = \frac{\log(k)}{k}$
 - 4: **if** the outcome is HEAD **then**
 - 5: $\mathcal{F}_1 = \{X \in \mathcal{F}'_1 : \nexists Y \in \mathcal{F}'_1 \text{ with } Y \supseteq X\}$ with
$$\mathcal{F}'_1 = \{Z \in \mathcal{S} : Z \subsetneq S_i \text{ and } \sum_{x \in Z} |w_Z(x) - w_{S_i}(x)| \text{ is minimum}\}$$
 - 6: $\mathcal{F}_2 = \{X \in \mathcal{F}'_2 : \nexists Y \in \mathcal{F}'_2 \text{ with } Y \subsetneq X\}$ with
$$\mathcal{F}'_2 = \{Z \in \mathcal{S} : Z \supseteq S_i \text{ and } \sum_{x \in S_i} |w_Z(x) - w_{S_i}(x)| \text{ is minimum}\}$$
 - 7: select a set S'_i uniformly at random from $\mathcal{F}_1 \cup \mathcal{F}_2$
 - 8: set $\mathcal{S}'_k = \mathcal{S}_k \setminus \{S_i\} \cup \{S'_i\}$
 - 9: **if** $\mathcal{U}(\mathcal{S}'_k) > \mathcal{U}(\mathcal{S})$ **then** $\mathcal{S}_k = \mathcal{S}'_k$
 - 10: **else if** $\mathcal{U}(\mathcal{S}'_k) = \mathcal{U}(\mathcal{S})$ **then**
 - 11: set $\mathcal{S}_k = \mathcal{S}'_k$ if the outcome of a biased coin flip is HEAD
 - 12: **return** \mathcal{S}_k
-

3 Applications

To demonstrate the practical usefulness of our approach, in this section we present the applications of the PWSC problem to two classical problems of data mining: To clustering and outlier detection. In case of clustering, the task is to identify subsets of observations (i.e., *clusters*) minimizing the inter-cluster

distances (i.e., the distance between instances within the same subset) and maximizing the inter-cluster distances (i.e., the distance between clusters). Note that this informal definition applies to soft clustering as well. Regarding outlier detection, we use the following definition: “*An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*” [1]. Thus, the goal of outlier detection is to distinguish the set of outlier observations from that of the inlier ones. Similarly for example to DBSCAN [2], we reduce the outlier detection problem to clustering; all instances not belonging to any of the clusters are regarded as outliers. In order to model the two problems above by the PWSC problem and to apply Algorithm 1, we need to construct an appropriate weighted set system and define the reward function P_O .

3.1 The PWSC Problem for Clustering and Outlier Detection

Both clustering and outlier detection use a concept of similarity between observations. We consider the case that the observations form a finite set $S \subseteq \mathbb{R}^d$ for some d and that the similarity between observations is defined by some metric D on \mathbb{R}^d . For each point $P \in S$, we consider a set of d -balls around P for all radii defined by the elements of a finite geometric progression for some scale factor. The set system \mathcal{S} over S is then defined by the family of subsets of S , each covered by such a ball centered around a point P for some $P \in S$. We will refer to the resulting PWSC as *BallCover*.

More precisely, we assume without loss of generality that the smallest distance between two different points in S is 1, i.e.,

$$\min_{P_1, P_2 \in S, P_1 \neq P_2} D(P_1, P_2) = 1 .$$

Given some positive real number θ defining the scale factor $1 + \theta$, we define $L \in \mathbb{N}$ by

$$L = \lceil \log_{1+\theta} R \rceil ,$$

where $R = \max_{P_1, P_2 \in S} D(P_1, P_2)$. Thus,

$$D(P, Q) \leq (1 + \theta)^L$$

for all $P, Q \in S$. For all $P \in S$, we determine an integer $0 \leq L_P \leq L$ that gives an upper bound on the set of balls of center P ; the algorithm calculating L_P is discussed below. Using these concepts, for S and θ above we define the set system \mathcal{S} over S by

$$\mathcal{S} = \{S_{P,l} : P \in S, 0 \leq l \leq L_P\}$$

with

$$S_{P,l} = \{P' \in S : D(P, P') < (1 + \theta)^l\} .$$

The definitions imply that $S \in \mathcal{S}$ and that $S_{P,L} = S$ for all $P \in S$.

For all $P \in S$ and for all $l = 0, 1, \dots, L_P$, we define the relative weights of the instances in $S_{P,l}$ by

$$w_{S_{P,l}} : Q \mapsto \frac{1}{i+1}$$

for all $Q \in S_{P,i} \setminus S_{P,i-1}$ and for all $i = 0, 1, \dots, l$, where $S_{P,-1} = \emptyset$. That is, $S_{P,l}$ is partitioned into $l+1$ annuli, where the 0th annulus is the point P , and the relative weight of a point Q belonging to the i th annulus is $1/(i+1)$, i.e., it is inversely proportional to the distance of the annulus from P . In this way, we disregard the exact distance for any two points belonging to the same annulus in $S_{P,l}$.

We now specify the generality (G) and the penalty function (P_O) for \mathcal{S} . Regarding the generality, we define it by

$$G(S_{P,l}) = \lambda(1 + \theta)^l$$

for all $P \in S$ and for all $l = 0, 1, \dots, L_P$, where $\lambda \geq 0$ is some user specified parameter. Regarding P_O , let \mathcal{S}' be a subset of \mathcal{S} and let $S' \subseteq S$ be the set of points contained by at least two sets in \mathcal{S}' . Then $P_O(\mathcal{S}_k)$ is defined by

$$P_O(\mathcal{S}_k) = \sum_{x \in S'} \left(\sum_{X \in \mathcal{S}'} w_X(x) - \max_{X \in \mathcal{S}'} w_X(x) \right) .$$

That is, according to the definition of the utility function in (1), we subtract all relative weights of a point x covered by more than one set, except for the highest one. Finally we note that if a subset of S has more than one ball representation, we take the ball (and the corresponding weights) that has the highest weight.

It remains to discuss the determination of the upper bound L_P for a point $P \in S$. Since balls having large annuli of low density are poor choices for clustering, we need to disregard them as candidates. To find the maximum ball around P that contains no annuli of low density, we observe the change in density as a function of the radius while growing the ball. The density is measured by the number of instances covered relative to the ball's radius. Accordingly, we sort the instances by their distance to P . The position in this list then provides the number of instances covered at the respective distance, giving rise to a monotone function sampled at finitely many non-equidistant positions. Our goal is to estimate the first plateau of this unknown continuous density function. To achieve this, we first interpolate the function value at equidistant positions using nearest neighbor interpolation. We then approximate the first derivative by folding the interpolated signal using a Sobel kern. Finally, we smooth the result and determine the first position that is numerically a zero point. This position defines the maximal radius L_P we consider for the ball around P . Due to space limitations we omit the formal definition of this algorithm.

3.2 Experiments

To demonstrate the usefulness of our *BallCover* approach to the tasks of outlier detection and clustering, we have conducted a series of experiments. We have

compared our results achieved on synthetic and real-world data from the UCI machine learning repository[3] to those obtained by state-of-the-art algorithms. The problems of outlier detection and clustering have been studied extensively in the past. As a result, a wide range of different concepts and algorithms have been proposed. For instance, there are outlier detection algorithms using information theoretical criteria, spectral decomposition, clustering, proximity, or density. For a recent overview of outlier detection algorithms, the reader is referred to [4].

An exhaustive comparison to all relevant outlier detection and clustering algorithms is beyond the scope of this work. Therefore, we focus only on algorithms using density criterion to identify outliers or clusters, as these are the most similar methods to the algorithm proposed in this paper. More precisely, we consider the following algorithms:

Local outlier factor (LOF) [5] The algorithm identifies outliers by comparing the density of a point with that of its surrounding points. The size of the surrounding neighborhood is specified by the user supplied parameter *MinPts*. Within the *MinPts* neighborhood around a point, the local outlier factor (LOF) is calculated as the average density of all points in the neighborhood normalized by the points own densities. Points with density much lower than their neighbors produce a high LOF value and are considered outliers. For our experiments we use the implementation available with the ELKI [6] toolkit. To identify the set of outliers, we order the instances according to their LOF value and select the top n instances, where n is the true number of outliers. In our experiments on synthetic data we set the *MinPts* parameter (i.e., the number of instances in a cluster) to 1000; other choices (10, 20, 100) of this parameter have not lead to better results. For the UCI datasets we follow [7] and set $MinPts = 10$.

Support vector novelty detection (SVND) [8] is an extension of support vector machines (SVM) to the case of unlabeled data. In SVM the maximal separating hyperplane is determined by the location of instances with different labels in the feature space. However, there are no labels in SVND. Therefore, the goal is to find a simple subset of instances such that the probability of an instance falling into this set meets a probability threshold parameter ν . The boundary of the set is expressed in terms of a kernel expansion and its complexity is controlled by empirical risk minimization. The algorithm takes the probability threshold ν and the kernel as parameters. For our experiments we set ν to the true fraction of outliers and use the Gaussian kernel. The Gaussian kernel itself requires the specification of the variance σ , which we set to the median distance of all points in the dataset, following the recommendation in [9]. In our experiments we use the open source implementation libsvm [10].

DBSCAN [2] constructs a clustering by expanding clusters around dense points, called core points. A point is dense if it has at least *MinPts* neighbors in a distance at maximum ϵ . All points in the neighborhood are recursively added to the cluster as long as they have *MinPts* neighbors. Points that do not belong to any cluster are considered as outliers. For our experiments we use

the implementation available from sklearn [11], i.e., we leave the parameter *MinPts* at the default choice of 10 and set ϵ to the medium of pairwise distances between two points in the data.

Isolation Forest [7] constructs an ensemble of trees, by randomly choosing attributes and splits at each inner node of a tree. The tree growing stops once each instance is isolated in a single leaf or the tree exceeds a height threshold. Each tree is grown on a random sample of data and the tree height is restricted to the height of a binary tree with number of leaves equal to the sample size. Each instance is scored by the expected average path length between the tree root and the node containing the instance. Instances with a short path length are isolated earlier from the rest of the data and are considered as outliers. For our experiments we use the implementation available in sklearn [11] and keep all parameters to their defaults (ensemble size 100, data sample size of 256). The authors propose instances with path length measure much smaller than 0.5 to be considered as inliers. We therefore use this threshold to determine the prediction of outliers.

For our empirical evaluation, we need datasets with known ground truth, i.e., for which we know which instances are outliers within the data. Therefore, we create synthetic datasets and use publicly available classification datasets from the UCI repository for comparison. For the synthetic data, we place k -Gaussians uniformly at random in a hypercube, each with a random diagonal co-variance matrix. The centers may be generated close to each other and the resulting distributions may have arbitrary overlap. We draw the same number of instances from each Gaussian and add uniform noise over the hypercube extended by the largest 3σ . The parameters for the data we generated here are as follows: the center coordinates of the Gaussians are drawn uniformly at random from the interval $[0, 20]$ and the variances σ^2 from $[0, 1]$. We generated ten 2-dimensional datasets with four Gaussians, having 1000 samples each and added 20% uniform noise samples as outliers. Regarding the real-world data from the UCI repository, we follow the transformation used in [7] to construct an outlier detection task from the corresponding multi-class classification task, i.e., we consider classes 3,4,5,7,8,9,14,15 as outliers for the *arrhythmia* set, classes 1,2 for *anthyroid*. The *pima* and *ionosphere* datasets are binary classification tasks, the minority class are considered as outliers. We further take *wilt* and a random sample of the *adult* dataset into account using the same proposition. Our selection includes datasets of small up to large size (350-7000) and of low and high dimensions (5-274), to cover a broad range of application scenarios.

We tested all algorithms described above by applying them to the full datasets including the outliers during the training stage. The ground truth labels are not presented to the learner; they are only used in the calculation of the performance measure. We use F_1 -score, with the normal instances as positive class, to assess the quality, as it accounts for the class imbalance. For our algorithm, we use the true number of balls with the synthetic datasets and report results for different choices of k on the UCI datasets. Further, we fix the parameters $\theta = 0.1$ and $\lambda = 0.05$ for all experiments. The results achieved for the different combinations

UCI Dataset	BallCover			LOF	SVND	DBSCAN	IFOREST
	k=2	k=8	k=20				
adult (sample)	0.79	0.86	0.87	0.80	0.79	0.86	0.87
arrythmia	0.93	0.92	0.87	0.88	0.57	0.89	0.92
annthyroid	0.52	0.74	0.68	0.93	0.93	0.95	0.96
ionosphere	0.70	0.84	0.68	0.91	0.88	0.83	0.87
pima	0.76	0.71	0.67	0.64	0.72	0.79	0.79
wilt	0.92	0.86	0.88	0.95	0.94	0.97	0.93
Synth. Dataset	k=4						
4-Gauss (mean)	0.95			0.98	0.95	0.97	0.97

Table 1. Summary of F_1 scores of outlier detection algorithms on UCI machine learning tasks and synthetic data sets. For the ten synthetic 2-dimensional Gaussian mixtures datasets we report the mean value of F_1 scores.

of the datasets and algorithms are summarized in Table 1. As we can see, the performance of our approach depends on the choice of k , but in most cases, there is a choice that matches the quality of the competitors (*adult*, *ionosphere*, *pima*, *wilt*, synthetic data). Only for the *annthyroid* dataset, the performance of our algorithm is clearly worse than that of any reference. In turn, we can report a slightly better performance as the best competitor for the *arrythmia* dataset.

For the clustering application we use the synthetic Gaussian datasets. Each of the Gaussians forms a separate cluster and each instance is labeled according to it. The uniform noise represents an additional component and is identified as another cluster id. For comparison, we use the DBSCAN and K-MEANS clustering algorithms as reference. We follow the same parameter selection scheme for DBSCAN as used within our outlier experiments. In contrast to DBSCAN, the K-MEANS algorithm creates a complete partition of all instances, especially without excluding any noise. Therefore we consider different choices of k setting it at least to the true number of Gaussian plus one additional for the noise. We treat the cluster id as target of a multiclass classification problem and assess the performance in terms of the weighted average over the F_1 scores for each of the classes. To match the cluster id used by the algorithm with that of the ground truth, we apply the following mapping: For the K-MEANS and DBSCAN algorithms we assign to each cluster the ground truth label of the majority vote of the instances in that cluster. For our algorithm, we use the center points to select the ground truth label for each ball. The results indicate, that our algorithm performs better than K-MEANS for small choices of k . Increasing k leads to results comparable to our approach for most datasets. There are some cases where K-MEANS performs better and some where our approach is slightly better (c.f. Table 2). Further, across all sets used in this experiment, the performance of our algorithm competes with the results of DBSCAN; in some cases, our algorithm performs much better. A closer look at those cases reveals that our algorithm has an advantage if the centers of the Gaussians are very close to each other, resulting in large overlaps. In such cases, those clusters are merged by DBSCAN and our algorithm is able to separate the corresponding data instances. An ex-

Dataset ID	BallCover	DBSCAN	K-MEANS		
			k=5	k=8	k=16
05d...923	0.74	0.71	0.62	0.84	0.87
0c8...9d2	0.96	0.66	0.60	0.89	0.94
3ac...309	0.97	0.98	0.80	0.89	0.94
420...c47	0.64	0.68	0.63	0.67	0.69
718...f80	0.78	0.68	0.62	0.87	0.89
984...3e4	0.96	0.87	0.60	0.90	0.95
a63...013	0.69	0.69	0.60	0.66	0.88
b1a...3af	0.96	0.98	0.81	0.89	0.93
d2d...0a5	0.95	0.70	0.59	0.87	0.93
d7b...9e2	0.96	0.97	0.62	0.91	0.95

Table 2. Weighted average F_1 scores unsupervised reconstruction of class structure.

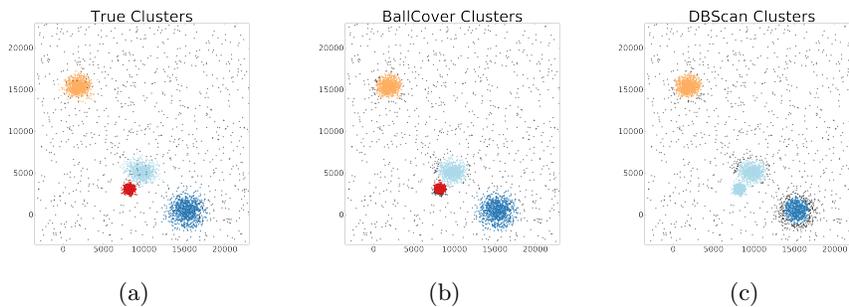


Fig. 1. Cluster structure of synthetic dataset with overlapping Gaussians. Colors indicate cluster memberships, noise cluster is black. True clusters are on the left (a), cluster structure found by BallCover in the middle (b), for DBSCAN on the right (c).

ample of this situation is depicted in Figure 1. The true association of points to clusters is at the left. Note the two overlapping Gaussians at the middle bottom region. Our algorithm seeks dense balls with low overlap and small radii and can detect the structure correctly. For DBSCAN, the two Gaussians are too close to each other and joined into one cluster.

4 Discussion

The approach presented in this paper is a first step towards a systematic study of its applications to other data mining/machine learning problems. The advantage of translating such problems into the partial weighted set system problem is that it allows for the application of techniques developed for combinatorial optimization problems. For example, one might transform the underlying problem into set systems of some advantageous structural properties (e.g., into matroids or greedoids) that can be utilized by the algorithm. In this way, new tractable subclasses of the problem could be identified. Another interesting research direction

is the restriction of the utility function to set function classes of advantageous algorithmic properties.

The distance discretization used in the applications considered in this work raises the question whether our approach can be combined with state-of-the-art techniques improving the speed, such as, for example, with locality sensitive hashing.

Our utility function includes two penalty terms. One of them is concerned with the generality of the elements in the set system. It is an interesting question whether and if so, how can we control the complexity of the output via these terms and the parameter k . Finally, we are going to investigate how to extend our method to an interactive one, in which the set system and the utility function are automatically adapted according to the feedback of an expert (c.f. [12]).

Acknowledgments This research was supported by the EU FP7-ICT-2013-11 project under grant 619491 (FERARI).

References

1. Hawkins, D.: Identification of Outliers. Monographs on applied probability and statistics. Chapman and Hall (1980)
2. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of 2nd International Conference on Knowledge Discovery and. (1996) 226–231
3. Lichman, M.: UCI machine learning repository (2013)
4. Aggarwal, C.: Outlier Analysis. Springer New York (2013)
5. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: Identifying density-based local outliers. SIGMOD Rec. **29**(2) (May 2000) 93–104
6. Schubert, E., Koos, A., Emrich, T., Züfle, A., Schmid, K.A., Zimek, A.: A framework for clustering uncertain data. PVLDB **8**(12) (2015) 1976–1979
7. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. (Dec 2008) 413–422
8. Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7) (July 2001) 1443–1471
9. Caputo, B., Sim, K., Furesjo, F., Smola, A.: Appearance-based object recognition using svms: Which kernel should i use? Proc of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision, Whistler **2002** (2002)
10. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12** (2011) 2825–2830
12. Boley, M., Mampaey, M., Kang, B., Tokmakov, P., Wrobel, S.: One click mining-interactive local pattern discovery through implicit preference and performance learning. In: KDD 2013 Workshop on Interactive Data Exploration and Analytics (IDEA). (2013)

Variable Attention and Variable Noise: Forecasting User Activity

César Ojeda, Kostadin Cvejoski, Rafet Sifa, and Christian Bauckhage

Fraunhofer IAIS, Germany
{name.surname}@iais.fraunhofer.de

Abstract. The study of collective attention is of growing interest in an age where mass- and social media generate massive amounts of often short lived information. That is, the problem of understanding how particular ideas, news items, or memes grow and decline in popularity has become a central problem of the information age. Recent research efforts in this regard have mainly addressed methods and models which quantify the success of such memes and track their behavior over time. Surprisingly, however, the aggregate behavior of users over various news and social media platforms where this content originates has large been ignored even though the success of memes and messages is linked to the way users interact with web platforms. In this paper, we therefore present a novel framework that allows for studying the shifts of attention of whole populations related to websites or blogs. The framework is an extension of the Gaussian process methodology, where we incorporate regularization methods that improve prediction and model input dependent noise. We provide comparisons with traditional Gaussian process and show improved results. Our study in a real world data set, uncovers hidden patterns of user behavior.

Keywords: Gaussian process, Regularization Methods

1 Introduction

Over the last couple of years, so called “question answering” (QA) sites have gained considerable popularity. These are internet platforms where users pose questions to a general population. Yahoo Answers, Quora and the Stack Exchange family establish internet communities which provide natural and seamless ways for organizing and providing knowledge [1]. So far, dynamical aspects of such questions answering sites have been studied in different contexts. Previous work in this area includes studying causality aspects through quasi experimental designs [8], user churn analysis through classification algorithms such as support vector machines or random forests [9], and predictions of the future value of questions answers pairs according to the initial activity of the question post [2]. In contrast to previous work where long term activity of users is being predicted, our focus in this paper is time series analysis related to user-defined tags. This

approach allows detailed daily analysis of the behavior of users and we concentrate on the QA site Stackoverflow. This platform has an established reputation on the web and boasts a community of over 5 million distinct active users who, so far, have provided more 18 million answers to more than 11 million questions. Thanks to the sheer size of the corresponding data set as well as because of the regular activity of the user base, we are able to mine temporal data in order to uncover defining aspects of the dynamics of the user behavior.

Due to the complexity of user-system interaction (millions of people discuss thousands of topics), flexible and accurate models are required in order to guarantee reliable forecasting. In recent years the Bayesian setting and the Gaussian Process (GP) framework [11, 5] has shown to provide an accurate and flexible tool for time series analysis. In particular, the possibility of incorporating error ranges as well as different models with the selection of different kernels, permits interpretability of the results. In this work, we model changes in attention as a variability in the fluctuation of the time series of occurrences of user defined tags which can be categorized as a special case of *heterocedasticity* or input dependent noise. We provide an extension of sparse input Gaussian Processes [15, 14] which allow us to model functional dependence in the time variation of the fluctuations. In practical experiments, we study the top 10 different tags for the Stackoverflow data set over different years, spanning a data set of over 2.9 million questions. We find that our model outperform predictions made by the simple GP model under variable noise. In particular, we uncover weekly and seasonal periodicity patterns as well as random behavior in monthly trends. All in all, we are able to forecast the number of questions within a 5 percent error 20 days in the future.

In the next section, we formally introduce the Gaussian Process framework and provide details regarding our extensions towards variable noise models. We then show an analysis of the periodicity of the time series of tag activity as apparent from the Stackoverflow data set. Next, we compare our prediction results with those of other models and discuss the advantages of introducing functional dependencies on noise terms. Finally, we provide conclusions and directions of future work.

2 A Model for Time Series Analysis

In this section, we propose a Gaussian process (GP) model for regression that extends the sparse pseudo-input Gaussian process (SPGP) for regression [14]. Our model deals with the problem of over fitting that hampers the SPGP model and makes it possible to analyze the function of the uncertainty added to every pseudo-input. Analyzing the uncertainty function, we indirectly analyze the effects of heteroscedastic noise.

A GP is a Bayesian model that is commonly used for regression tasks [11]. The main advantages of this method are its non-parametric nature, the possibility

of interpreting the model through flexible kernel selection, and the confidence intervals (error bars) obtained for every prediction. The non-parametric nature of this method has a drawback, though. The computational cost of the training is $\mathcal{O}(N^3)$, where N is the number of training points. There are many sparse approximation methods of the full GP that try to lower the computational cost of the training to $\mathcal{O}(M^2N)$ where M is the size of the subset of the training points that are used for approximation (i.e. the active set) and typically $M \ll N$ [13, 12]. The M points for the approximation are chosen according to various information criteria. This leads to difficulties w.r.t. learning the kernel hyperparameters by maximizing the marginal likelihood of the GP using gradient ascent. The re-selection of the active set causes non-smooth fluctuations of the gradients of the marginal likelihood, which results likely convergence to sub-optimal local maxima [14].

2.1 Gaussian Process for Regression

Next, we first briefly review the GP model for regression, yet, for a detailed discussion we refer to [11, 10].

Consider a data set \mathcal{D} of size N containing pairs of input vectors $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ and real value target points $\mathbf{y} = \{y_n\}_{n=1}^N$. In order to apply the GP model to regression problems, we need to account for noise in the available target values, which are thus expressed as

$$y_n = f_n + \epsilon_n \quad (1)$$

where $f_n = f(x_n)$ and ϵ_n is a random noise variable which is independently chosen for each n . We shall consider a noise process following a Gaussian distribution defined as

$$p(\mathbf{y} | \mathbf{f}) = \mathcal{N}(\mathbf{y} | \mathbf{f}, \sigma^2 \mathbf{I}) \quad (2)$$

where $\mathcal{N}(\mathbf{y} | \mathbf{m}, \mathbf{C})$ is a Gaussian distribution with mean \mathbf{m} and covariance \mathbf{C} . The marginal distribution of $p(\mathbf{f})$ is then given by another Gaussian distribution, namely $p(\mathbf{f} | \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f} | \mathbf{0}, \mathbf{K}_N)$. The covariance function that determines \mathbf{K}_N is chosen to express the property that, if points \mathbf{x}_n and \mathbf{x}_m are similar, the value $[\mathbf{K}_N]_{nm}$ should express this similarity. Usually, this property of the covariance function is controlled by small number of hyperparameters $\boldsymbol{\theta}$. Integrating out over \mathbf{f} , we obtain the marginal likelihood as

$$\begin{aligned} p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}) &= \int p(\mathbf{y} | \mathbf{f}) p(\mathbf{f} | \mathbf{X}, \boldsymbol{\theta}) \\ &= \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_N + \sigma^2 \mathbf{I}_N), \end{aligned} \quad (3)$$

which is used for training the GP model by maximizing it with respect to $\boldsymbol{\theta}$ and σ^2 . The distribution of the target value of a new point \mathbf{x} will then be

$$\begin{aligned} p(y | \mathbf{x}, \mathcal{D}, \boldsymbol{\theta}) &= \mathcal{N}\left(y | \mathbf{k}_x^T (\mathbf{K}_N + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, K_{\mathbf{x}\mathbf{x}} \right. \\ &\quad \left. - \mathbf{k}_x^T (\mathbf{K}_N + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_x + \sigma^2\right), \end{aligned} \quad (4)$$

where $[\mathbf{k}_x]_n = K(\mathbf{x}_n, \mathbf{x})$ and $K_{xx} = K(\mathbf{x}, \mathbf{x})$. In order to predict with GP model, we need to have all the training data available during run-time, which is why the GP for regression is referred to as a non-parametric model.

2.2 SPGP and SPGP+HS Models

An approximation of the full GP model for regression is presented in [14] in which the authors propose the sparse pseudo-input Gaussian process (SPGP) regression model that enables a search for the kernel hyper-parameters and the active set in a single joint optimization process. This is possible because it is allowed for the active set (pseudo-inputs M) to take any position in the data space, not only to be a subset of the training data. Parameterizing the covariance function of the GP by the pseudo-inputs, gives the possibility for learning the pseudo-inputs using gradient ascent. This is a major advantage, because it improves the model fit by fine tuning the locations of the pseudo-inputs. Let, $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}_m\}_{m=1}^M$ be the pseudo-inputs and $\bar{\mathbf{f}} = \{\bar{f}_m\}_{m=1}^M$ are the pseudo targets, the predictive distribution of the model for a new input \mathbf{x}_* will be given by

$$\begin{aligned} p(y_* | \mathbf{x}_*, \mathcal{D}, \bar{\mathbf{X}}) &= \int p(y_* | \mathbf{x}_*, \bar{\mathbf{X}}, \bar{\mathbf{f}}) p(\bar{\mathbf{f}} | \mathcal{D}, \bar{\mathbf{X}}) d\bar{\mathbf{f}} \\ &= \mathcal{N}(y_* | \mu_*, \sigma_*^2), \\ \mu_* &= \mathbf{k}_*^\top \mathbf{Q}_M^{-1} \mathbf{K}_{MN} (\mathbf{\Lambda} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \\ \sigma_*^2 &= K_{**} - \mathbf{k}_*^\top (\mathbf{K}_M^{-1} - \mathbf{Q}_M^{-1}) \mathbf{k}_* + \sigma^2, \end{aligned} \quad (5)$$

where \mathbf{K}_N is the covariance matrix of the training data, \mathbf{K}_M is the covariance matrix of the pseudo inputs, σ^2 is the noise, \mathbf{Q} is defined as

$$\mathbf{Q} = \mathbf{K}_{MN} (\mathbf{\Lambda} + \sigma^2)^{-1} \mathbf{K}_{NM} \quad (6)$$

and $\mathbf{\Lambda}$ is defined as

$$\mathbf{\Lambda} = \text{diag}(\mathbf{K}_N - \mathbf{K}_{NM} \mathbf{K}_M^{-1} \mathbf{K}_{MN}). \quad (7)$$

Finding the pseudo input locations $\bar{\mathbf{X}}$ and the hyperparameters (kernel parameters and noise) $\Theta = \{\theta, \sigma^2\}$ can be done by maximizing the marginal likelihood (8) with respect to the parameters $\{\bar{\mathbf{X}}, \Theta\}$.

$$\begin{aligned} p(\mathbf{y} | \mathbf{X}, \bar{\mathbf{X}}, \Theta) &= \int p(\mathbf{y} | \mathbf{X}, \bar{\mathbf{X}}, \bar{\mathbf{f}}) p(\bar{\mathbf{f}} | \bar{\mathbf{X}}) d\bar{\mathbf{f}} \\ &= \mathcal{N}(\mathbf{y} | 0, \mathbf{K}_{NM} \mathbf{K}_M^{-1} \mathbf{K}_{MN} + \mathbf{\Lambda} + \sigma^2 \mathbf{I}) \end{aligned} \quad (8)$$

One positive effect of the sparsity of the SPGP model is the capability of learning data sets that have variable noise where the term *variable noise* refers to noise which depends on the input. However, it is important to note, that this capability is limited and an improvement of the SPGP model is presented in

[15]. Introducing an additional uncertainty h_m parameter to every pseudo-input point makes the model more flexible and allows for improved representations of heteroscedastic data sets. The covariance matrix of the pseudo-inputs is defined by

$$\mathbf{K}_M \rightarrow \mathbf{K}_M + \text{diag}(\mathbf{h}), \quad (9)$$

where \mathbf{h} is a positive vector of uncertainties that needs to be learned and $\text{diag}(\mathbf{h})$ represents a diagonal matrix whose elements are those of the \mathbf{h} vector. This extension allows the possibility of gradual influence on the pseudo inputs. This means that if uncertainty $h_m = 0$, then the pseudo input m behaves like in the standard SPGP. Yet, as h_m grows the particular pseudo input has less influence on the predictive distribution. This possibility of partially turning off the pseudo inputs allows a larger noise variance in the prediction. The authors of [15] refer to this as heteroscedastic (input dependable noise) extension SPGP+HS.

2.3 SPGP+FUNC-HS

Introducing the heteroscedastic extension to the SPGP empowers the model to learn from data sets with varying noise. However, making the model this flexible may cause problems of over fitting. Also, using the SPGP+HS to predict user and website activities, does not allow us to interpret the behavior of the noise because noise is represented as a positive vector \mathbf{h} of uncertainties and attempts of interpreting these values do not yield meaningful information about the behavior of the noise.

One way of solving the problems of over fitting and lack of interpretability will be to put a prior distribution over the vector \mathbf{h} of uncertainties. However, taking this approach leads to computationally intractable integrals.

The solution which we propose for these problems is to make use of an uncertainty function that depends on the pseudo-inputs. Our covariance function of the pseudo-inputs is defined as

$$\mathbf{K}_M \rightarrow \mathbf{K}_M + \text{diag}(f_h(\bar{\mathbf{x}}_m)), \quad (10)$$

where f_h is the uncertainty function and $\bar{\mathbf{x}}_m$ is a pseudo-input. By defining the heteroscedastic extension in this way, it is possible for the parameters of the uncertainty function to be learned by the gradient based maximum likelihood approaches. Hence, later on, we are able to interpret the parameters of the heteroscedastic noise function as parameters that govern the noise in the model. Another advantage of having a heteroscedastic function is that it restricts the parameter search space when learning the model. This restriction can be beneficial when learning the model, because, it removes unnecessary local maxima. This results in much faster convergence when learning the model and also in improved chances of reducing over fitting. In the following, we will refer to our new heteroscedastic function model as **SPGP+FUNC-HS**.

For modeling the Stackoverflow data set, we introduce two heteroscedastic noise functions. In general, we may use any function that can describe the noise of the given data set. The first heteroscedastic noise function which we consider is the simple sine function defined by

$$f_h(\bar{\mathbf{x}}_m) = a \sin(2\pi\omega\bar{\mathbf{x}}_m + \varphi), \quad (11)$$

where a is the amplitude, ω is the frequency and φ is the phase. We refer to this model as **SPGP+SIN-HS**. The second heteroscedastic noise function we investigate is a product of the sine function and an RBF kernel, namely

$$f_h(\bar{\mathbf{x}}_m, \mathbf{h}_m) = c^2 e^{-\frac{(\bar{\mathbf{x}}_m - \mathbf{h}_m)^2}{2l^2}} \sin(2\pi\omega\bar{\mathbf{x}}_m + \varphi), \quad (12)$$

where c is the variance, \mathbf{h}_m is a mean associated with every pseudo-input $\bar{\mathbf{x}}_m$ in the RBF kernel, and l is the length scale of the RBF kernel. The mean in the RBF kernel can be initialized at random or set by the user if the user has corresponding prior knowledge. Setting a mean for every pseudo-input point divides the whole input space into regions where, in each region, we have a function governing the uncertainty associated with every pseudo input. The uncertainty function defined like this then behaves like mixture of experts and we refer to this model as **SPGP+RBFSIN-HS** model.

3 Results

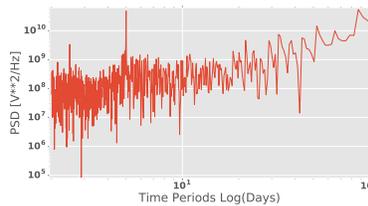


Fig. 1: Spectral Density Estimation on the Stackoverflow dataset using a periodogram. We observe two peaks at two and a half and five days, where the latter peak is the doubled period of the former peak.

In the previous section, we presented the Gaussian process method and two extensions of this method, the SPGP+HS and the SPGP+FUNC-HS. In this section, we present results we obtained when using these models on our Stackoverflow data set.

In order to test our models, we used publicly available data-dumps of Stackoverflow¹. The data set contains the number of questions and answers of postings

¹ Downloadable URL: www.archive.org/details/stackexchange

	MSE		NLPD				NLML					
	GP	RBFSIN	HS	SIN	GP	RBFSIN	HS	SIN	GP	RBFSIN	HS	SIN
android	960.88	692.03	887.45	720.75	4.65	4.49	4.61	4.49	-1076.37	-948.40	-1149.22	-993.23
c#	1029.06	881.11	950.64	894.61	4.70	4.62	4.64	4.62	-1003.23	-949.54	-962.43	-961.62
c++	1216.94	533.68	5068.20	675.84	4.84	4.45	6.02	4.66	-717.14	-698.50	-756.95	-716.58
html	681.57	678.19	774.17	754.95	4.47	4.45	4.51	4.50	-841.93	-784.78	-798.28	-820.60
ios	2598.35	1474.72	3064.63	1660.90	5.82	4.81	5.53	4.86	-757.36	-737.24	-750.69	-740.49
java	1917.86	1431.70	3446.30	1782.17	5.12	4.90	5.79	4.95	-1098.13	-1034.83	-1087.29	-1068.30
javascript	2992.30	1869.61	2396.68	2102.05	6.09	4.97	5.52	5.22	-1493.42	-883.31	-1054.49	-1044.76
jquery	808.26	825.28	989.07	1163.88	4.57	4.77	4.69	4.73	-957.31	-932.99	-866.17	-862.45
php	5892.26	907.07	5379.89	2745.40	6.83	4.60	6.13	5.15	-1042.95	-883.65	-945.85	-853.21
python	604.89	702.25	744.28	881.65	4.44	4.58	4.53	4.62	-782.68	-842.76	-787.24	-788.14

Table 1: Results showing the MSE and NLPD (smaller better) on the 2014 question test set and NLML (larger is better) on the 2014 question training set. GP indicates a pure Gaussian process, HS indicates a sparse pseudo-input Gaussian process with heteroscedastic noise, SIN-HS refers to a sparse pseudo-input Gaussian process with sine functional noise, and RBFSIN-HS refers to a sparse pseudo-input Gaussian process with sine and RBF kernel functional noise.

classified by tag for every business day. The models are trained on a data set containing information about daily postings in the time between 01.02.2014 to 31.08.2014. The evaluation of the models is done on a test set containing postings for the first 21 working days in September 2014.

The performance of the presented models depends on the choice of the kernels used for the covariance matrix. When working with GPs, an additional analysis is required to select proper kernels for the covariance matrix. Because we work with a data set that reflects user behavior, we supposed that it may show a form of periodicity in the behavior of the users. Accordingly, we performed *spectral density estimation* analysis [6, 4, 7] of the time series using a *periodogram* analysis [6, 4, 7]. This analysis shows the power (amplitude) of the time series as a function of frequency and in this way we are able to verify if there are indeed periodicities and, if so, at what frequency they occur.

A periodogram of the time series data that we are analyzing is shown in Figure 1. Since all our tag related time series tag have almost the same periodogram, we only show one of them. For better interpretability we converted the frequencies into periods to observe in how many days the periods occur. There are two apparent peaks, the first occurring at two and a half days and the second at five days. In this case the period of five days appears as an echo of the two and a half days period, therefore we dismiss the second period and we only take into account the first period. Additional characteristics of this data set are minor irregularities and a long term rising trend in the overall time series.

Given these observations, our models that show the best performance as a covariance function use a sum of four kernels

$$k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x') + k_4(x, x'). \quad (13)$$

The question of how to choose these kernels and the particular role of each kernel in the learned model will be discussed in the next section.

Next, we present the result achieved for the top ten tags according to the number of posted questions and answers in the 2014 Stackoverflow data set. Table 1 presents the results of the different models of the posted questions time series and Table 2 presents the results of the different models for the posted answers time series. In order to compare the prediction models, we considered the following measures:

- **Mean Square Error (MSE)** to accounts for the accuracy of the prediction of an unseen data point
- **Negative Log Predictive Distribution (NLPD)** to obtain a confidence for the predicted values on an unseen data point
- **Negative Log Marginal Likelihood (NLML)** to account for how well the model fits the training data.

	MSE			NLPD			NLML					
	GP	RBFSIN	HS	SIN	GP	RBFSIN	HS	SIN	GP	RBFSIN	HS	SIN
android	1097.05	1098.29	1041.58	1031.10	4.80	4.79	4.81	4.78	-903.82	-919.78	-913.40	-927.61
c#	2889.76	2723.95	2998.26	2878.46	5.24	5.18	5.24	5.22	-1007.62	-983.75	-989.81	-995.13
c++	1602.27	1436.71	3491.81	3010.85	4.89	4.85	6.21	5.15	-825.76	-805.98	-886.82	-775.62
html	1856.82	2016.96	2162.96	1904.25	4.98	4.99	5.02	4.96	-1082.09	-957.67	-907.46	-954.44
ios	3944.90	1541.55	5156.82	5017.53	5.74	4.87	5.48	5.41	-831.93	-839.15	-778.98	-777.68
java	3207.22	2987.25	4085.50	3090.13	5.38	5.19	5.35	5.20	-1283.56	-1016.72	-1024.00	-1047.48
javascript	5360.20	4869.97	5434.37	5374.24	5.61	5.50	5.68	5.51	-1141.66	-1110.28	-1139.14	-1131.77
jquery	1817.16	1728.42	1749.74	1725.81	5.12	5.03	5.07	5.00	-976.82	-1021.99	-1009.31	-1023.81
php	2950.13	2948.65	3076.88	2982.74	5.16	5.16	5.20	5.17	-1011.84	-1015.56	-995.81	-994.36
python	911.70	606.00	1660.13	605.22	4.64	4.64	4.90	4.64	-867.67	-820.73	-792.96	-813.29

Table 2: Results showing the MSE and NLPD (smaller is better) on the 2014 answers test set and NLML (larger is better) on the 2014 answers training set. GP indicates a pure Gaussian process, HS indicates a sparse pseudo-input Gaussian process with heteroscedastic noise, SIN-HS refers to a sparse pseudo-input Gaussian process with sine functional noise, and RBFSIN-HS refers to a sparse pseudo-input Gaussian process with sine and RBF kernel functional noise.

For the MSE and the NLPD measures, smaller values are better, and for the NLML larger values are better. The best model for each tag has been chosen using the Akaike information criterion (AIC) [3]. We observe that models with functional noise perform better in nine of the ten tags in the answer time series, and eight of the ten tags in the question time series. The superior performance of the SPGP+FUNC-HS over the full GP can be attributed to the fact that the data set contains variable noise. Note that for this data set, SPGP+FUNC-HS performs better, because of the sparsity of the model and the additional functional noise that is added to the pseudo-inputs. SPGP+HS performs worse than the best models because, adding only a positive vector of uncertainty increases

the flexibility of the covariance function, which at the end can lead to over fitting and convergence to bad local maxima. Using a functional noise constraint, the optimization space shrinks and implicitly prunes bad local maximas. The drawback is that the function of the noise should follow the distribution of the noise in the data set, otherwise the model will perform poorly. This is probably the case why the SPGP+FUNC-HS performs worse on one tag for the answers and on two tags for the questions.

In Fig. 2, we present two learned models, one for the tag “Java” (Fig. 2a) and one for the tag “iOS” (Fig. 2b). We observe that the model that models the Java tag strives to predict the test point using the mean. In contrast, the model that models the iOS tag predicts the test points in terms of noise.

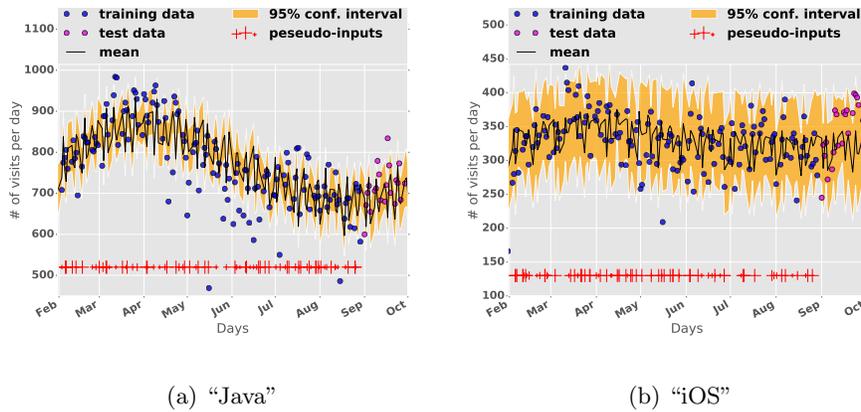


Fig. 2: Models learned with SPGP+SIN-HS for the tags “Java” and “iOS” in the 2014 data set.

4 Analysis

The different kernels in Eq. (13) allow us to dissect the dynamical behavior of the population w.r.t. different scales and patterns. In order to portrait these behaviors, we calculated the mean function and variance Eq. (5) by generating vector \mathbf{k}_*^T using independent kernels. We present the values of each kernel in the “android” question data set in Fig. 3

- **Mean trends** (Fig. 3a) characterize the behavior of the population of users over scales measured in months and represent the global mean behavior of the population. We hypothesize that they are driven by the sheer size of the user base. The more people interested in the tag are visiting the site, the higher the average number of questions per month. Further, this overall trend might

represent the changes in the dominance of this particular tag of questions in the data set. Because the tag refers to a programming language, trends like this indicate changes in attention to various languages. Such dynamics are modeled using the rational quadratic kernel

$$k_1(x, x') = \theta_6^2 \left(1 + \frac{(x - x')^2}{2\theta_8\theta_7^2} \right)^{-\theta_8} \quad (14)$$

- **Seasonal trends** (Fig. 3b) arise on a time scale smaller than major trends and show both periodical and stochastic patterns. They represent changes in the population behavior throughout the different months of the year which can be uncovered with the Ornstein-Uhlenbeck kernel

$$k_2(x, x') = \theta_1 \exp\left(-\frac{|x - x'|}{\theta_2}\right). \quad (15)$$

- **Weekly periods** (Fig. 3c) as obtained from the periodogram represent weekly usage patterns and fine grained periods of activity in our data set. We hypothesize that such behaviors are related natural work patterns during the working week and model them using the following kernel (16).

$$k_3(x, x') = \theta_3^2 \exp\left(L_1 + L_2\right) \quad (16)$$

where we define L_1 and L_2 as

$$L_1 = -\frac{(x - x')^2}{2\theta_4^2} \quad (17)$$

and

$$L_2 = -\frac{2 \sin^2[\pi(x - x')/P]}{\theta_5^2}. \quad (18)$$

- **Weekly noise** (Fig. 3d) are fluctuations in the weakly behavior which can be expected due to the statistical nature of our data set. Randomness in the behavioral pattern of each user might give rise to fluctuations which we model using the following kernel

$$k_4(x, x') = \theta_9^2 \left(1 + \frac{(x - x')^2}{2\theta_{11}\theta_{10}^2} \right)^{-\theta_{11}} \quad (19)$$

5 Conclusion and Future Work

In this paper, we addressed the problem of forecasting the daily posting behavior of users of the Stackoverflow question answering web platform. In order to accomplish this task, we extended the variable noise pseudo inputs Gaussian

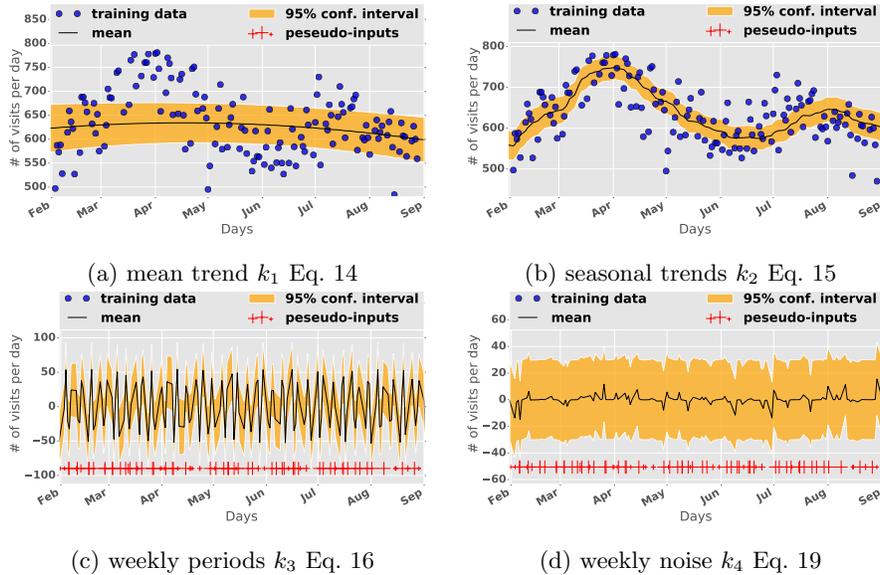


Fig. 3: Decomposition of the SPGP+SIN-HS model for the “android” tag in different kernels. We observe four main behaviors: mean trends, seasonal trends, weekly periods, and weekly noise.

Process framework by introducing a functional noise variant. The idea of using functional descriptions of noise allowed us to study periodic patterns in collective attention shifts and was found to act as a regularizer in model training.

Our extended Gaussian Process framework with functional representations of various kinds of noise provides the added advantage of increased interpretability of results as the different kernels defined for this purpose can uncover different kinds of dynamics. In particular, our kernels revealed major distinct characteristics of the question answering behavior of users. First of all, there are major trends on time scales of about six months showing growing and declining interest in particular topics or corresponding tags. Second of all, these major trends are perturbed by seasonal behavior, for example overall activities usually drop during the summer season. Third of all, on a fine grained scale, there are weekly patterns characterized by periods of 2.5 days. Fourth of all, there are noisy fluctuations in activities on daily scales.

Given the models and results presented in this paper, there various directions for future work. First and foremost, we are currently working on implementing a distributed Gaussian Process framework in order to extend our approach towards massive amounts of behavioral data (use of tags, comments, and likes) that can be retrieved from similar social media platforms such as Twitter or Facebook.

References

1. L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proc. of ACM WWW*, 2008.
2. A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow. In *Proc. of ACM KDD*, 2012.
3. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
4. J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
5. K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th international conference on Machine learning*, pages 393–400. ACM, 2007.
6. D. G. Manolakis, V. K. Ingle, and S. M. Kogon. *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing*. Artech House Norwood, 2005.
7. C. Ojeda, R. Sifa, and C. Bauckhage. Investigating and Forecasting User Activities in Newsblogs: A Study of Seasonality, Volatility and Attention Burst. *Work On Progress*, 2016.
8. H. Oktay, B. J. Taylor, and D. D. Jensen. Causal Discovery in Social Media Using Quasi-experimental Designs. In *Proc. of ACM Workshop on Social Media Analytics*, 2010.
9. J. S. Pudipeddi, L. Akoglu, and H. Tong. User Churn in Focused Question Answering Sites: Characterizations and Prediction. In *Proc. of ACM WWW*, 2014.
10. C. E. Rasmussen. *Evaluation of Gaussian Processes and Other Methods for Non-linear Regression*. PhD thesis, University of Toronto, 1996.
11. C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2005.
12. M. Seeger. Pac-bayesian Generalisation Error Bounds for Gaussian Process Classification. *J. Mach. Learn. Res.*, 3:233–269, 2003.
13. M. Seeger, C. Williams, and N. Lawrence. Fast Forward Selection to Speed Up Sparse Gaussian Process Regression. In *Proc. of Workshop on Artificial Intelligence and Statistics*, 2003.
14. E. Snelson and Z. Ghahramani. Sparse Gaussian Processes Using Pseudo-inputs. In *Proc. of NIPS*, 2005.
15. E. Snelson and Z. Ghahramani. Variable Noise and Dimensionality Reduction for Sparse Gaussian Processes. *arXiv preprint arXiv:1206.6873*, 2012.

Mining Subgroups with Exceptional Transition Behavior

(Abstract)

Florian Lemmerich^{1,5}, Martin Becker², Philipp Singer^{1,5},
Denis Helic³, Andreas Hotho^{2,4}, and Markus Strohmaier^{1,5}

¹ GESIS, Cologne, Germany

{florian.lemmerich, philipp.singer, markus.strohmaier}@gesis.org

² University of Würzburg, Germany

{becker, hotho}@informatik.uni-wuerzburg.de

³ Graz University of Technology, Graz, Austria

dhelic@tugraz.at

⁴ L3S Research Center, Hannover, Germany

⁵ University of Koblenz-Landau, Mainz, Germany

Abstract. In this talk, we present a recently developed method for detecting interpretable subgroups with exceptional transition behavior in sequential data [1]. Potential applications of this technique include, e.g., studying human mobility or analyzing the behavior of internet users. To approach this task, we extend exceptional model mining. Exceptional model mining provides a framework for mining interpretable data subsets with unusual interactions between a set of target attributes considering a user-chosen model class. However, previously investigated model classes cannot capture transition behavior. Thus, we introduce first-order Markov chains as a novel model class for exceptional model mining and present a new interestingness measure that quantifies the exceptionality of transition subgroups. The measure compares the distance between the Markov transition matrix of a subgroup and the respective matrix of the entire data with the distance of random dataset samples. In addition, our method can be adapted to find subgroups that match or contradict given transition hypotheses. We demonstrate that our method is consistently able to recover subgroups with exceptional transition models from synthetic data and illustrate its potential in two application examples. Our work is relevant for researchers and practitioners interested in detecting exceptional transition behavior in sequential data.

References

1. Lemmerich, F., Becker, M., Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Mining subgroups with exceptional transition behavior. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), http://www.kdd.org/kdd2016/papers/files/Paper_185.pdf

Harmony Assumptions: Extending Probability Theory for Information Retrieval

Thomas Roelleke

Queen Mary University of London

In many applications, independence of event occurrences is assumed, even if there is evidence for dependence. Capturing dependence leads to complex models, and even if the complex models were superior, they fail to beat the simplicity and scalability of the independence assumption. Therefore, many models assume independence and apply heuristics to improve results. Theoretical explanations of the heuristics are seldom given or generalisable.

[1] reports that some of these heuristics can be explained as encoding dependence in an exponent based on the generalised harmonic sum. Unlike independence, where the probability of subsequent occurrences of an event is the product of the single event probability, harmony is based on a product with decaying exponent.

For independence, the sequence probability is $p^{1+1+\dots+1} = p^n$. For harmony, the probability is $p^{1+1/2+\dots+1/n} \approx p^{1+\log(n)}$. The generalised harmonic sum is the exponent of p , and this leads to a spectrum of *harmony assumptions*. We will discuss that settings of the term frequency (TF) in IR correspond to harmony assumptions. We will focus on four settings of the TF:

$$\text{TF}(t, d) := \begin{cases} \text{tf}_d & \text{total TF: corresponds to assuming independence} \\ \sqrt{\text{tf}_d + 1} - 1 & \text{sqrt TF: middle between total TF and log-TF} \\ \log(\text{tf}_d + 1) & \text{log-TF: assumes a form for harmony} \\ \text{tf}_d / (\text{tf}_d + K_d) & \text{BM25 TF: assumes a strong form of harmony} \end{cases}$$

[1] shows series-based explanations of the TF settings, and these lead to new insights regarding the relationships between IR and probability theory. From an IR point of view exciting is the finding that the BM25-TF is the harmonic sum of Gaussian sums.

$$\frac{\text{tf}_d}{\text{tf}_d + 1} = \frac{1}{2} \cdot \left[1 + \frac{1}{1+2} + \dots + \frac{1}{1+2+\dots+\text{tf}_d} \right]$$

This finding provides a probabilistic interpretation of the BM25-TF quantification.

An experimental study for IR and social media investigates assumptions that explain the dependence between term occurrences. Interestingly, the assumption sqrt-harmony, i.e. the middle between the total-TF and log-TF, is on average a better assumption than independence or the strong harmony assumptions corresponding to log-TF and BM25-TF. The potential impact of harmony assumptions lies beyond IR, since many scientific disciplines and applications rely on probability theory and apply heuristics to compensate the independence assumption. Given the concept of harmony assumptions, the dependence between multiple occurrences of an event can be reflected in an intuitive and effective way.

References

1. Thomas Roelleke, Andreas Kaltenbrunner, and Ricardo A. Baeza-Yates. Harmony assumptions in information retrieval and social networks. *Comput. J.*, 58(11):2982–2999, 2015.

Comparing contextual and non-contextual features in ANNs for movie rating prediction

Ghulam Mustafa and Ingo Frommholz

Institute for Research in Applicable Computing
University of Bedfordshire, Luton, UK
ghulam.mustafa4@study.beds.ac.uk, ifrommholz@acm.org

Abstract. Contextual recommendation goes beyond traditional models by incorporating additional information. Context aware recommender systems (CARs) correspond to not only the user’s preference profile but also consider the given situation and context. However, the selection and incorporation of optimal contextual features in context aware recommender systems is always challenging. In this paper we evaluate different representations (feature sets) from the given dataset (LDOS-CoMoDa) for contextual recommendations, in particular looking into movie rating prediction as a subproblem of recommendation. We further cross-compare these representations to select useful and relevant features and their combination. We also compare the performance of standard matrix factorization to Artificial Neural Networks (ANNs) in CARs. Our evaluation shows that dynamic, contextual features are dominant compared to non-contextual ones for the given task in the given data set. We also show that ANNs slightly outperform matrix factorization approaches typically used in CARs.

Keywords: Artificial Neural Networks, Feature Selection, Rating Prediction, Context-aware Recommender Systems, Matrix Factorization

1 Introduction

Context aware recommender systems (CARs) incorporate additional contextual information into recommender systems and have emerged as one of the hottest topics in the domain of recommender systems [2, 18]. Traditionally recommender systems focus on recommending the most relevant items to the users or the most appropriate users to the items [1]. While traditional recommendation approaches have performed well in many applications [10], in a number of other applications and contexts, such as location and time based service recommender systems and travel recommendations, it may not be sufficient to consider only users and items [21]. It is also important to incorporate additional contextual information into the recommendation process [10]. A context can be defined as a dynamic set of factors that further describe the state of a user at the moment of user’s experience [6]. Nowadays, CARs have become very popular for many applications such as movie, music and mobile recommendations, services for learning, travel

and tourism, shopping assistance and multimedia [5, 11]. Most CAR approaches assume the contextual information does not change significantly and remains static, but some dynamic contextualization approaches have been proposed.

In CARs some additional contextual information is available to influence the rating behaviour. A context in this case can be defined as a set $c \in C$, e.g. $c_1 = \{happy, sad, ..\}$, the time at which the movie was watched e.g. $c_2 = \{Morning, Afternoon, \dots\}$ or the location $c_3 = \{Home, Public, \dots\}$. In this case multiple contexts C_1, \dots, C_m are available besides users U and items I , so the function y to estimate the rating R can be expressed as $y : U \times I \times C_1 \times \dots \times C_m \rightarrow R$.

Instead of providing the user with a recommendation decision, in this work we focus on an important sub-problem — the prediction of the ratings (e.g., 1 to 5 stars) a user might give a certain item. In a later step this prediction can be used for recommendation, for instance by recommending items with a predicted rating of 5 stars. To predict ratings, machine learning algorithms are reported in the literature to develop models and find patterns based on training data. Some of the well-known model based techniques are clustering, associating rules, matrix factorization, restricted Boltzmann machines and others. In context aware recommender systems, the selection of the appropriate context feature remains a persisting challenge [15]. In CARs, using too many context features may result in low accuracy and high dimensionality in the process of recommendation. Recommendation algorithms usually depend on the assumption that the features selected in advance will result in better accuracy [23].

To aim of our study is to gain more insights to aid the feature selection process. We investigate different feature sets (which we call *representations*) and their performance either as single representation or combined. To conduct our studies we use LDOS-CoMoDa, which is the most prominent collection for contextual movie recommendation. This is a very specific collection for the evaluation of CARs as it contains *dynamic contextual features* like location, mood, etc. Previous work has shown that applying dynamic features leads to highly accurate results. However, said previous work has only considered dynamic contextual features, but did not look at other available non-contextual ones (like gender, movie type, etc). Hence one aim of this study is to check the performance of non-contextual features, either alone or combined with contextual ones. Furthermore we show that utilising Artificial Neural Networks (ANNs) instead of matrix factorization, which is prominent in CARs, improves the performance of the rating categorization.

The remainder of this paper is structured as follows. In the next section we discuss some related work in context recommender systems. Subsequently in Section 4, we present our ANN-based approach to predict the ratings and some information about the data set used. We also introduce the contextual and non-contextual representations applied in our work. Furthermore, we present and discuss results of our experiments combining different representations with ANNs in Section 4.2, before we conclude.

2 Related Work

Context aware recommender systems have become very popular in many areas such as movies, music, mobile recommendations, services for learning, travel and tourism, shopping assistance and multimedia [18]. Feature selection in context aware recommender systems is always a challenging task. Since all the features and contexts do not contribute equally to generating valuable recommendations, it is very important to analyse the contextual features to choose the best ones. A number of studies have focused on the selection of contextual features [23, 20]. Different approaches of context aware recommender systems can be categorized by the contextual factors they are considering [16]. Many approaches assume that the contextual information does not change significantly and remains static. This assumption is made in most of the cases, while some recent research has been proposed for dynamic contextualization [9]. Recent work on CARs has focused on developing the models by integrating the contextual information with the user/item relations and models the user and item as well as context interactions [19]. To date, different approaches have been proposed under different categories of CARs including Hybrid Recommender [3], Tensor Factorization and Factorization Machine (FM).

In CARs researchers also suggest the incorporation of meta data such as user or item attributes into the prediction, however meta data normally yields only small improvements over the strong baseline methods that are used for the prediction of ratings [22].

In contextual recommender systems, machine learning algorithms are used to develop models and find patterns based on training data. Most of models are based on using a cluster technique for identification of a user based on test set. Some of the well-known model based techniques are Clustering, Associating rules, Matrix Factorization, Restricted Boltzmann Machines and others [24, 8, 13, 12]. In our approach, we are using ANNs, which haven't been used in detail for contextual recommendations we are dealing with yet.

3 Contextual Recommendations with ANNs

In order to compare the different contextual features, we introduce our ANN-based approach which is composed of a three layers architecture as illustrated in Fig. 1, consisting of an input, hidden and output layer. The input layer is composed of the 6 representations, which are provided as input to ANNs to predict the output y that represents the ratings from 1 to 5. These representations are manually formed based on the nature of the different contextual attributes and explained at the end of this section. The different representations are also combined as input, for example Dynamic representation is combined with the Category and User to find a better match in terms of accuracy. Each of the representations and their combinations are evaluated against the target data, which is the ratings data that comes from the users. The user rating (1-5) is

transformed into binary rating as explained in the following section, so that the ANNs can be trained. The optimal set of representations of the context features will be identified and recommended based on the accuracy that comes from the ANNs for each input. A brief description of the different representations with respect to the list of features is given in the Table 1.

In order to train ANNs on the different manually formed representations, we normalize the contextual features. This results in better accuracy for different features and their combinations. The hidden layer, a feed-forward multi-layer perception neural network, is used to map the input into the output binary classes y .

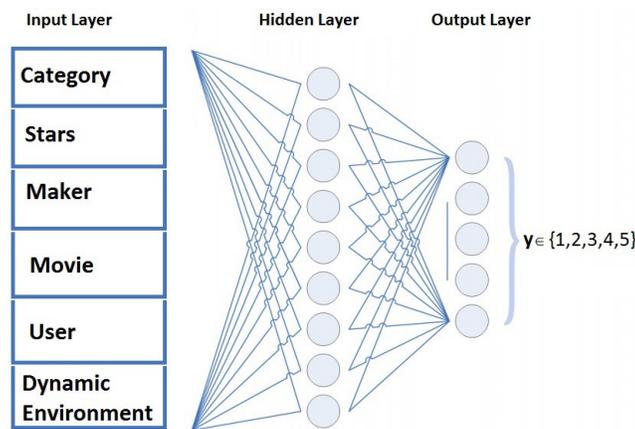


Fig. 1. Architectue of the proposed ANN approach

We pre-process the data to train a model using neural networks. The features available in the dataset are a dynamic set of features and static features. The different features available in the dataset are manually categorized into the 6 representations based on the type of contextual attributes as described in Table 1. Each of the representations is evaluated against the target data (user ratings) in our experiments. Since ANNs are binary classifiers, the target data is converted into binary representations for comparison and evaluation. To turn each of the 5 classes into a binary classification decision, each of the 5 possible ratings is compared to the rest as a yes/no decision (e.g. “Class 1 / not Class 1” to decide if the rating was 1 or not), resulting in 5 classes Class 1 to Class 5. In comparison, matrix factorization considers the ratings provided by the users for the items to map the users and the items in a joint latent feature space [4]. Different representations are also combined to find a better match in terms of accuracy, with less error rate. After evaluating the different representations and their combinations, the optimal feature set combination with highest accuracy will be considered for the recommendation process. Note that an item might be

classified into more than one of the above classes (e.g., the ANN may predict 1 star and 4 stars based on the single binary decisions). In this case, our policy is to select the highest rating prediction. The prediction of the ratings for items will also allow to rank items. Since the other approaches such as probabilistic neural networks are slower than multilayer perceptron networks and require more memory space to store the model, they are not better options at this stage.

Table 1. Representations and Features from LDOS-Comoda dataset

Representation	List of Features
Category	g1, g2, g3 (Genres of the movie)
Stars	a1, a2, a3 (Actors of the movie)
Maker	dir, budget
Movie	movie language, movie year, movie country
User	age, gender, city, country
Dynamic Environment	time, day-type, season, location, weather, season, dominant emotions, end emotion, mood, physical, decision, interaction

Following the representations given in the Table 1, the contextual features are distributed among 6 representations. The *Category* representation consist of movie genres which shows each movie is presented by three genres. The representation of *Stars* consist of the cast of movies, whereas the *Maker* representation contains information about the director of the movies as well as the their budget. The representation *Movie* consist of information about the movie country, movie language and movie year. The representation *User* consists of the static information of users including age, gender, city and country of the user. The *Dynamic Environment* representation contains dynamic variables such as time, day type, season, location, weather, social, dominant emotions, end emotions, mood, physical, decision and interaction. Different representations with the associated contextual information from the LDOS Comoda dataset are shown in Table 2. Once the different representations are identified, a neural network is trained to compare every single representation and combinations thereof with the target data to evaluate the performance and accuracy of the different context features. The optimal set of representations of the context features will be identified and recommended based on the experiments. Further details are provided in the next section.

4 Evaluation

In this section we briefly describe the dataset and the method used for our experiments. First of all we examine the dataset to find which information can be used as potential context from it. Based on the structure of the dataset we define a method how different representation can be formed based on the nature of the contextual features.

4.1 Dataset

The chosen dataset LDOS-CoMoDa¹ consists of 4381 movies which are rated by 121 users. The number of ratings available in this dataset are 2296 and the maximum number of ratings provided by a single user is 220; the minimum number of ratings is 1. The dataset consist of 12 contextual variables in addition to static user information. The basic statistics are given in Table 2.

Table 2. Basic Statistics of LDos-Comoda

Users/Items	121/4381	Ratings	2296
Rating scales	1-5	Context factors	12
User attributes	4	Item attributes	7

In order to evaluate the performance of different representations using binary classification, the true positive rate vs. false positive rate are more helpful than other predictive accuracy matrices [17].

4.2 Results and Discussion

Results In this section the different representations derived from the given contextual variables in the LDOS-Comoda dataset are evaluated, presented and discussed. The work presented in [15] on detecting the relevant context in movie recommender systems provides the relevance and irrelevance of contextual variables. However, we can categorize the contextual variables into the 6 different representations discussed above and cross-compare the representations as well as their combinations to find successful sets of contextual representations. In order to train the neural network on the chosen dataset, the data is preprocessed and normalized in the first stage. The rating data is transformed into a binary form as the neural network performs better using binary classifications. Different features available in the dataset are then normalized with respect to the number of available context features. Then, the ANN is trained using the method described in Section 3 and the samples are divided among the training data, selection data and the validation data. The statistics from the ANN samples division are giving in the Table 3. The number of the samples used by the Neural Network for training purpose is 1608 (70%), 344 for the selection purpose and 344 for the validation. The cross entropy during the training stage of ANN is measured as 1.4213 which shows a small fraction of error occurs during the training stage. The error percentage in the training stage is 3.17% which shows a small fraction of samples are mis-classified during the training stage. Similarly, the selection stage of the Neural Network utilizes 244 (15%) samples with the cross entropy 3.85 and error percentage 2.03. The validation stage also utilizes 344 samples (15%) with cross entropy 3.87 and the percentage of error at 3.19. We have also

¹ <http://www.ldos.si/comoda.html>

tried the combinations of all representations and observed the higher error rate of 62.20% which shows that it is not an ideal condition to use the features from all representations. A full intersection of the all six representations is not better matched, however, a combination given in Table 4 performed at the rate of 80% which shows the intersection of the Category, User and Dynamic can perform better in the scenario.

Table 3. Sampling from ANN

Entire dataset size	No. of Samples	Cross-Entropy Measure	% Error
Training dataset size	1608	1.4213	3.17
Selection dataset size	344	3.85	2.03
Validation dataset size	344	3.87	3.19

Using the method described in Section 3, we cross-compared the different representations with the target data to find the relevant representation which is a set of features. Features are cross-compared one by one by training the neural network which learns over 2296 samples (70% for training, 15% for testing and 15% for validation). The results from the experiments shows the performance of the Dynamic Environments representation performs better than the other representations Maker, Category, User, Movie and Stars. The performance of the Dynamic Environment representation remains above the threshold line when the binary classes are used to obtain the performance. The representations other than Dynamic Environment struggle with the errors and shown inferior performance, so the set of the features given as part of Dynamic Environment are a good set of features that can be used potentially for generating the recommendations. So the recommended stand alone representation is the set of features given in Dynamic Environment.

As we can see in the Table 4, the context features available in the Dynamic Environment representation performed better while the other representations struggle with respect to the performance and errors. So the Dynamic Environment representation is picked as the single optimal set of features. We also tried combinations of Dynamic Environment with other representations such as Category, Makers, Stars and User Statics to study combinations of representations. The comparison of combinations given in Figure 2 shows that the performance of the Dynamic Environment is not improved when combining this representation with others; the representations do not seem to complement each other. This means Dynamic Environment is indeed the dominant representation in the LDOS-CoMoDa collection.

We use the results reported in [14], using matrix factorization (MF), as baseline to compare the performance of our ANN approach since it utilizes the contextual attributes which are part of the Dynamic Environment in our method. The results comparison in Fig. 3 between the contextual attributes in the baseline method on the one hand and the ANNs on the other hand shows that contextual attributes performed better with ANN.

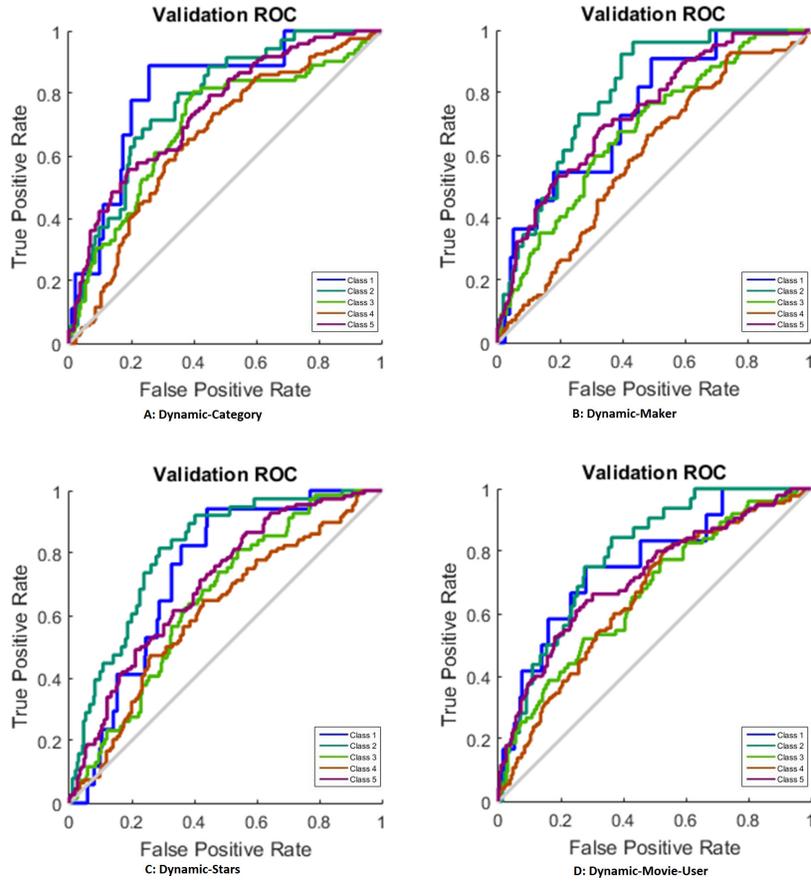


Fig. 2. True Positive Rate vs. False Positive Rate for combination of Dynamic Environment representation with other representations

Table 4. Performance of different features from ANN

Representation	Performance (Accuracy)	MF
Dynamic	97.12	96.9
Category	80.68	Not Reported
Makers	65.8	Not Reported
Where	66.58	Not Reported
User	64.9	Not Reported
Category + User + Dynamic	80.81	Not Reported

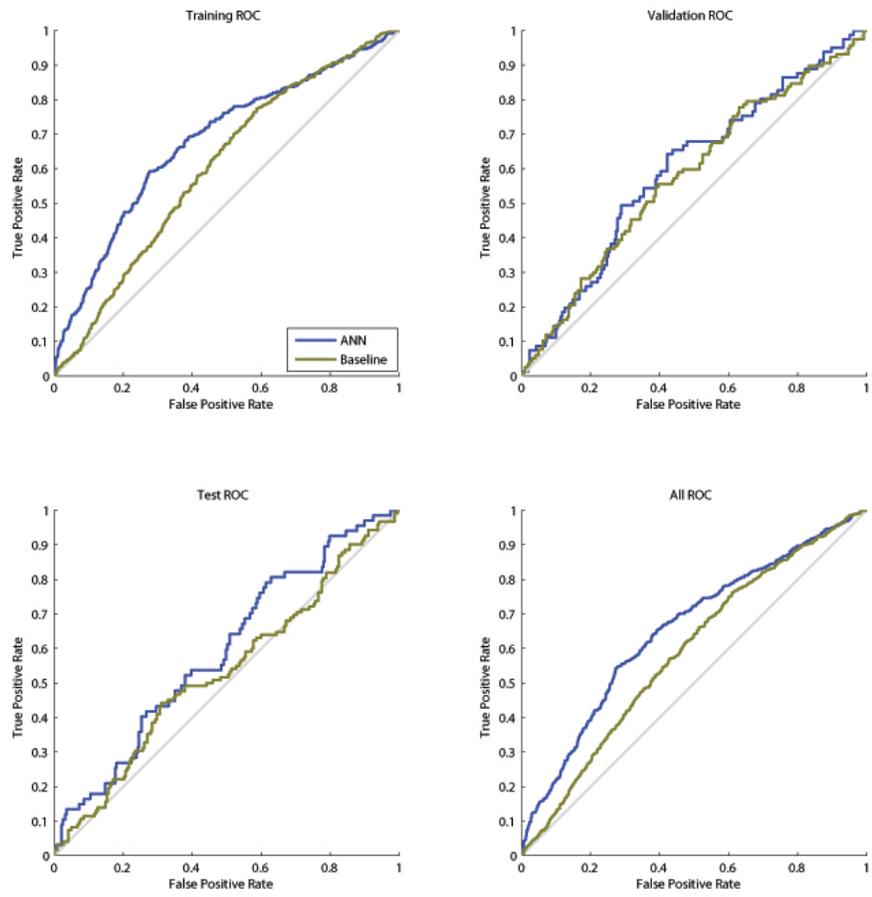


Fig. 3. Comparison of Contextual variables from Dynamic Environments with Baseline

The performance of the ANN is also evaluated by computing the Cross-Entropy which helps to evaluate the performance of three different stages of ANN (Train, Validation and Test) against the best performance. The results presented in the Fig. 4 shows that the performance of ANN remains better for all three stages when the ANN is trained for 22 epochs. In ANNs, an epoch is used to present the set of training vectors to the network for the calculation of new weights. The best performance is achieved in validation, as can be seen in the circle and gradient line in the figure, which means the performance is deemed acceptable according to the literature.

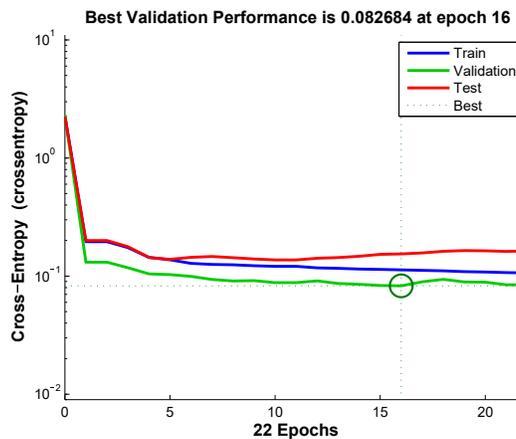


Fig. 4. Performance of ANN during Train, Validation and Test stages

Discussion The results have shown that contextual Dynamic Environment features by far outperform the static non-contextual features in the chosen LDOS-CoMoDa collection when it comes to rating prediction. The results also show that applying ANNs instead of matrix factorization improves the rating prediction accuracy even further when using the Dynamic Environment features. It confirms the important role of contextual features for CARs and the rather inferior role non-contextual features play, at least in the given data set. ANNs are indeed a very effective method for rating prediction, which is crucial for context-based recommendation.

5 Conclusion and Future Work

In this paper we introduced ANNs for rating prediction in contextual recommendations. We presented how to form different representations from a chosen

dataset LDOS-CoMoDa. Different representations are cross-compared and concluded that the Dynamic Environment context features performed best when applied alone, also outperforming the chosen matrix factorization baseline method. We further cross-compared combinations of the Dynamic Environment with other representations and observed that they do not perform well and are even not able to further complement the dynamic features, at least not with the combinations of the different representations.

The LDOS-CoMoDa dataset is an interesting data set when it comes to providing a rich set of dynamic contextual features. The dominance of such features for the given rating prediction task is remarkable. In the future we will look into similar data sets and investigate the role of dynamic contextual features compared to static, non-contextual ones. In this respect, we will also check if there is still a way to combine non-contextual features with dynamic, contextual ones, given that other data sets do not possess a dominant feature set like we find with LDOS-CoMoDa. One potential idea is borrowed from the principle of polyrepresentation [7], which is also a reason why we called feature sets *representations* in this work. If documents are recommended by different classifiers using different representations (feature sets), we would expect that the set of documents recommended by all classifiers exhibits a high precision. This would also give rise to a more interactive approach to recommendation, for instance by presenting to the user those recommendations first that are confirmed by different representations and let the user decide which set of recommendations to visit next (for instance those that match the current mood vs. those that match other features like age, location or genre). Whether we can actually observe something ‘polyrepresentation-like’ in machine learning based recommendation is subject to further investigation.

References

1. Nana Yaw Asabere. Towards a viewpoint of context-aware recommender systems (CARS) and services. *International Journal of Computer Science and Telecommunications*, 4(1), 2013.
2. Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breiting. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 2015.
3. Claudio Biancalana, Fabio Gasparetti, Alessandro Micarelli, Alfonso Miola, and Giuseppe Sansonetti. Context-aware movie recommendation based on signal processing and machine learning. *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation - CAMRa '11*, pages 5–10, 2011.
4. Maunendra Sankar Desarkar and Sudeshna Sarkar. Rating prediction using preference relations based matrix factorization. In *UMAP Workshops*, 2012.
5. Negar Hariri and Robin Burke. Context-aware music recommendation based on latent topic sequential patterns. In *Proceedings RecSys 2012*, pages 131–138, 2012.
6. Balázs Hidasi and Domonkos Tikk. Fast ALS-based tensor factorization for context-aware recommendation from implicit feedback. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 67–82, 2012.

7. Peter Ingwersen and Kalvero Järvelin. *The turn: integration of information seeking and retrieval in context*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.
8. Salil Kanetkar, Akshay Nayak, Sridhar Swamy, and Gresha Bhatia. Web-based personalized hybrid book recommendation system. In *International Conference on Advances in Engineering and Technology Research*, pages 1–5, 2014.
9. Pavlos Kefalas, Panagiotis Symeonidis, and Yannis Manolopoulos. New perspectives for recommendations in location-based social networks. *Proceedings MEDES 2013*, pages 1–8, 2013.
10. Fabrício D. A. Lemos, Rafael A. F. Carmo, Windson Viana, and Rossana M. C. Andrade. Towards a context-aware photo recommender system. *CEUR Workshop Proceedings*, 889, 2012.
11. Asher Levi, Osnat Ossi Mokryn, Christophe Diot, and Nina Taft. Finding a needle in a haystack of reviews: Cold start context-based hotel recommender system. In *Proceedings RecSys 2012*, pages 115–122, 2012.
12. Jiankou Li and Wei Zhang. Conditional Restricted Boltzmann Machines for cold start recommendations. *arXiv preprint arXiv:1408.0096*, 2014.
13. Ziqi Liu, Yu-Xiang Wang, and Alexander Smola. Fast differentially private matrix factorization. In *Proceedings RecSys 2015*, pages 171–178, 2015.
14. Ante Odić, Marko Tkalčić, Jurij F. Tasić, and Andrej Košir. Relevant context in a movie recommender system: Users’ opinion vs. statistical detection. *CEUR Workshop Proceedings*, 889, 2012.
15. Ante Odić, Marko Tkalčić, Jurij F. Tasić, and Andrej Košir. Predicting and detecting the relevant contextual information in a movie-recommender system. *Interacting with Computers*, 25(1):74–90, 2013.
16. Simon Renaud-Deputter, Tengke Xiong, and Shengrui Wang. Combining collaborative filtering and clustering for implicit recommender system. In *27th International Conference on Advanced Information Networking and Applications (AINA 2013)*, pages 748–755. IEEE, March 2013.
17. Antonio J. Serrano, Martin José D. Soria, Emilio, Rafael Magdalena, and Juan Gómez. Feature selection using ROC curves on classification problems. In *International Joint Conference on Neural Networks (IJCNN 2010)*, pages 1–6, 2010.
18. Gediminas Adomavicius Tuzhilin and Alexander. Advances in Collaborative Filtering. In *Recommender Systems Handbook*, pages 217–253. 2011.
19. Shu Lin Wang and Chun Yi Wu. Application of context-aware and personalized recommendation to implement an adaptive ubiquitous learning system. *Expert Systems with Applications*, 38(9):10831–10838, 2011.
20. Hao Wu, Kun Yue, Xiaoxin Liu, Yijian Pei, and Bo Li. Context-aware recommendation via graph-based contextual modeling and postfiltering. *International Journal of Distributed Sensor Networks*, 2015:7–9, 2015.
21. Liu Xin and Karl Aberer. Personalized Point-of-Interest Recommendation by Mining Users Preference Transition. *Proceedings CIKM 2013*, pages 733–738, 2013.
22. Mao Ye, Peifeng Yin, and Wang-Chien Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 458–461, 2010.
23. Saloua Zammali, Khedija Arour, and Amel Bouzeghoub. A Context Features Selecting and Weighting Methods for Context-Aware Recommendation. *2015 IEEE 39th Annual Computer Software and Applications Conference*, 5:575–584, 2015.
24. Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis Categories and Subject Descriptors. In *Proceedings SIGIR 2014*, pages 83–92, 2014.

Author Index

- Aamodt, Agnar, 153
Abedjan, Ziawasch, 225
Althoff, Klaus-Dieter, 249, 291
Amin, Kareem, 257
Anna Schett, Maria, 103
Archambeau, Cédric, 2
Arnu, David, 283
Assent, Ira, 234
Atzmueller, Martin, 121, 157, 283
Ayzenshtadt, Viktor, 291
- Bach, Kerstin, 153
Barrat, Alain, 121
Barros, Alistair, 29
Bauckhage, Christian, 311, 347
Baumeister, Joachim, 145, 165, 177
Becker, Martin, 359
Belli, Volker, 145, 165, 177
Bergamini, Elisabetta, 51
Bergmann, Ralph, 27, 43, 54
Bizer, Christian, 7
Bothe, Sebastian, 335
Braun, Daniel, 235
Breß, Sebastian, 233
Bruns, Werner, 136
Buczowski, Przemyslaw, 122
Butzin, Björn, 83
- Campello, Ricardo, 234
Campos, Guilherme, 234
Castro Fernandez, Raul, 212
Cattuto, Ciro, 121
Christen, Victor, 227
Conrad, Stefan, 235
Costa, Paolo, 212
Cvejoski, Kostadin, 347
- Dörsch, Tobias, 186
Düver, Jonas, 74
Dees, Jonathan, 9
Dengel, Andreas, 291
Drumond, Lucas, 89
Duong-Trung, Nghia, 89
- Eric, Westenberger, 9
Erik Gundersen, Odd, 153
Eschenbach, Sebastian, 136
- F. Ilyas, Ihab, 225
Fürsich, Michael, 155
Faerber, Franz, 9
Fier, Fabian, 194
Freytag, Johann-Christoph, 194
Frommholz, Ingo, 361
Fuchs, Florian, 54
- Funke, Henning, 233
Furth, Sebastian, 145, 165, 177
- Garg, Pranav, 156
Gauraha, Niharika, 53
Godde, Christian, 299
Groß, Anika, 200, 227
Grumbach, Lisa, 43
Grunert, Hannes, 83
- Höfer, Eva, 194
Habich, Dirk, 30
Hartz, Marc, 9
Helbing, Dirk, 12
Helic, Denis, 359
Henkel, Wolfram, 249
Heuer, Andreas, 83
Hildebrandt, Juliana, 30
Hinneburg, Alexander, 265
Hirzel, Ann-Kathrin, 29
Horvath, Tamas, 42, 335
Hotho, Andreas, 359
Houle, Michael E., 234
Huellermeier, Eyke, 323
- Illig, Jens, 121
- Jähnichen, Patrick, 101
Jarle Mork, Paul, 153
Jerzak, Zbigniew, 6
Jugel, Uwe, 6
- K Maier, Ronald, 103
Kübler, Eric, 123
Kasparick, Martin, 83
Kettunen, Kimmo, 124
Kibanov, Mark, 121
Kiefer, Cornelia, 62
Klettke, Meike, 26
Kloft, Marius, 101
Knöll, Daniel, 157
Kohlhase, Andrea, 155
Koliouisis, Alexandros, 212
Krestel, Ralf, 299
Kriegel, Hans-Peter, 226
Kuhn, Norbert, 43
Kuokkala, Juha, 124
- Lazaridou, Konstantina, 299
Legler, Alexander, 177
Lehner, Wolfgang, 30
Lemmerich, Florian, 359
Lenz, Richard, 206
Leser, Ulf, 5, 213
Leyer, Michael, 29

- Lommatzsch, Andreas, 52, 74, 186
Looz, von Moritz, 51
Luis Wiegandt, David, 213
- Mäkelä, Eetu, 124
Müller, Gilbert, 27
Madhusudan, P., 156
Maier, Edith, 136
Makary, Mireille, 175
Markl, Volker, 233
Meyer-Wegener, Klaus, 206
Meyerhenke, Henning, 51
Micenková, Barbora, 234
Mikyas, Ada, 291
Minor, Mirjam, 123
Moormann, Jürgen, 29
Morcos, John, 225
Mu, Mu, 52
Mustafa, Ghulam, 361
- Neider, Daniel, 156
Nentwig, Markus, 200
Neumann, Thomas, 11
Niemi, Jyrki, 124
- Oakes, Michael, 175
Ojeda, Cesar, 347
Ouzzani, Mourad, 225
- Pölit, Christian, 14
Papotti, Paolo, 225
Parui, Swapan, 53
Pfister, Maximilian, 54
Piatkowski, Nico, 13
Pietzuch, Peter, 212
Ploch, Danuta, 74
- Rahm, Erhard, 200, 227
Reimer, Ulrich, 136
Reuss, Pascal, 249
Rieder, Constantin, 157
Rietzke, Eric, 43
Roelleke, Thomas, 360
Rosner, Frank, 265
Roth, Dan, 156
Roth, Lea, 145
Ruokolainen, Teemu, 124
Russell, Nick, 29
- Sander, Jörg, 234
Saqib Bukhari, Syed, 291
Sattler, Kai-Uwe, 10
Schäfer, Dirk, 323
- Scheel, Christian, 271
Scherer, Klaus-Peter, 157
Scherzinger, Steffi, 26
Schilling, Nicolas, 89
Schmid, Ute, 28
Schmidt, Andreas, 283
Schmidt-Thieme, Lars, 89
Schmitz, Claudia, 271
Scholz, Christoph, 121
Schubert, Erich, 226, 234
Schwab, Peter, 206
Schwarzer, Malte, 74
Schwinn, Markus, 43
Sifa, Rafet, 347
Singer, Philipp, 359
Singhof, Michael, 235
Sobkowicz, Antoni, 122
Sombach, Stephanie, 26
Störl, Uta, 26
Starlinger, Johannes, 213
Staudt, Christian, 51
Stede, Manfred, 1
Stonebraker, Michael, 225
Striffler, Albrecht, 177
Strohmaier, Markus, 359
Stumme, Gerd, 121
Szczepanski, Tomasz, 153
- Teubner, Jens, 233
Thalmann, Stefan, 103
Thiede, Felix, 111
Timm, Felix, 111
Timmermann, Dirk, 83
- Wagner, Dorothea, 3
Wahl, Andreas, 206
Weidlich, Matthias, 212
Weiler, Michael, 226
Welke, Pascal, 42
Wenzel, Florian, 101
Wiech, Katharina, 26
Wieczorek, Sebastian, 4
William De Luca, Ernesto, 247, 271
Wolf, Alexander, 212
Wrobel, Stefan, 42
- Yamout, Fadi, 175
Yuan, Jing, 52
- Zasada, Andrea, 111
Zeller, Christina, 28
Zimek, Arthur, 234