Jahresbericht 2012

Fachgebiet
Informationssysteme

Prof. Dr. Felix Naumann

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Inhaltsverzeichnis

1 Personelle Zusammensetzung

Leiter des Fachgebiets

- Prof. Dr. Felix Naumann

Assistentin der Fachgruppe

- Katrin Heinrich

Senior Researcher

- Dr. Gjergji Kasneci

Wissenschaftliche Mitarbeiter

- Alexander Albrecht
- Christoph Böhm
- Uwe Draisbach
- Toni Grütze
- Arvid Heise
- Dustin Lange

PhD-Stipendiaten und Postdocs

- Ziawasch Abedjan
- Maximilian Jenders
- Anja Jentzsch
- Johannes Lorey
- Dr. Saeedeh Momtazi
- Tobias Vogel
- Zhe Zuo

2 Lehrveranstaltungen

2.1 Vorlesungen

Sommersemester 2012

- Datenbanksysteme I (Naumann, 4 SWS, Bachelor)
- Natural Language Processing (Momtazi, 2 SWS, Master)
- Data Mining and Probabilistic Reasoning (Kasneci, 2 SWS, Master)
- Information Integration (Naumann, 2 SWS, Master)
- Mainframe Computing Summit (Naumann, Blockkurs, Bachelor / Master, gemeinsam mit Prof. Polze)

Wintersemester 2012/2013

In diesem Semester befand sich Prof. Naumann im Forschungsfreisemester.

- Information Retrieval (Kasneci, 2 SWS, Master)

2.2 Seminare

Sommersemester 2012

- Beauty is our Business (2 SWS, Bachelor)
- Algorithms for Pattern Mining (4 SWS, Master)

2.3 Masterprojekte

Wintersemester 2012/2013

- FactScore: Global Relevance Scores for DBpedia Facts (Kasneci, Masterprojekt)
  Betreuer: Gjergji Kasneci, Ziawasch Abedjan
  Studenten: Steffan George, Matthias Kohnen, Philipp Langer, Tobias Metzke, Patrick Schulze
  Abstract: Knowledge bases have become ubiquitous assets in today's Web. They provide access to billions of statements about real-world entities derived from governmental, institutional, product-oriented, bibliographic, biochemical and different general-purpose datasets. The sheer amount of statements that can be retrieved for a given entity calls for ranking techniques that return the most salient statements as top results.
  In this paper we analyse and compare various ranking strategies; some of them synergetically combine complementary aspects such as frequency and inverse frequency with structural features, yet others rely on authority-based measures (e.g., PageRank) and Web-based co-occurrence statistics for entity pairs. A user-based evaluation of all approaches has been conducted on the popular DBpedia knowledge base with statistics derived from an indexed version of the Clueweb corpus.

3 Betreuung von Studierenden und Dissertationen

3.1 Betreuung von Bachelorprojekten und -arbeiten

3.1.1 Bachelorprojekte (abgeschlossen in 2012)

- A Cloud Platform for On-Demand Access to Open Data
  Betreuer: Naumann, Böhm, Lorey
  Partner: Fluid Operations AG
  Studenten: Christian Godde, Robert Lehmann, Magdalena Noffke, Dominic Petrick, Benjamin Reissaus
  Abstract: To leverage the benefits of open data, fluid Operations is developing a cloud platform that will simplify and considerably accelerate the entire data utilization process, namely support in data discovery, self-service deployment of data sources, scalable data storage, data integration, data curation, as well as analytics and custom application development on top of the data. In the context of this cloud platform, the goal of the project is to create a repository of open data sources. Concrete challenges and tasks include the population of the repository from public sources, the extraction and generation of metadata about data sets, the generation of statistics about data sets, link discovery across data sources, and data cleansing to deal with imperfections in the data, especially when integrating heterogeneous sources.
- CelebDB: Harvesting Celebrity Data
  Betreuer: Naumann, Momtazi
  Partner: Celebrity Performance GmbH
  Studenten: Julien Bergner, Fabian Eckert, Nicolas Fricke, Maria Neise, Kai-Adrian Rollmann, Robert Schäfer
  Abstract: The goal of the project is to populate the CelebDB database and present its data in a celebrity portal. The project partner provides a list of 5,000 celebrities of which only 500 celebrities are included in the database. The current version of the database is populated manually and it only contains factual information about celebrities. In the bachelor project, we want to expand the database automatically in order to cover more

celebrities and include more information about them

3.1.2 Laufende Bachelorprojekte (Abschluss in 2013)

- VIP 2.0: Celebrity Exploration
  Betreuer: Kasneci, Naumann, Grütze
  Partner: Celebrity Performance GmbH
  Studenten: Daniel Dummer, Johannes Eschrig, Manuel Hegner, Florian Moritz, Johannes Wolf, Martin Zabel
  Abstract: The goal of the project is to harvest web data to add connections between people and/or companies into a given database of German celebrities. The project partner provides a database with basic information about many well-known people in Germany. This version of the database is populated manually and it only contains factual information about celebrities. During the project this database will be expanded automatically in order to represent connections of the celebrities.

3.2 Betreuung von Masterarbeiten

3.2.1 Abgeschlossene Masterarbeiten (Abgabe 2012)

- David Wenzel: „Manuelle Duplikaterkennung mittels Crowdsourcing“
- Benjamin Emde: „Context-aware Recommendations in Social Networks“
- Fabian Lindenberg: „Generating Query Suggestions by Exploiting Latent Semantics in Query Logs“
- Maximilian Jenders: „Analyzing and Predicting Viral Tweets“
- Eyk Kny: „Erweiterung und Optimierung eines Graph-Clustering-Verfahrens“
- Michael Leben: „Email Classification with Contextual Information“
- MinhTuan Nguyen: „Summarizing Extract-Transform-Load Workflows“
- Sebastian Kölle: „Automatic Data Normalization Using Pattern-Based Repairs“

3.2.2 Laufende Masterarbeiten (Abgabe 2013)

- Tobias Rawald: „Iterative Data Cleansing“
- Thorsten Papenbrock: „Progressive Duplicate Detection“
- Florian Thomas: „Optimierung regelbasierter Duplikaterkennung“
- Sven Viehmeier: „Incremental Data Profiling“
- Thomas Kaske: „Automatische Generierung eines Doktorvater-Stammbaumes“

3.3 Betreuung von Dissertationen (intern, extern)

3.3.1 Abgeschlossene/ eingereichte Dissertationen in 2012

keine

3.3.2 Laufende Dissertationsprojekte

- Ziawasch Abedjan: „Data Mining for RDF Data“
- Alexander Albrecht: „Managing and Integrating ETL Processes“
- Jana Bauckmann: „Semi-automatic Integration of Life Sciences Data“ (eingereicht 02/2013)
- Christoph Böhm: „Profiling Heterogeneous Data“
- Uwe Draisbach: „Efficient Entity Resolution“
- Toni Grütze: „Web Data Extraction“
- Arvid Heise: „Parallel and Declarative Data Cleansing“
- Maximilian Jenders: „Microblog Analysis“
- Dustin Lange: „Effektive und effiziente Suche und Identifikation von Entitäten in Relationen“ (eingereicht 02/2013)
- Johannes Lorey: „Linked Open Data Services in the Cloud“
- Tobias Vogel: „Provisioning and Adaption of Data Quality Web Services“
- Zhe Zuo: „Relationship Extraction“

4 Bearbeitete Forschungsthemen

Die verschiedenen Forschungsaktivitäten des Lehrstuhls haben sich inhaltlich auf die folgenden Schwerpunkte konzentriert:

4.1 Data Cleansing

Forschungsprojekt: Data Quality as a Service (DAQS)
Betreuer: Tobias Vogel
Abstract: Datenreinigungsprozesse sind traditionell auf menschliche Experten angewiesen, die die komplexen Verfahren konfigurieren und die Programme bedienen. Sowohl Experten als auch die Programme sind ein großer Kostenfaktor und nur von größeren Organisationen einsetzbar. Data Quality as a Service (DAQS) ist ein Projekt, das für diese Problemstellung einen Webservice bereitstellt. Dabei werden Konfigurationsentscheidungen autonom durch den Service getroffen (z.B. die Erkennung von Attributklassen oder die Wahl von Partitionierungsschlüsseln). Damit eignet sich DAQS zum Datenreinigen auch für kleinere Organisationen oder für Ad-hoc-Reinigungsprojekte.

Forschungsprojekt: Annealing Standard
Team: Uwe Draisbach, Arvid Heise, Dustin Lange, Felix Naumann, Tobias Vogel

Abstract: Die Evaluierung von Duplikaterkennungssystemen benötigt normalerweise einen Goldstandard. Diese sind häufig aufwändig zu beschaffen, nicht repräsentativ oder vertraulich. Ein Annealing Standard ist eine wohldefinierte Menge an Record-Paaren, die manuell oder von vielen (maschinellen) Classifiern übereinstimmend klassifiziert wurden. Jeder weitere Classifier unterstützt die bestehende Klassifizierung der vorherigen Classifier oder leitet eine manuelle Überprüfung für einzelne Paare ein. Damit werden automatisch nur die für den Computer schwierig zu klassifizierenden Duplikatkandidaten manuell überprüft; die eindeutigeren Duplikat- und Nicht-Duplikatpaare werden vom Computer klassifiziert. Der Annealing Standard konvergiert somit gegen einen Goldstandard.

4.2 Data Profiling

Heterogene Datenquellen sind nicht nur von unterschiedlicher Qualität, sondern auch von unterschiedlicher Struktur. Um mit unbekannten Datenquellen arbeiten zu können, müssen diese daher zunächst auf ihre individuellen Eigenschaften (Metadaten) untersucht werden. Data Profiling ist ein automatisierter, analytischer Prozess, der Metadaten zu weitgehend unbekannten Echtdaten generiert. Der Prozess umfasst auch die Disziplinen Data Mining und Data Cleansing. Die gewonnen Metadaten beschreiben das Schema der Daten (Spaltennamen, Datentypen, etc.) und weitere wichtige Eigenschaften der Daten, wie etwa wiederkehrende Muster, Regeln oder Gütekriterien. Besonders relevant ist Data Profiling beispielsweise für wissenschaftliche Datenbanken, wie molekularbiologische oder astronomische Datenbanken. Auch wird es intensiv für vernetzte Datenquellen und Wissensbanken im World Wide Web und für das Web of Data eingesetzt.

Forschungsprojekt: Discovery of unique column combinations

Team: Ziawasch Abedjan, Arvid Heise, Anja Jentzsch, Jorge Quiane-Ruiz (QCRI), Felix Naumann

Abstract: Die Entdeckung von Spaltenkombinationen, die nur eindeutige Werte enthalten, ist für die Datenmodellierung, die Datenbankoptimierung und auch für die Datenintegration von großem Interesse. Schon Lösungen zu diesem Problem sind expoentieller Natur, da die die Anzahl der Spaltenkombinationen in $O(2^n)$ liegt. Insofern bieten sich parallelisierte, verteilte Verfahren an, etwa auf der Hadoop oder auf der Stratosphere Plattform.
Dieses Projekt wird in Kooperation mit dem Qatar Computing Research Institute (QCRI) durchgeführt.

4.3 Open Data

Im Allgemeinen werden unter dem Begriff Open Data jene Datensätze zusammengefasst, die frei verfüg- und nutzbar sind. Darunter fallen solche Informationsbestände, die von Regierungen, Regierungsbehörden oder privaten Unternehmen veröffentlicht werden, aber auch andere ohne Einschränkung bereitgestellte Dokumente wie wissenschaftliche Publikationen oder Wetteraufzeichnungen. Insbesondere das Untersuchen strukturierter und untereinander verknüpfter Datensätze, die als sogenannte RDF-Tripel veröffentlicht werden, steht dabei im Zentrum unserer Forschung. Im Rahmen unserer Projekte analysieren wir diese Datenbestände und den Zugriff darauf. Mithilfe der dabei gewonnen Erkenntnisse verbessern wir den Umfang, die Qualität und die Handhabung der enthaltenen Informationen.

4.3.1 Data Mining

Forschungsprojekt: Association Rule Mining on RDF Data
Betreuer: Ziawasch Abedjan
Abstract: Linked Open Data umfasst sehr viele und oft sehr große offene Datenmengen, die oft in der RDF Struktur vorzufinden sind. Die Heterogenität der vorhandenen Datenquellen erfordern jedoch unabwendbare Integrationsschritte bevor sie durch Anwendungen genutzt werden können. Eine vielversprechende und neue Methode um solche Daten zu untersuchen und aufzubereiten ist Association Rule Mining. Basierend auf eine auf Mining Konfigurationen Methodology entwickeln wir verschiedene Möglichkeiten RDF Daten zu analysieren und nützliche Metadaten zu generieren. Dabei entwickeln wir Methoden für Attributvorschläge, Datenkomplementierung, Ontology-Verbesserungen und Anfrageoptimierung. Diese Verfahren verbessern einerseits die Qualität der Daten und anderseits erlauben sie Nutzern mit Inkonsistenz in der Datenmenge umgehen zu können.

4.3.2 Data Provisioning

Forschungsprojekt: Linked Data as a Service
Betreuer: Johannes Lorey
Abstract: Es existiert eine Vielzahl an Projekten, die es sich zum Ziel gesetzt haben, verschiedenartige Informationen als sogenannte Linked Data bereitzustellen. Dabei wird das Publizieren solcher Daten durch eine Reihe von Standards und Konventionen formell klar geregelt. Andererseits ist das Verwenden dieser Daten oftmals mit großen Hindernissen verbunden. Während einige Datenquellen beispielsweise lediglich als herunterladbare Dateien zur Verfügung stehen, sind

andere nur als öffentlich zugänglicher SPARQL-Endpunkt erreichbar. Das Ziel dieses Forschungsprojekts ist der Aufbau einer Plattform für einen einfacheren Zugriff auf diese Datenmengen basierend auf einer skalierbaren Umgebung. In diesem Kontext untersuchen wir typische Zugriffsmuster auf Daten und entwickeln darauf basierend einen optimierten Zugriff auf vorhandene Ressourcen.

4.3.3 Data Matching

Forschungsprojekt: Linked Data Alignment
Betreuer: Christoph Böhm
Team: Gerard de Melo (ICSI Berkley), Nils Rethmeier, Felix Naumann, Gerhard Weikum (MPII)
Abstract: Das Linked Data Alignment (LINDA) Projekt beschäftigt sich mit dem Auffinden verschiedener Repräsentationen eines Reale-Welt-Objektes in RDF Daten aus dem sog. Web of Data. Die Schwierigkeit dieses Projektes liegt insbesondere in der Größe und Heterogenität der vorhandenen Datenquellen im Web. Ziel des Projektes war es unter den gegebenen Voraussetzungen Ansätze zu entwickeln, die nicht lediglich isolierte Entscheidungen für zwei Entitäten treffen, sondern die Gesamtheit der in den Daten zur Verfügung stehenden Beziehungen ausnutzen - d.h. sog. Joint Reasoning zu betreiben. So wurden drei Methoden mit unterschiedlichen Eigenschaften entworfen, implementiert und getestet. Methode 1 liefert ein vollständig konsistentes Ergebnis, hat aber Laufzeitdefizite. Methode 2 arbeitet deutlich schneller, verletzt aber u.U. die Konsistenz des Ergebnisses. Methode 3 ist skalierbar, denn sie lässt sich (wie Methode 2) auf Rechnerclustern ausführen und liefert ein konsistentes Ergebnis.

4.4 Web Mining

Today's Web contains ever more information about entities like companies, products, and persons. This information bare the potential for new applications that would not be possible with traditional information systems (i.e., database systems of limited scope). We apply different data mining and machine learning techniques to solve challenging problems for different domains. Due to the ever-growing size of the Web, a primary focus is set on the scalability of our approaches.

4.4.1 Disambiguation

This research area addresses the problem of entity disambiguation. Entity disambiguation is concerned with the process of identifying which entity is mentioned in a text, when there are several entities that could be referred to by the same mention.

Forschungsprojekt: Web Search Result Disambiguation
Betreuer: Toni Grütze
Team: Gjergji Kasneci, Zhe Zuo
Abstract: A wealth of useful information about people occurs in Web 2.0 platforms, such as Wikipedia, LinkedIn, Facebook, etc. Being human-generated, the information on these platforms is clean, focused, and already disambiguated. We aim at exploiting the above information to improve search results to queries about ambiguous person names.

Forschungsprojekt: Named Entity Linkage
Betreuer: Zhe Zuo
Team: Gjergji Kasneci, Toni Grütze
Abstract: Many entities that are contained in general-purpose and domain-specific knowledge bases have ambiguous names and are difficult to recognize and disambiguate in free text. The goal of this project is to automatically identify mentions of named entities in Web documents and link them to their knowledge base pendants, thus enabling a better analysis of the underlying documents.

4.4.2 Web-based Prediction and Recommendation

This research topic addresses the problem of predicting the preference of a user for items (e.g., products, news, ...) he has not yet seen or considered. Thus, it addresses the reasoning over reduced data without losing relevant information for users or items.

Forschungsprojekt: Predicting viral spread of information
Betreuer: Gjergji Kasneci
Team: Maximilian Jenders
Abstract: We address important questions concerning the spread of information in social networks. More specifically we are interested in predicting whether a post will become “viral”, i.e., will be frequently shared within the network? To answer these questions we conduct extensive analysis of a wide range of structural, information-, and sentiment-oriented features from the underlying social network.

Forschungsprojekt: Cross-Plattform Recommendation
Betreuer: Maximilian Jenders
Team: Gjergji Kasneci
Abstract: The wealth of information shared through social network platforms calls for techniques that can adequately filter topic-specific information in form of (topically) related posts and blogs. Our goal is to design algorithms that can reliably align posts across platforms. The problem is cast into a recommendation problem where for any selected post the user is presented with topically related posts from other networks.

Forschungsprojekt: Identifying temporally coherent topics from evolving sources
Betreuer: Toni Grütze
Team: Gjergji Kasneci
Abstract: Much of the information shared through social networks undergoes a temporal evolution, as posts are read, reedited, or reposted by users within a certain time frame. The aim of this project is to detect temporal as well as topical dependencies between information fragments and shared concepts. This would allow us to better predict uprising, interesting topics and dependencies between them.

4.5 Similarity Search

Similarity Search (Ähnlichkeitssuche) bezeichnet die Aufgabe, in einer Menge von Objekten diejenigen zu finden, die ausreichend ähnlich zu einem gegebenen Anfrageobjekt sind. Herkömmliche relationale Datenbanksysteme bieten nur Mittel zum effizienten Finden exakter Treffer zu einer gegebenen Anfrage. Enthält die Anfrage z.B. Tippfehler, fehlende oder vertauschte Attributwerte, könnten Algorithmen, die nur exakte Suche unterstützten, nicht alle relevanten Objekte finden.

Forschungsprojekt: Ähnlichkeitssuche auf Personendaten
Betreuer: Dustin Lange
Abstract: In einem Forschungsprojekt mit der SCHUFA Holding AG entwickelt unsere Gruppe neue Algorithmen zur effektiven und effizienten Ähnlichkeitssuche. Effektive Ähnlichkeitssuche wird erreicht durch die Definition eines geeignten Ähnlichkeitsmaßes, welches die Ähnlichkeit zweier Objekte berechnet. Zur effizienten Anfrageausführung werden Indexstrukturen benötigt, die ähnliche Attributwerten zur einer Anfrage berechnen, sodass die Anfrage so schnell wie möglich beantwortet werden kann. In unserem Projekt haben wir u.a. Algorithmen zur statischen und dynamischen Indexauswahl für Ähnlichkeitsanfragen entwickelt.

4.6 ETL

Extract-Transform-Load (ETL) Tools werden insbesondere bei der Erstellung, Wartung und Evolution von Data Warehouse Systemen verwendet. ETL Workflows befüllen diese Systeme mit Daten aus vielen unterschiedlichen Quellsystemen. Über die Zeit entstehen so viele hundert ETL Workflows, insbesondere da in heutigen Data Warehouse Szenarien kontinuierlich neue Quellsysteme und Anforderungen berücksichtigt werden müssen. Das Erstellen und Warten von ETL Workflows ist ein manueller Prozess, der viel Zeit kostet. In Kooperation mit der Firma InfoDyn AG erforschen wir daher Methoden für das bessere Verständnis komplexer ETL Workflows. So entwickeln wir beispielsweise Methoden für das Finden aussagefähiger Bezeichner für kryptische ETL-Schemata (Schema Decryption).

5 Auftragsforschung und sonstige Projekte

5.1 DAQS mit SAP

Forschungsprojekt: Business Object Data Services
Projektpartner: SAP Innovation Center Potsdam
Projektleiter: Mohammed AbuJarour
Projektteam: Mohammed AbuJarour, Tobias Vogel
Abstract: Relationale Eingangsdaten in Datenreinigungs- und Datenprofilingprozessen sind oft schwierig zu benutzen, weil es an Metadaten mangelt. So können z.B. Attribute ohne Titel auftreten oder dieser zumindest nicht maschinenlesbar sein. Dadurch fällt es beispielsweise schwer, geeignete Ähnlichkeitsmaße oder ein Mapping zu definieren. Dies erfordert dann ein manuelles Eingreifen in den Prozess. In diesem Projekt wurden die Ergebnisse des Klassifizierungsmoduls aus dem DAQS-Projekt (siehe Abschnitt 4.1 auf Seite 10) in die Business Object Data Services Suite prototypisch integriert, um den Automatisierungsgrad zu steigern und den Algorithmus anhand von Echtweltdaten zu evaluieren.

5.2 GovWILD

Forschungsprojekt: GovWILD - Government Web Integration for Linked Data
Projektpartner: IBM mittels eines Scalable Data Analytics Award
Projektleiter: Prof. Dr. Felix Naumann
Projektteam: Claudia Lehmann, Andrina Mascher, Christoph Böhm, Arvid Heise
Abstract: Viele demokratisch geführten Staaten veröffentlichen erhobene Daten allgemeinen Interesses, um die Transparenz zu erhöhen. Leider sind die verschiedenen Datenquellen so heterogen, dass eine einheitliche Anfrage an die Daten praktisch unmöglich ist. In GovWILD haben wir verschiedene Datenquellen der USA, von Deutschland und der EU zu einem einheitlichen Datensatz integriert. Auf diesem Datensatz ist es nun möglich umfangreiche Statistiken zu berechnen aber auch interessante Querverbindungen zwischen den involvierten Politikern und Firmen zu finden.

5.3 iDuDe

Forschungsprojekt: iDuDe - iPhone Adress Book Duplicate Detection
Projektpartner: CAS Software AG
Projektleiter: Felix Naumann
Projektteam: Uwe Draisbach, Dustin Lange, Tobias Vogel
Abstract: Adressbücher auf mobilen Geräten enthalten oft Duplikate von Personen oder Organisationen. Kontaktangaben (Telefonnummern, E-Mail-Adressen) können so über mehrere Einträge verstreut sein. Dadurch werden Konversationen (Telefonanrufe, Kurznachrichten, E-Mails), nicht zusammengeführt und es ist schwieriger, die Kontakteinträge zu pflegen, gerade bei geschäftlichen und privaten Kontakten. Ziel des Projekts war es, eine iPhone-Anwendung zu erstellen, die Duplikate im Adressbuch finden und zusammenführen kann. Die aktuelle Version findet sich im App Store unter https://itunes.apple.com/us/app/onecontact/id597402028.

5.4 METL

Projekt: METL - Managing and Integrating ETL Workflows
Projektpartner: InfoDyn AG
Projektleiter: Prof. Dr. Felix Naumann
Projektteam: Alexander Albrecht, Paul Möller, Minh Tuan Nguyen
Abstract: Das Erstellen und Warten von ETL Workflows ist ein manueller Prozess, der viel Zeit und Geld kostet. In Kooperation mit der Firma InfoDyn AG erforschen wir daher Methoden und Techniken für das bessere Verständnis komplexer ETL Workflows. So entwickeln wir beispielsweise Methoden für das Finden aussagefähiger Bezeichner für kryptische ETL-Schemata (Schema Decryption) und Techniken für die kompakte Darstellung komplexer ETL Workflows (ETL Workflow Abstraction).

5.5 Relate

Projekt: Entwicklung einer Ähnlichkeitsfunktion für Firmen- und Personenadressen
Projektpartner: Relate GmbH
Projektleiter: Uwe Draisbach
Projektteam: Uwe Draisbach, Florian Thomas
Abstract: Ziel der Kooperation zwischen der Relate GmbH und dem Hasso-Plattner-Institut war die Entwicklung und Optimierung einer Ähnlichkeitsfunktion zur Berechnung der Ähnlichkeit zweier Personen- oder Firmen-Datensätzen. Die Relate GmbH stellte Testdaten zur Verfügung, auf Basis derer eine entsprechende Ähnlichkeitsfunktion gemäß den vereinbarten Anforderungen entwi-

ckelt und evaluiert wurde. Die Ähnlichkeitsfunktion verwendet verschiedene Algorithmen, die dem aktuellen Stand der Wissenschaft entsprechen. Zusätzlich wurde eine Parallelisierung der Verarbeitung implementiert, um die Leistungsfähigkeit moderner Multicore-Prozessoren besser auszunutzen. Die Ähnlichkeitsfunktion wird regelmäßig zur Deduplizierung von ca. 60 Mio. Adressdatensätzen eingesetzt und ist zusätzlich Bestandteil der Software reDUB.

5.6 Similarity Search

Projekt: Similarity Search
Projektpartner: SCHUFA Holding AG
Projektleiter: Prof. Dr. Felix Naumann
Projektteam: Dustin Lange
Abstract: Die SCHUFA Holding AG ist die größte deutsche Wirtschaftsauskunftei mit Informationen über die Kreditwürdigkeit von etwa 66 Millionen Personen. Täglich werden etwa 275.000 Anfragen zur Kreditwürdigkeit von der SCHUFA beantwortet, die von den Geschäftspartnern (etwa Banken oder Online-Versandhäuser) an die SCHUFA gestellt werden. Die Anfragen enthalten in einigen Fällen Abweichungen von den Werten im Datenbestand, etwa Tippfehler im Namen oder abweichende Anschriften.
In diesem Forschungsprojekt haben wir Verfahren zur effektiven Anfragebeantwortung entwickelt, sodass für möglichst jede Anfrage die korrekte Person im Datenbestand identifiziert werden kann. Darüber hinaus gewährleistet das Verfahren eine effiziente Anfrageausführung, um den Geschäftspartnern der SCHUFA eine schnelle Antwort zu liefern. Die SCHUFA plant den produktiven Einsatz der von uns entwickelten prototypischen Implementierung.

5.7 Stratosphere

Forschungsprojekt: Data Cleansing in Stratosphere
Projektpartner: DFG
Projektleiter: Prof. Dr. Felix Naumann
Projektteam: Arvid Heise, Tommy Neubert, Fabian Tschirschnitz
Abstract: Stratosphere ist ein verteiltes System zur Datenanalyse von großen Datenmengen, das in Kooperation mit der TU und HU Berlin entwickelt wird. Wir entwickeln Datenreinigungsoperationen, die in eine komplexe, deklarative Anfrage eingebettet werden können. Ein besonderes Augenmerk liegt auf der Erforschung von Interaktionen zwischen diesen und anderen Operationen, um die Ausführung der Anfragen zu optimieren und damit Zeit und Geld sparen zu können.

6 Publikationen

6.1 Begutachtete Konferenzartikel

- Christoph Böhm and Daniel Hefenbrock and Felix Naumann. Scalable Peer-to-Peer-based RDF Management. In Proceedings of the 8th Int. Conference on Semantic Systems, Graz, Austria, 9 2012.
- Toni Gruetze and Christoph Böhm and Felix Naumann. Holistic and Scalable Ontology Alignment for Linked Open Data. In Proceedings of the 5th Linked Data on the Web (LDOW) Workshop at the 21th International World Wide Web Conference (WWW), Lyon, France, 4 2012.
- Dandy Fenz, Dustin Lange, Astrid Rheinländer, Felix Naumann, and Ulf Leser. Efficient Similarity Search in Very Large String Sets. In Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM), Chania, Crete, Greece, 2012.
- Christoph Böhm, Markus Freitag, Arvid Heise, Claudia Lehmann, Andrina Mascher, Felix Naumann, Mauricio Hernandez, Vuk Ercegovac and Peter Haase. GovWILD: Integrating Open Government Data for Transparency (demo). In Proceedings of the International World Wide Web Conference (WWW), Lyon, France, 2012.
- Martin Köppelmann, Dustin Lange, Claudia Lehmann, Marika Marszalkowski, Felix Naumann, Peter Retzlaff, Sebastian Stange, Lea Voget. Scalable Similarity Search with Dynamic Similarity Measures. In Proceedings of the 6th International Workshop on Ranking in Databases (DBRank) in conjunction with VLDB, Istanbul, Turkey, 2012.
- Alexander Albrecht, Felix Naumann. Schema Decryption for Large Extract-Transform-Load Systems. In Proceedings of the 31st International Conference on Conceptual Modeling (ER 2012), Florence, Italy, 2012.
- Saeedeh Momtazi. Fine-grained German Sentiment Analysis on Social Media. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 2012.
- Tobias Vogel and Felix Naumann. Automatic Blocking Key Selection for Duplicate Detection based on Unigram Combinations. In Proceedings of the 10th International Workshop on Quality in Databases (QDB) in conjunction with VLDB, 2012.
- Arvid Heise, Astrid Rheinländer, Marcus Leich, Ulf Leser, Felix Naumann. Meteor/Sopremo: An Extensible Query Language and Operator Model. In

Proceedings of the International Workshop on End-to-end Management of Big Data (BigData) in conjunction with VLDB 2012, Istanbul, Turkey, 2012.

- Christoph Böhm, Gerard de Melo, Felix Naumann, and Gerhard Weikum. LINDA: Distributed Web-of-Data-Scale Entity Matching. In Proceedings of the International Conference on Information and Knowledge Management (CIKM), Maui, Hawaii, 2012.
- Ziawasch Abedjan, Johannes Lorey, and Felix Naumann. Reconciling Ontologies and the Web of Data. In Proceedings of the International Conference on Information and Knowledge Management (CIKM), Maui, Hawaii, 2012.
- Jana Bauckmann, Ziawasch Abedjan, Heiko Müller, Ulf Leser, and Felix Naumann. Discovering Conditional Inclusion Dependencies. In Proceedings of the International Conference on Information and Knowledge Management (CIKM), Maui, Hawaii, 2012.
- Christoph Böhm and Gjergji Kasneci and Felix Naumann. Latent Topics in Graph-Structured Data. In Proceedings of the Conference on Information and Knowledge Management (CIKM), 2012.
- Gjergji Kasneci. Reasoning about Knowledge from the Web - (Extended Abstract). In ICWE Workshops, pages 186-188, 2012. Springer.
- Enkelejda Tafaj and Gjergji Kasneci and Wolfgang Rosenstiel and Martin Bogdan. Bayesian online clustering of eye movement data. In Proceedings of the 2012 Symposium on Eye-Tracking Research and Applications, pages 285-288, 2012. ACM.
- Uwe Draisbach, Felix Naumann, Sascha Szott, Oliver Wonneberg. Adaptive Windows for Duplicate Detection. In Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, D.C., USA, 2012.

6.2 Zeitschriftenartikel

- Arvid Heise and Felix Naumann, Integrating Open Government Data with Stratosphere for more Transparency, Web Semantics: Science, Services and Agents on the World Wide Web 14(1):45 - 56, 2012.
- Alberto Abello, Jerome Darmont, Lorena Etcheverry, Matteo Golfarelli, Jose-Norberto Mazon, Felix Naumann, Torben Bach Pedersen, Stefano Rizzi, Juan Trujillo, Panos Vassiliadis, Gottfried Vossen, Fusion Cubes: Towards Self-Service Business Intelligence, International Journal of Data Warehousing and Mining (IJDWM), 2012.
- Daniel Rinser, Dustin Lange, Felix Naumann, Cross-lingual Entity Matching and Infobox Alignment in Wikipedia, Information Systems (IS), 2012.
- George Beskales, Gautam Das, Ahmed K. Elmagarmid, Ihab F. Ilyas, Felix Naumann, Mourad Ouzzani, Paolo Papotti, Jorge Quiane-Ruiz, and Nan Tang, The Data Analytics Group at the Qatar Computing Research Institute, SIGMOD Record 41(4), 2012.

- Melanie Herschel and Felix Naumann and Sascha Szott and Maik Taubert, Scalable Iterative Graph Duplicate Detection, Transactions on Knowledge and Data Engineering (TKDE) 24(11):2094-2108, 2012.

6.3 Buchkapitel

- Uwe Draisbach, Partitionierung zur effizienten Duplikaterkennung in relationalen Daten, of Ausgezeichnete Arbeiten zur Informationsqualität. Springer Vieweg, 2012

6.4 Technische Berichte

- Uwe Draisbach and Felix Naumann. Adaptive Windows for Duplicate Detection. Technical Report 49, Hasso-Plattner-Institut für Softwaresystemtechnik an der Universität Potsdam, 2012. ISBN 978-3-86956-143-1, ISSN 1613-5652.
- Jana Bauckmann and Ziawasch Abedjan and Ulf Leser and Heiko Müller and Felix Naumann. Covering or complete? Discovering conditional inclusion dependencies. Technical Report 62, Hasso-Plattner-Institut für Softwaresystemtechnik an der Universität Potsdam, 2012. ISBN 978-3-86956-212-4, ISSN 1613-5652.
- Alexander Albrecht and Felix Naumann. Understanding Cryptic Schemata in Large Extract-Transform-Load Systems. Technical Report 60, Hasso-Plattner-Institut für Softwaresystemtechnik an der Universität Potsdam, 2012. ISBN 978-3-86956-201-8, ISSN 1613-5652.

7 Vorträge

Es werden nur Vorträge genannt, die nicht in Zusammenhang mit einer Konferenzveröffentlichung stehen.

Prof. Felix Naumann

- Keynote: International Conference on Information Quality (ICIQ)
  "The Quality of Web Data"
- Keynote: International Conference on Web Engineering (ICWE)
  "Extreme Web Data Integration"

Dr. Gjergji Kasneci

- Keynote: International Workshop on Quality in Web Engineering (colocated with ICWE 2012)
  "Reasoning about Knowledge from the Web"

8 Web-Portale und -Services

- Website des Fachgebiets

  http://www.hpi.uni-potsdam.de/naumann/

- GovWILD: Government Web Integration for Linked Data

  http://www.govwild.org

- Black Swan: Discovering Events that Matter

  http://blackswanevents.org/

- HPI's open data initiatives

  http://www.hpi.uni-potsdam.de/opendata.html

9 Mitgliedschaften, Programmkomitees, Gutachtertätigkeiten

9.1 Mitgliedschaften

Prof. Felix Naumann

- Association for Computing Machinery (ACM)
- ACM Special Integrest Group Management of Data (SIGMOD)
- Gesellschaft für Informatik (GI)
- GI Fachgebiet Datenbanken (FGDB)
- Deutsche Gesellschaft für Informationsqualität (DGIQ)

9.2 Mitarbeit in Boards und Programmkomitees

Prof. Felix Naumann

- Area Editor of Information Systems (IS)
- Associate Editor of ACM Journal on Data and Information Quality (JDIQ)
- Demonstrations Chair for International Conference on Extending Database Technology (EDBT)
- Program committee member of the following conferences: ADBIS, ESWC
- Program committee member of the following workshops: ODBASE, DBSocial, LDOW, QDB, WOD

Dr. Gjergji Kasneci

- Reviewer for the Microsoft Research PhD Scholarship Programme
- Committee member of the International World Wide Web Conference (WWW)

- Committee member of the International Conference on Extending Database Technology (EDBT)