Datenreinigung

(Seminar)

Prof. Felix Naumann, Alexander Albrecht, Jana Bauckmann, Jens Bleiholder, Frank Kaufer, Melanie Weis

Hasso-Plattner-Institut für Softwaresystemtechnik GmbH
FB Informationssysteme

Prof.-Dr.-Helmert-Str. 2-3
D-14482 Potsdam

Überblick

Einführung und Motivation

Datenreinigungsthemen

1. Data Profiling
2. Duplikaterkennung
3. Datenfusion
4. Datenreinigungssysteme

Organisatorisches

Schlagzeilen

„Falsche Preisangaben in Warenwirtschaftssystemen des Einzelhandels kosten Konsumenten in den USA jährlich [illegible],5 Milliarden Dollar.“

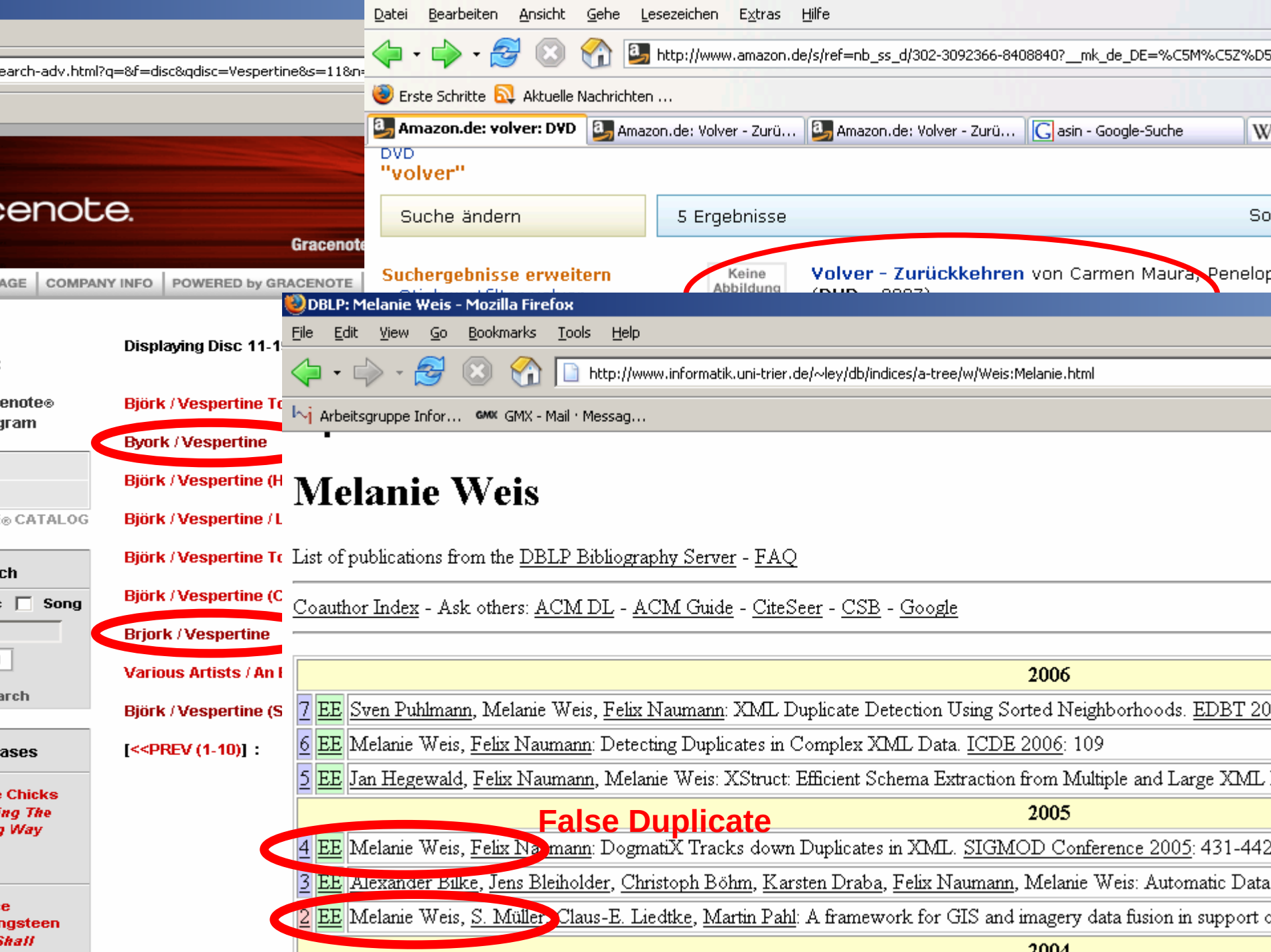
„Das Finanzamt der [illegible] barschecks mit Steuerrücke[illegible]gaben nicht zustellen.“

„Zwischen 50% und 80 [illegible] nregister der USA sind fehlerhaft, unvollständ[illegible], ode[illegible]g.“

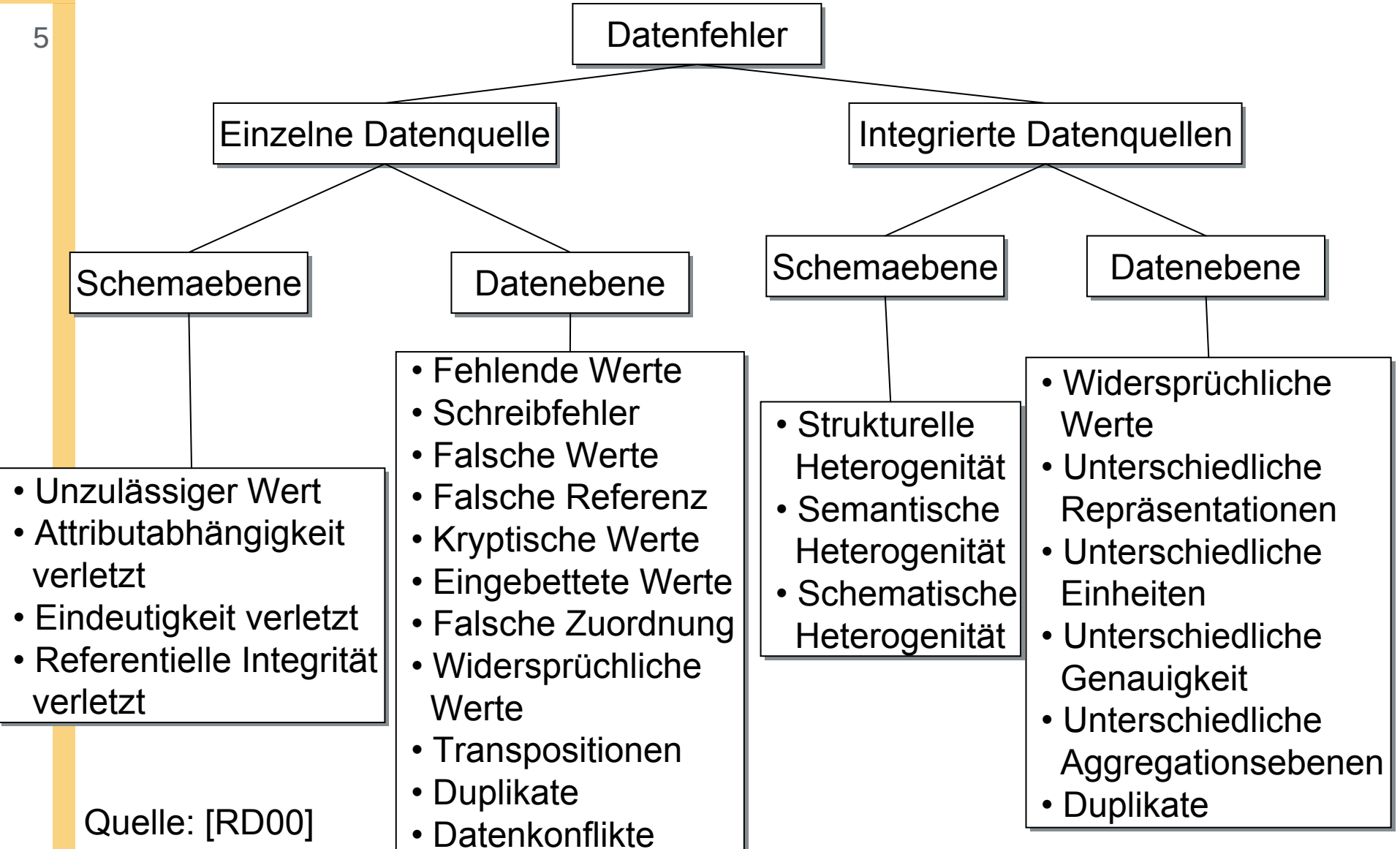
Datenqualität ist wichtig!
→ Datenreinigung

...

Quelle: [LN07]


Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
http://www.amazon.de/s/ref=nb_ss_d/302-3092366-8408840?__mk_de_DE=%C5M%C5Z%D5
Erste Schritte Aktuelle Nachrichten ...
Amazon.de: volver: DVD
Amazon.de: Volver - Zurü...
Amazon.de: Volver - Zurü...
asin - Google-Suche
DVD
"volver"
Suche ändern
5 Ergebnisse
Suchergebnisse erweitern
Keine Abbildung
Volver - Zurückkehren von Carmen Maura, Penelop
COMPANY INFO
POWERED by GRACENOTE
Displaying Disc 11-1
Björk / Vespertine To
Byork / Vespertine
Björk / Vespertine (H
Björk / Vespertine / L
Björk / Vespertine To
Björk / Vespertine (C
Brjork / Vespertine
Various Artists / An I
Björk / Vespertine (S
[<<PREV (1-10)] :
DBLP: Melanie Weis - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/w/Weis:Melanie.html
Arbeitsgruppe Infor...
GMX - Mail · Messag...
Melanie Weis
List of publications from the DBLP Bibliography Server - FAQ
Coauthor Index - Ask others: ACM DL - ACM Guide - CiteSeer - CSB - Google
2006
7 EE Sven Puhlmann, Melanie Weis, Felix Naumann: XML Duplicate Detection Using Sorted Neighborhoods. EDBT 20
6 EE Melanie Weis, Felix Naumann: Detecting Duplicates in Complex XML Data. ICDE 2006: 109
5 EE Jan Hegewald, Felix Naumann, Melanie Weis: XStruct: Efficient Schema Extraction from Multiple and Large XML
2005
False Duplicate
4 EE Melanie Weis, Felix Naumann: DogmatiX Tracks down Duplicates in XML. SIGMOD Conference 2005: 431-442
3 EE Alexander Bilke, Jens Bleiholder, Christoph Böhm, Karsten Draba, Felix Naumann, Melanie Weis: Automatic Data
2 EE Melanie Weis, S. Müller, Claus-E. Liedtke, Martin Pahl: A framework for GIS and imagery data fusion in support

Klassifikation von Datenfehlern

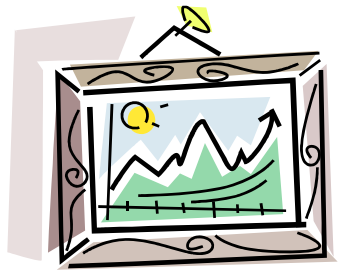


Quelle: [RD00]

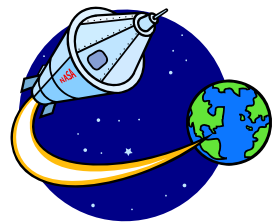
Entstehung von Datenfehlern

Vier Ursachen

1. Fehlerhafte Dateneingabe oder Erfassung
   - Manuelle Tippfehler (Weis, Weiss, Weiß, …)
   - Dummywerte (Titel1, Titel2, Titel3, …)
   - Messfehler (Barcodescanner, Schrifterkennungsalgorithmen, …)
2. Veralterung
   - Umzüge
   - Aktienkurse



3. Fehlerhafte Transformation oder Aggregation
   - Falscher Wechselkurs zwischen Währungen
   - Falsche angegebene Wertkorrespondenzen
   - Falsches Schema Mapping



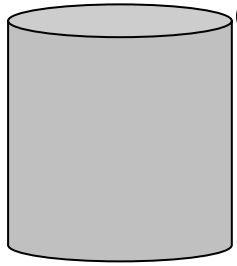
4. Integration
   - Duplikate und Konflikte entstehen durch Überlappung der Quellen

Auswirkungen von Datenfehlern

Analyst

Garbage-in-garbage-out Prinzip

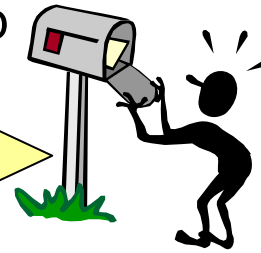
Data Mining



fehlerhafte Kunden DB

Customer relationship management

Kundenbetreuung



Kunde

- Fehlerhafte Daten im Data Warehouse
- Data Mining Algorithmus analysiert die Daten

→ Ergebnis basiert auf falschen Daten

→ Falsche strategische Entscheidungen

- Fehlerhafte Daten im Data Warehouse
- Kunde bekommt gleiches „persönliches Angebot“ mehrfach

→ Verärgerter Kunde

→ Druck, Umschlag, Versand als zusätzliche Kosten für die Firma

Umgang mit Fehlern

1. Profiling - Erkunden des Datenbestandes
   - Manuell & Toolbasiert
   - Statistiken (Minima, Maxima)
   - Häufigkeitsverteilungen
   - Musteranalyse (Telefonnummer)

2. Assessment
   - Definition von Bedingungen, denen die Daten entsprechen sollen (Alter < 150).
   - Messung, wie gut diese Bedingungen erfüllt sind.
   - Ergebnis: Bericht über die Anzahl/Verteilung von Fehlern

3. Monitoring
   - Maßnahmen zur Fehlerbeseitigung evaluieren und kotrollieren.

Maßnahmen der Fehlerbeseitigung

2 Phasen

1. Data scrubbing: Beseitigung einfacher Fehler in einem Datensatz
   - Normalisierung (Datumsformat, Abkürzungen, Adressen, …)
   - Konvertierung
   - Fehlende Werte und Ausreißer
   - Referenztabellen
2. Behandlung von tupelübergreifenden Fehlern
   - Duplikaterkennung
   - Datenfusion

Überblick

Einführung und Motivation

Datenreinigungsthemen

1. Data Profiling
2. Duplikaterkennung
3. Datenfusion
4. Datenreinigungssysteme

Organisatorisches

Data Profiling

- Analyse von Struktur und Inhalt der Daten
- Finden von Besonderheiten
- Bestimmung der Datenqualität
- Kein Data Mining (Analyse von Business Informationen)

ID	NAME	ZIP	CITY
1	Alexander	10997	Berlin
2	Frank	10629	Berlin
3	Jana	12333	Belin
4	Jens	10247	Berlin
5	Melanie	102$7	Berlin

Häufigkeitsverteilung

Musteranalyse

Data Profiling

Methoden (Auszug):

- Analyse von Abhängigkeiten (Primär- und Fremdschlüsseln)
  - Beziehungen zwischen unterschiedlichen Business Objekten
  - Beispiel: Abhängigkeiten von Kundendaten und Bestellungen aus verschiedenen Datenquellen finden.
- Analyse des Inhalts der Attribute, z.B. Datentyp oder Musteranalyse (mm/dd/YY)
- Finden fehlender und falscher Werte
- Erkennen von Daten- und Eingabefehlern
- Erkennen von Duplikaten

Thema 1: Bellman – Finden von FKs [JMD03, DJMS02]

- Erstellen von „summaries“ über Attributen (deren Werten) mittels Signatur basierend aufs Hashfunktionen
- Nutzung der summaries zum
  - Finden von Join-Pfaden
  - Finden zusammengesetzter Felder
  - Finden von heterogenen Tabellen
- Implementieren des Findens von Join-Pfaden

Thema 2: Finden von mehrattributigen INDs [KR03, KR02]

- Level-weiser Algorithmus über Wertigkeit der IND
- Abbilden von unären (A ⊆ B) und binären INDs (AB ⊆ CD) in einen Hypergraphen und Finden von Cliquen in diesem Graph; eine Clique ist Kandidat für IND
- Abbilden von binären (AB ⊆ CD) und ternären INDs (ABC ⊆ DEF) in einen Hypergraphen…
- Vorteil: Reduzieren des Suchraumes

- Implementieren des Findens von mehrwertigen INDs (ab 3-wertigen INDs, also ABC ⊆ DEF)
- unäre (A ⊆ B) und binäre INDs (AB ⊆ CD) werden gegeben

Thema 3: TANE – Finden von FDs [HKPT99]

- Unterteilen der Relation in Partitionen
- Level-weiser Algorithmus über Anzahl der Attribute in Partition:
  - Generieren der möglichen FDs
  - Pruning
  - Test
- Erweiterung für approximative FDs

- Implementation des Ansatzes zum Finden exakter FDs
- Vergleich mit IBM Information Server und Microsoft SQL Server Integration Services

Duplikaterkennung

(Semi-)Automatische Erkennung verschiedener Repräsentationen eines gleichen realen Objekts.

Verschiedene Bezeichungen

- Object Identification
- Record Linkage
- Reference Reconciliation
- Entity Resolution
- ...

MOVIE_RELATION

Title	Genre	Year	Director
Troy	Action	2004	Petersen
Typo	Wiederspruch	Fehlende Daten	
Troj	History		Petersen

Definition

Gegeben zwei Tupelmengen A und B

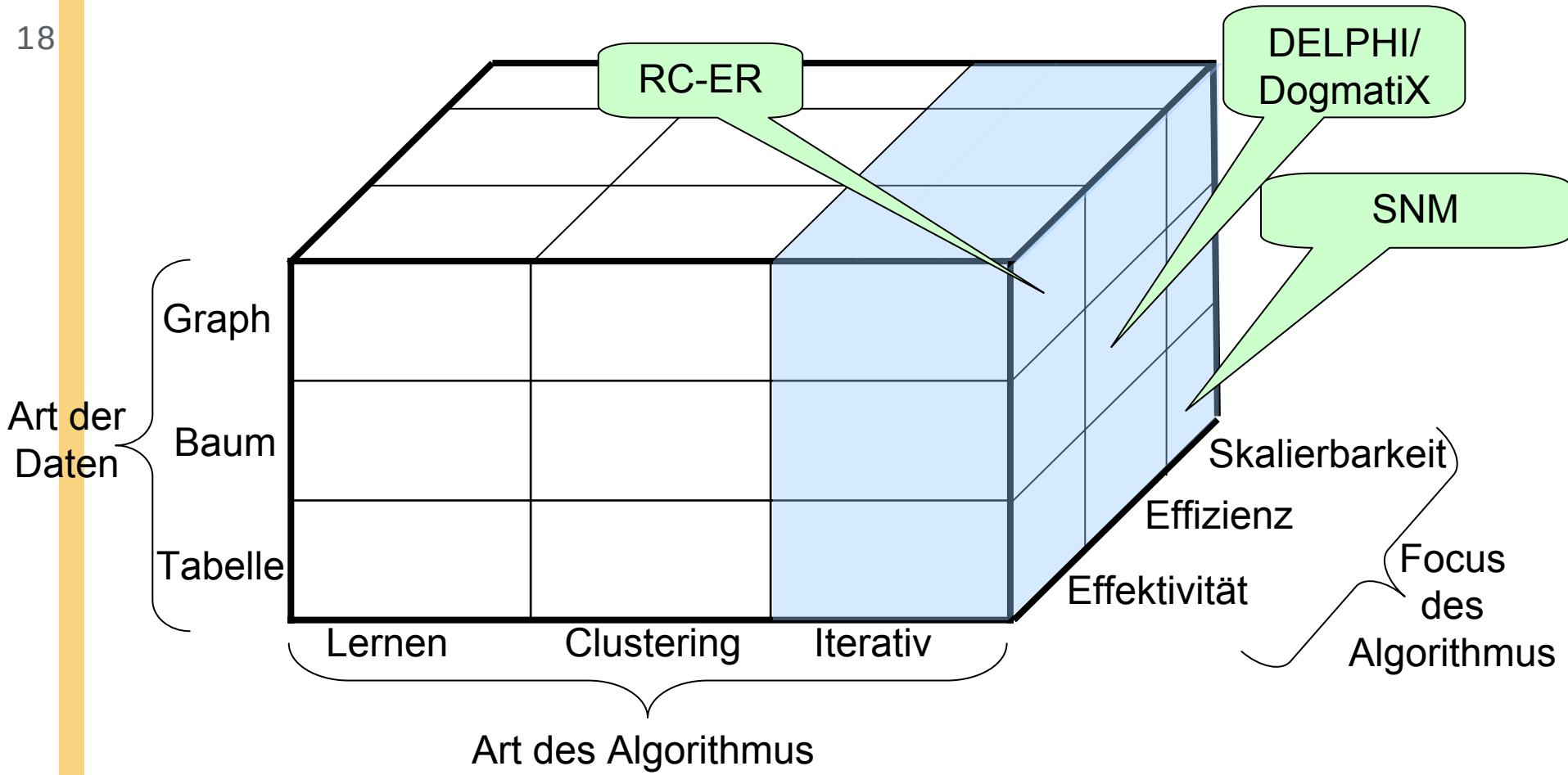
Kernidee:

- Bilde Kreuzprodukt aller Tupel.
- Für jedes Paar berechne Ähnlichkeit
  - Z.B. bzgl. Attributwerte
  - Z.B. bzgl. Beziehungen zu anderen Tabellen (Fremdschlüssel)
  - usw.
- Wähle Duplikatpaare aus
  - Ähnlichste Paare bis Schwellwert
  - Nebenbedingungen
- Bilde Duplikatcluster
  - Transitive Hülle

Probleme

- Anzahl und Komplexität der Vergleiche (Effizienz)
- Güte des Ähnlichkeitsmaßes (Effektivität)
- Große Datenmengen (Skalierbarkeit)

Themen


RC-ER
DELPHI/
DogmatiX
SNM
Graph
Baum
Tabelle
Art der
Daten
Skalierbarkeit
Effizienz
Effektivität
Focus
des
Algorithmus
Lernen
Clustering
Iterativ
Art des Algorithmus

Sorted Neighborhood Methode [HS98]

ID	Title	Year	Genre
17	Mask of Zorro	1998	Adventure
18	Addams Family	1991	Comedy
25	Rush Hour	1998	Comedy
31	Matrix	1999	Sci-Fi
52	Return of Dschafar	1994	Children
113	Adams Family	1991	Comedie
207	Return of Djaffar	1995	Children

Key generation

1.

ID	Key
17	MSKAD98
18	DDMCO91
25	RSHCO98
31	MTRSC99
52	RTRCH94
113	DMSCO91
207	RTRCH95

2. Sort keys

ID	Key
18	DDMCO91
113	DMSCO91
17	MSKAD98
31	MTRSC99
25	RSHCO98
52	RTRCH94
207	RTRCH95

Slide window & compare

3.

ID	Key
18	DDMCO91
113	DMSCO91
17	MSKAD98
31	MTRSC99
25	RSHCO98
52	RTRCH94
207	RTRCH95

sim(18,113) → duplicates

sim(52,207) → duplicates

Sorted Neighborhood Methode [HS98]

Erweiterungen

- Multi-pass
- Incrementell
- Domänenunabhägig
- Effizienzsteigerung
  - Gruppierung während der Vergleiche (UNION)
  - Finden von Duplikaten zu einer Gruppe (FIND)

Duplikaterkennung in Hierarchischen Daten (DELPHI) [AC02]

Beispiel:
Kundendimension

Relation R1

OrgId	Name	Address	CityId
1	Clintstone Assoc.	#1,Lake View Blvd.	C1
2	Compuware	#20,Main Street	C2
3	Compuwar	#20,Main Street	C3
4	Clintstone Associates	#1,Lake View	C4
5	Ideology Corp.	#10,Vancouver Pl.	C5
6	Victoria Films	#5,Victoria Av.	C6
7	Ideology Corporation	#10,Vanc. Pl.	C7
8	Clark Consultants Ltd.	#8,Cherry Street	C8
9	Clark Consultants	#8,Cherr St.	C9

R2

CityId	City	StateId
C1	Joplin	S1
C2	Jopin	S2
C3	Joplin	S4
C4	Joplin	S3
		S5
		S6
		S5
		S7
		S8

R3

StateId	State	CtryId
S1	MO	1
S2	MO	2
S3	MO	3
	Missouri	3
	BC	4
	British Columbia	4
	Aberdeen shire	5
	Aberdeen	5

R 4

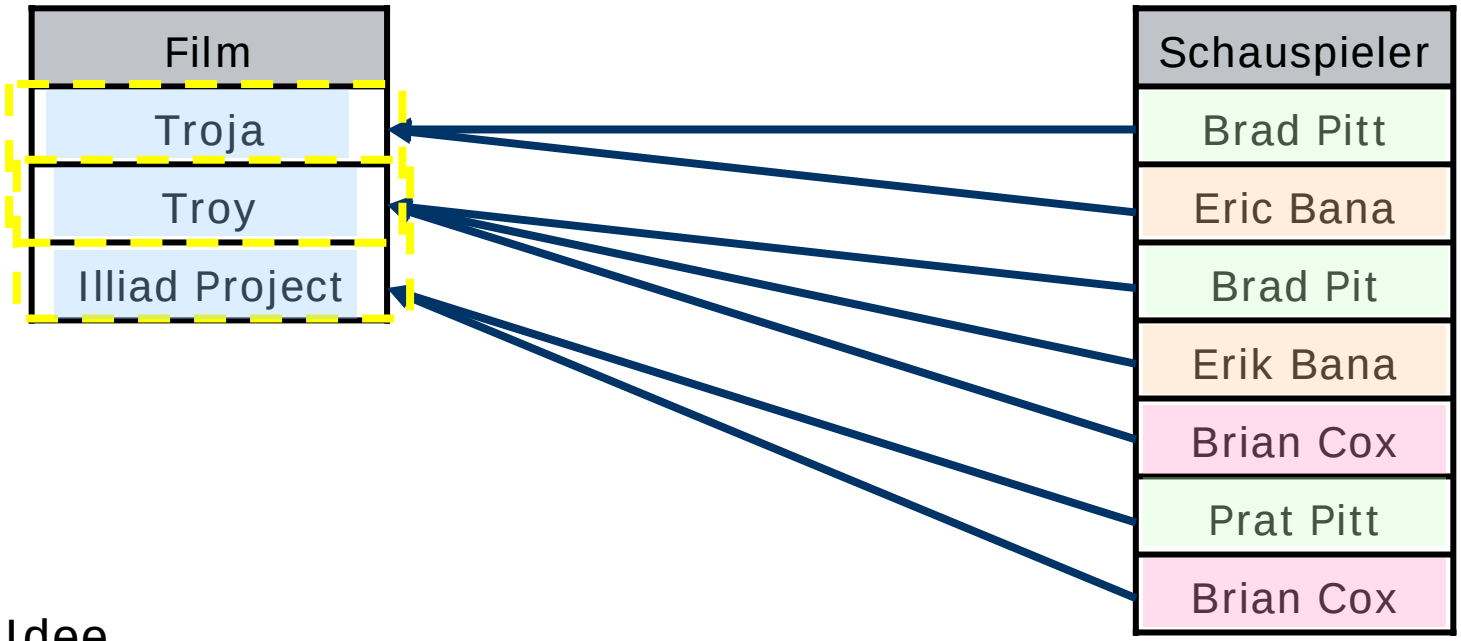
CtryId	Country
1	United States of America
2	United States
3	USA
4	Canada
5	UK

Hierarchische Ordnung durch Fremdschlüssel

Erweiterungen

- Filter zur Effizienzsteigerung
- Unterstützung beim Festsetzen eines Schwellwerts
- Duplikaterkennung in XML [WN05]
  - Definition
  - Ähnlichkeitsmaß
  - Algorithmus

Duplikaterkennung in Graph Daten (RC-ER) [BG07]

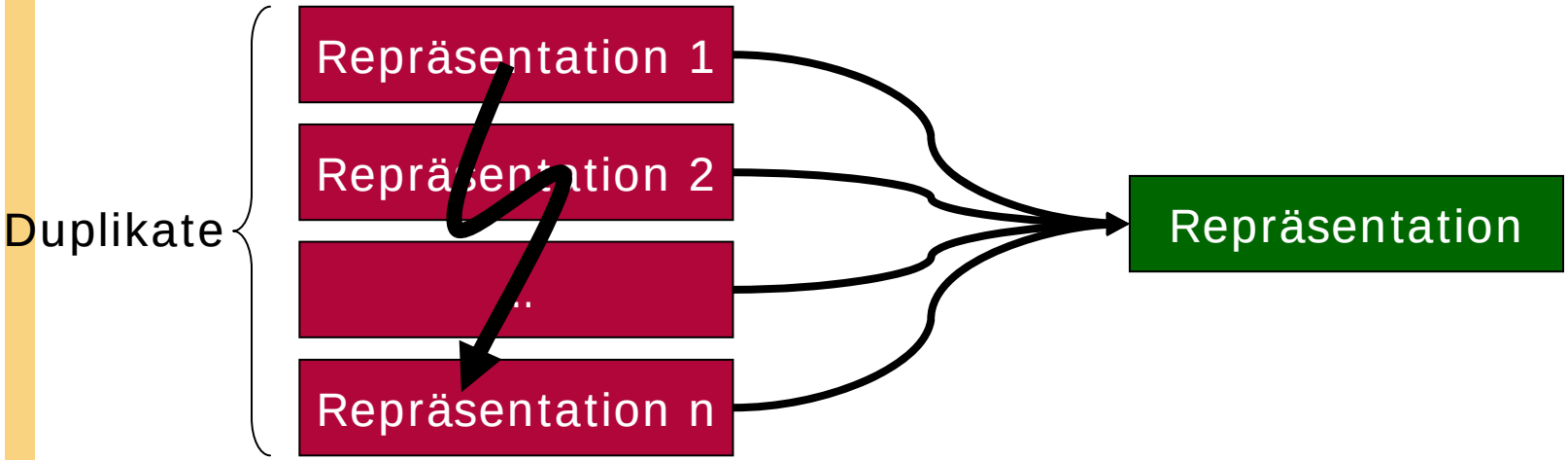


Idee

- Paare werden in einer beliebiegen Reihenfolge verglichen.
- Wenn ein Duplikat gefunden wird, vergrößert sich die Ähnlichkeit seiner Nachbarn
  → Nachbarn müssen (wieder) verglichen werden.

Datenfusion

Das Zusammenfügen mehrerer unterschiedlicher Repräsentationen ein und desselben Objektes der realen Welt zu einer einzigen konsistenten Repräsentation wird als Datenfusion bezeichnet.



Beispiel

IMDB

ID	Title	Year	Director	Genre
1	Snatch	2000	Ritchie	Crime
2	Troy	2004	Petersen	⊥
3	Vanilla Sky	2001	Crowe	Sci-Fi
4	Shrek	2001	Adamson	Anim.
5	M		wski	

Filmdienst

ID	Titel	Jahr	Rating	Genre
1	Snatch	1999	R	Crime
2	Troja	2004	R	History
3	Vannila Ski	2001	R	Sci-Fi
3	Vannile Sky	2000	16	Comedy

Choose(IMDB) | Max | Coalesce | Coalesce | Concat

Ergebnis:

ID	Title	Year	Director	Rating	Genre
1	atch	2000	Ritchie	R	Crime
2	Troy	2004	Petersen	R	History
3	Vanilla Sky	2001	Crowe	R	Sci-Fi, Comedy
4	Shrek	2001	Adamson	⊥	Anim.
5	Matrix	1999	Wachowski	16	Fantasy

Eindeutig | Duplikat | Unsicherheit | Widerspruch

Thema 1: TSIMMS und HERMES

Ausarbeitung

- Wie funktionieren die beiden Systeme?
  - Wie fusioniert TSIMMIS [PaAG96]?
    - Wie erkennt es Duplikate?
  - Wie fusioniert HERMES [SABE95]?
    - Welche Strategien verwendet es?
  - Unterschiede, Gemeinsamkeiten

ID	**Title**	**Year**	**Rating**	**Director**	**Genre**
1	Snatch	2000	R	Ritchie	Crime
2	Troy	2004	R	Petersen	⊥
3	Vanilla Sky	2001	R	Crowe	Sci-Fi
4	Shrek	2001	⊥	Adamson	Anim.
5	Matrix	1999	16	Wachowski	Fantasy

Thema 2: Konsistente Antworten

Ausarbeitung

- Was sind konsistente Antworten?
  - Update- vs. Insert/Delete-Semantik
  - Komplexität
  - Berechnung
- ConQuer, ein CQA System [FuFM05a]
  - Funktionsweise
  - Komplexität
  - Vgl. zu anderen Systemen

ID	**Title**	**Year**	**Director**	**Rating**	**Genre**
4	Shrek	2001	Adamson	⊥	Anim.
5	Matrix	1999	Wachowski	16	Fantasy

Thema 3: Partial Natural Outerjoin

Implementierung

- Probabilistic partial values
  - Was ist das?
  - Wie kann man es umsetzen?
  - Wie funktionieren relationale Operatoren?
- Partial Natural Outerjoin [TsCh00]
  - Als Fusionsoperator

ID	**Title**	**Year**	**Director**	**Rating**	**Genre**
1	Snatch	$[2000^{0.5}/1999^{0.5}]$	Ritchie	R	Crime
2	$[Troy^{0.5}/Troja^{0.5}]$	2004	Petersen	R	$[\perp^{0.5}/History^{0.5}]$
3	$[Vanilla\ Sky^{0.33}/Vannila\ Ski^{0.33}/Vannile\ Sky^{0.33}]$	$[2001^{0.66}/2000^{0.33}]$	Crowe	$[R^{0.5}/16^{0.5}]$	Sci-Fi/Comedy
4	Shrek	2001	Adamson	$\perp$	Anim.
5	Matrix	1999	Wachowski	16	Fantasy

Thema 4: Full Disjunction

Implementierung

- Operator zur Datenfusion
  - Full Disjunction [CoSa05]
  - Unterschied zu Outerjoin?
  - Schnelle Implementierung
  - Verwendbarkeit relationaler Operatoren?
  - Clou hier: Subsumption

ID	**Title**	**Year**	**Director**	**Rating**	**Genre**
5	Matrix	1999	Wachowski	⊥	⊥
5	Matrix	1999	Wachowski	16	Fantasy

ID	**Title**	**Year**	**Director**	**Rating**	**Genre**
5	Matrix	1999	Wachowski	16	Fantasy

Datenreinigungssysteme

- Software-Plattformen für Datenreinigung
- IBM® WebSphere® Information Server
- Microsoft® **S**QL Server **I**ntegration **S**ervices (SSIS)
- **E**(xtraction)**T**(tansform)**L**(oad) Tools
- Bieten Techniken der Duplikaterkennung (Fuzzy Grouping) und Datenfusion (Survive Stage)

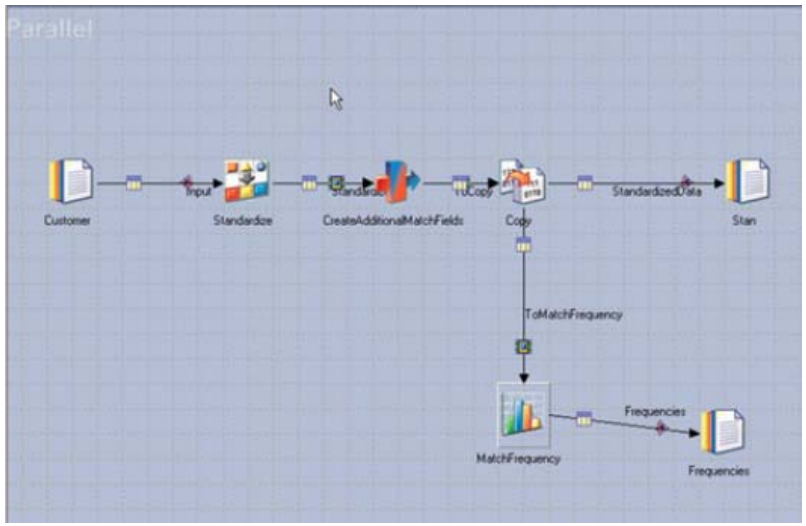
IBM®

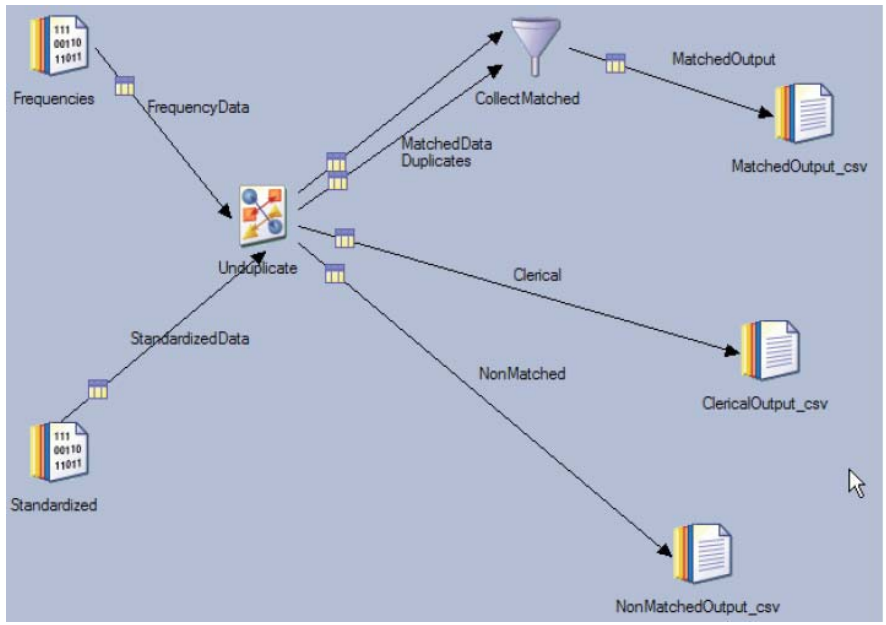
Microsoft® SQL Server™ 2005

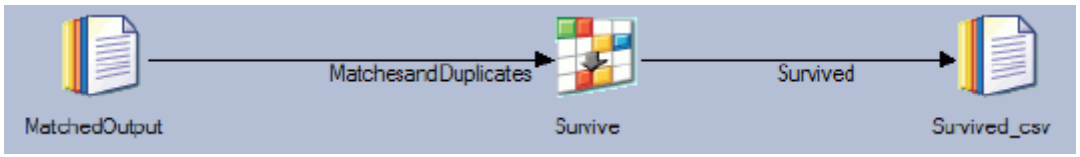
Beispielszenario:

- Bank will Kunden über neue Finanzdienstleistungen informieren
- Nur ein Angebot pro Haushalt verschicken
- Kundeninformationen werden über das Konto erfasst
- Beispiel: Ein Ehepaar hat vier Konten (zwei Girokonten, private Altersvorsorge, Aktienfond)
- Kundendatenbank enthält Duplikate für Namen und Adressen identischer Haushalte

Beispielszenario:





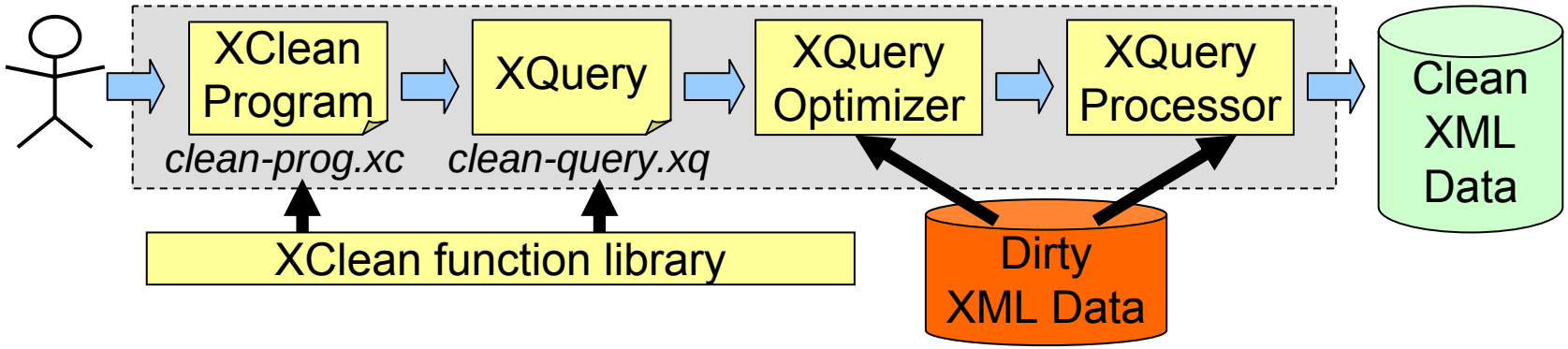


Datenreinigungssysteme

Aufgabe:

- Implementierung eines Beispielszenarios auf unterschiedlichen Software-Plattformen
- Qualitative und Quantitative Analyse
- Vergleich der Implementierungen / Software-Plattformen
- Literatur
  - Muller, H. and Freytag, J. C. Problems, Methods and Challenges in Comprehensive Data Cleansing, Technical Report, HUB-1B-164, 2003
  - J. Barateiro, H. Galhardas: A Survey of Data Quality Tools, Datenbank-Spektrum, 2005

XClean: Datenreinigung in XML [WM07]



- Spezifikation einer Abfolge von Operatoren (Scrubbing, Duplikaterkennung, Fusion, ...)
- Deklarative Spezifikationssprache
- Kompiliertes Programm (XQuery) auf der Datenbank ausführbar
- Ergebnis der ausgeführten Xquery sind bereinigte Daten

Forschungsprototypen zur Deklarativen Datenreinigung

AJAX: Datenreinigung in relationalen Daten [GF01]

- XClean wurde durch AJAX inspiriert, daher ähnlicher Ansatz
- Spezifikation von Transformationen in SQL
- Überbrückung von SQL Einschränkungen
- Data Lineage

Überblick

Einführung und Motivation

Datenreinigungsthemen

1. Data Profiling
2. Duplikaterkennung
3. Datenfusion
4. Datenreinigungssysteme

Organisatorisches

Basics

- Feste Termine
  - Dienstags, 17:00 Uhr – 18:30 Uhr
  - Raum A-1.2
  - Vom 24.04.2007 bis 18.07.2007
- Pro Termin
  - Zwei Vorträge
  - Diskussion über die Themen
  - letzter Termin: Demo Session
- Sprechstunde: Montags, 11 – 12 Uhr

Themen

Data Profiling

- Data-Profilung Konzepte und Methoden kommerzieller Produkte, z.B. IBM® WebSphere® Information Analyzer und Microsoft® SSIS Einordnung und Bewertung *(Implementierung)*

Doublettenerkennung

- Die Sorted Neighborhood Methode + Erweiterungen für Doublettenerkennung in relationalen Daten *(Implementierung)*
- DELPHI und DogmatiX: Duplikaterkennung in hierarchischen relationalen und XML Daten *(Ausarbeitung)*
- Doublettenerkennung in Graph Daten *(Implementierung)*

Datenfusion

- TSIMMIS und Hermes, 2 Fusionssysteme *(Ausarbeitung)*
- Consistant Query Answering (Theorie / Systeme) *(Ausarbeitung)*
- Probabilistische Fusion *(Ausarbeitung)*
- Partial Natural Outerjoin *(Implementierung)*

Datenreinigungssysteme

- Data-Cleaning Konzepte und Methoden kommerzieller Produkte, z.B. IBM® WebSphere® QualityStage und Microsoft® SSIS Einordnung und Bewertung *(Implementierung)*
- Deklaratives Data Cleaning in relationalen un XML Daten, z.B. AJAX vs. XClean *(Ausarbeitung)*

Themenvergabe

- Möglichst gleichmäßige Verteilung der Themen über fünf Betreuer
  - Alexander Albrecht: kommerzielle Produkte
  - Jana Bauckmann: Erkennen von INDs, FDs
  - Jens Bleiholder: Datenfusion
  - Frank Kaufer: Matching
  - Melanie Weis: Duplikaterkennung
- Themenwunschliste
  - per Email an Frank (frank.kaufer@hpi.uni-potsdam.de) bis spätestens 1.5.2007
  - mindestens 2 Themen
    - aus mindestens zwei Bereichen
    - nach absteigender Präferenz sortiert
    - bei Implementierung Teampartner angeben
- Themenvergabe: 2.5.2007

Leistungserfassung

Anwesenheit an allen Terminen

Implementierung

- 2er Teams
- Implementierung in Java zum gewählten Thema
  + ca. 5 Seiten Dokumentation
- 20-minütiger Vortrag pro Teammitglied + Diskussion
- 15-minütige Präsentation der Implementierung

Ausarbeitung

- Ein Thema pro Student
- Ausarbeitung zum gewählten Thema (ca. 15 – 20 Seiten in LaTeX Vorlage)
- Halbstündiger Vortrag + Diskussion

Abgabe bis spätestens 1.08.2007, 23:59:59 Uhr

Termine im Überblick

Termin	Themen
28.5.2007	Vortrag1 Vortrag2 *Data Profiling*
4.6.2007	Vortrag3
	Vortrag4
18.6.2007	Vortrag5 *Doublettenerkennung* Vortrag6
25.6.2007	Vortrag7 Vortrag8 *Datenfusion*
3.7.2007	Vortrag9
	Vortrag10
10.7.2007	Vortrag11 *Datenreinigungssysteme* Vortrag12
17.7.2007	Demo Session

Werbung

- Vorlesung Informationsintegration
  - Dienstags 9:15 - 10:45
  - Donnerstags 13:30 - 15:00
- Forschungsseminar MO 15:15-16:45 Uhr
- Masterarbeiten
  - Automatische Generierung von Fusionsanfragen
  - Partitionierung zur effizienten Duplikaterkennung in relationalen Daten
  - Skalierbares Schema Matching mittels Tree-Edit Distance
  - ...
- Tutoren
  - Datenintegration bei exCentos
  - Unterstützung im ALADIN Projekt

Referenzen

- [JMD03] Johnson, Marathe, Dasu: Database Exploration and Bellman. IEEE Daa Eng. Bull. 2003.
- []DJMS02] Dasu, Johnson, Muthukrishnan, Shkapenyuk: Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD 2002.
- [KR03] Koeller, Rundensteiner: Discovery of High-Dimensional Inclusion Dependencies. ICDE 2003.
- [KR02] Koeller, Rundensteiner: Discovery of High-Dimensional Inclusion Dependencies. Technical Report WPI-CS-TR-02-15. Worchester Polytechnic Institute. 2002.
- [HKPT99] Huhtala, Kärkkäinen, Porkka, Toivonen: TANE: An Efficient Algorithm for Discovering Functional and Approximate Dependencies. The Computer Journal. Vol 42. 1999.

Referenzen (cont.)

[RD00] Rahm and Do. Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, vol. 23, 2000.

[LN07] Leser and Naumann. Informationsintegration - Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. dpunkt.verlag, 2007.

[WM07] Weis and Manolescu. Declarative XML Data Cleaning with XClean. CAISE 2007 (to appear).

[HS98] Hernandez and Stolfo. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. Data Mining and Knowledge Discovery, vol. 2, nr. 1, 1998.

[AC02] Ananthakrishna, Chaudhuri, and Ganti. Eliminating Fuzzy Duplicates in a Data Warehouse. VLDB 2002.

[BG07] Bhattacharya and Getoor. Collective Entity Resolution in Relational Data. ACM Transactions on Knowledge Discovery from Data, vol. 1, nr. 1, 2007.

[GF01] Galhardas,Florescu, Shasha, Simon, and Saita. Declarative Data Cleaning: Language, Model, and Algorithms. VLDB 2001.

[WN05] Weis and Naumann. DogmatiX Tracks Down Duplicates in XML. SIGMOD 2005.

[PaAG96] Papakonstantinou, Abiteboul, and Garcia-Molina. Object Fusion in Mediator Systems. VLDB 1996.

[SABE95] Subrahmanian et. al. Hermes: A heterogeneous reasoning and mediator system. Techreport University of Maryland 1995.

[FuFM05a] Fuxman, Fazli, and Miller. ConQuer: efficient management of inconsistent databases. SIGMOD 2005.

[TsCh00] Tsai and Chen. Partial Natural Outerjoin – An Operation for Interoperability in a Multidatabase Environment. Journal of Information Science and Engineering. 16(4). July 2000.

[CoSa05] Cohen and Sagiv. An Incremental Algorithm for Computing Ranked Full Disjunctions. PODS 2005.

Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Noch Fragen?