

Seminar Data Cleansing

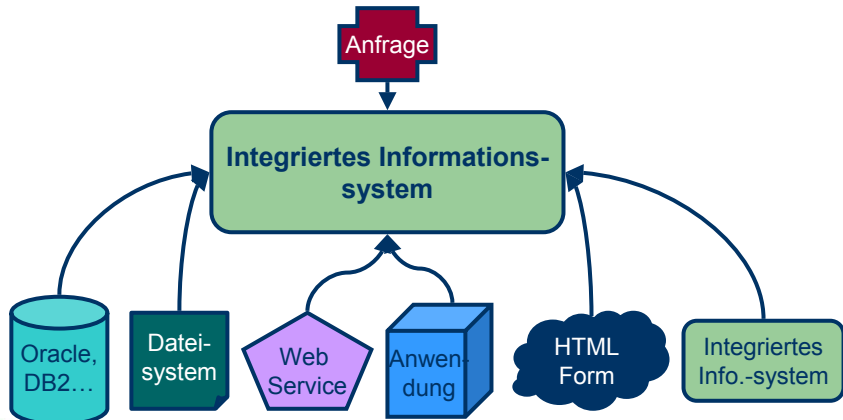
22.10.2003

Felix Naumann

Forschungsgruppe Informationsintegration

- Juniorprofessor: Felix Naumann
- Mitarbeiter
 - Jens Bleiholder
 - Melanie Weis (ab 1.11.)
- Themen
 - Objektidentifikation
 - Informationsintegration
 - Optimierung
 - Visualisierung

Integrierte Informationssysteme



3



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Data Cleansing

- Reinigt verschmutzte Daten
 - Datenqualität
 - Datenfehler
- Verschmutzung
 - Bei Dateneingabe
 - Bei Daten-transformation / -aggregation / -darstellung
 - Bei Informationsintegration
 - selbst wenn integrierte Quellen sauber sind

4



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Organisation

- Voraussetzungen
 - Themen werden allein oder zu zweit bearbeitet
 - Aufteilung der einzelnen Themen selbst überlegen und dann mit Betreuer besprechen
 - Referat, ca. 45 min (bei 1 Person), oder je 30 min
 - + Diskussion
 - Vorbereitung mit Betreuer jederzeit
 - Pflicht: Folienbesprechung 1 Woche vor dem Vortrag
 - Hausarbeit, je ca. 15 Seiten
 - Frist: 20.Feb 2004
 - Pflicht: Besprechung 2 Wochen nach Vortrag
 - LaTeX
 - Regelmäßige Teilnahme

5



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Literaturrecherche

- NEC ResearchIndex
 - Volltext, cross-references
 - <http://citeseer.nj.nec.com/>
- DBLP
 - Relativ vollständig, Home-Page Links
 - <http://www.informatik.uni-trier.de/~ley/db/>
- Bibfinder
 - Simultanes Durchsuchen diverser Archive
 - <http://kilimanjaro.eas.asu.edu/>
- Bibliothek!

6



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Allgemeine Hinweise

- Dozenten sind ansprechbar!
 - Vorbesprechung zur Einengung des Themas
 - Folien durchgehen
 - Evt. Eingrenzung der Ausarbeitung
 - Abgrenzung von existierenden Arbeiten
- Diskussion erwünscht
 - Keine Angst vor Fragen: Fragen sind keine Kritik
 - Eine Frage nicht beantworten können ist in Ordnung – bemühen Sie sich, sonst klären wir das in der Diskussion
- Tiefe, nicht Breite
 - Lieber das Thema einengen und dafür Details darstellen
 - „Aha“-Erlebnis erzeugen!
- Bezug nehmen
 - Einordnung der Arbeit im Seminarthema vornehmen
 - Vergleich zu anderen Arbeiten im Seminar, soweit angemessen

7



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Allgemeine Hinweise

- Werten und bewerten
 - Keine Angst vor nicht ganz zutreffenden Aussagen
 - Begründen und argumentieren
 - Kritikloses Abschreiben ist fehl am Platz
- Literaturrecherche ist erwünscht
 - Die ausgegebenen Arbeiten sind Anker
 - Weiterführende Arbeiten müssen herangezogen werden
 - Auch Grundlagen nachlesen
- Bibliotheken besuchen ist erlaubt; nicht alles ist im WWW (aber das meiste)

8



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Hinweise zum Vortrag

- Ca 45 Minuten
- Klare Gliederung
- Ab und ab Hinweise geben, wo man sich befindet
- Bilder und Grafiken; **Beispiele**
- Font: mind. 16pt
- Keine langen Erklärungen / Sätze auf den Folien: Stichwörter
- Vorträge müssen auch unterhaltend sein – Gimmicks, Gags, Rhythmuswechsel, Einbeziehen der Zuhörer, etc.
- Adressat sind alle Seminarteilnehmer, nicht nur die Betreuer

Hinweise zur Ausarbeitung

- Eine gedruckte Version abgeben
- **Selbständigkeitserklärung unterschreiben**
- Referenzen
 - Alle verwendeten und nur die
 - Liste am Schluss
- Korrekt zitieren
 - Vorsicht vor Übernahme von Textpassagen, wenn, dann deutlich kennzeichnen
 - Keine Aussagen ohne Evidenz oder Verweis auf Literatur
- Achten sie auf Rechtschreibfehler und Sprache!
- LaTeX – Vorlage demnächst

Hinweise zur Ausarbeitung –2-

- Gezielt und sachlich schreiben
 - Füllwörter vermeiden
 - Knappe Darlegung, präzise Sprache
 - Argumentieren und referenzieren
- Gliederung!
 - Beginnen Sie mit einer Gliederung und diskutieren Sie die mit dem Betreuer
- Kommen Sie zu Aussagen
 - Vorteile, Nachteile, verwandte Arbeiten, mögliche Erweiterungen, Anwendbarkeit, eigene Erfahrungen, ...

Warnung zum Thema „Selbständigkeit“

- Abschreiben von Arbeiten oder Übernahme von Folien erzwingt Ausschluss aus dem Seminar
 - Zitate sind natürlich möglich!
- Wir kennen das Gebiet (wahrscheinlich) besser als Sie ...
- Seminararbeit mit unterschriebener Selbständigkeitserklärung versehen

Überblick über den Vortrag

- Definition: Verschmutzte Daten
 - Datenqualität
 - Fehlerarten
- Duplikate
 - Algorithmen und Techniken zur Erkennung
- Weitere Cleansing Schritte (nach Rahm/Do'00)
- Einordnung der Seminarthemen

Datenqualität / Informationsqualität

- Was ist Datenqualität ?
 - „Fitness for use“
 - Anwendungsabhängig
- Folgen geringer Datenqualität
 - Falsche Prognosen
 - Verpasstes Geschäft
- Qualität ist besonders bei integrierten Informationen interessant
 - Oft keine Kontrolle über Informationsquellen (Autonomie!)
 - Oft zweifelhafte Qualität
 - Internet macht Publikation leicht
 - Vielzahl verfügbarer Quellen

Informationsqualität

IQ := { Verständlichkeit, Ansehen,
Zuverlässigkeit, Alter,
Verfügbarkeit, Preis,
Konsistenz, Deckung,
Antwortzeit, Dichte,
Vollständigkeit, Menge,
Genauigkeit, Relevanz, ... }

"Even though quality cannot be defined, you know what it is."
Robert Pirsig

Informationsqualität

IQ := { Verständlichkeit, Ansehen,
Zuverlässigkeit, Alter,
Verfügbarkeit, Preis,
Konsistenz, Deckung,
Antwortzeit, Dichte,
Vollständigkeit, Menge,
Genauigkeit, Relevanz, ... }

Informationsqualität

IQ := { Verständlichkeit, Ansehen,
Zuverlässigkeit, Alter,
Verfügbarkeit, Preis,
Konsistenz, Deckung,
Antwortzeit, Dichte,
Vollständigkeit, Menge,
Genauigkeit, Relevanz, ... }

Datenqualität vs. Datenfehler

- Qualität kann nicht einzig durch Data Cleansing erhöht werden.
 - Ansehen, Objektivität, ...
- Accuracy ≠ Quality

Datenqualität

- DWH besonders anfällig für Qualitätsprobleme
 - Probleme akkumulieren
 - Qualität der Ursprungsdaten (Eingabe, Fremdfirmen, ...)
 - Qualität der Quellsysteme (Konsistenz, Constraints, Fehler, ...)
 - Qualität des ETL (Parsen, Transformieren, ...)
 - Probleme treten erst bei konsolidierter Sicht zu Tage
 - DWH unterstützt strategische Entscheidungen: hohe Folgekosten bei Fehlentscheidungen

19



Quelle: Prof. Ulf Leser (VL Data Warehouses)

Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Beispiel: Kampagnenmanagement

- Probleme im CRM eines Multi-Channel Vertriebs
 - Kunden doppelt geführt
 - Kunden falsch bewertet
 - Falsche Adressen
 - Haushalte / Konzernstrukturen nicht erkannt
- Folgen
 - „False positives“: Verärgerte Kunden durch mehrere / unpassende Mailings
 - „False negatives“: Verpasste Gelegenheiten durch fehlende / falsche Zuordnung (Cross-Selling)
 - Sinnlose Portokosten bei falschen Adressen

20



Quelle: Prof. Ulf Leser (VL Data Warehouses)

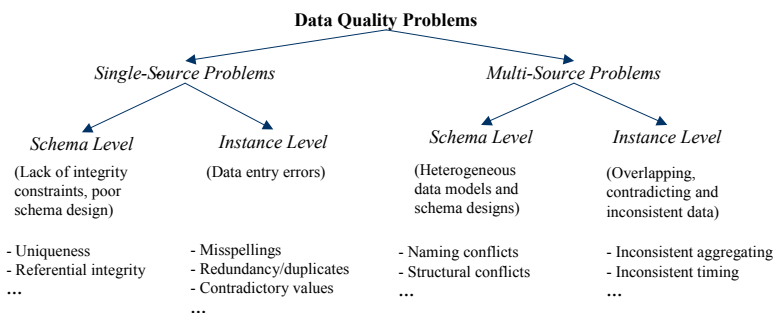
Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Kosten verschmutzter Daten

- A.T. Kearny: 25%-40% der operativen Kosten entstehen durch schlechte Datenqualität.
- Data Warehouse Institute: Industrie und Verwaltung in den USA verlieren jährlich 600 Milliarden USD.
- SAS Studie: Nur 18% der Deutschen Betriebe vertrauen ihren Daten.
- AT&T (70er): 20-30% aller Anschlüsse unbenutzt wegen schlechter Daten.
- 80% aller Krankenhaus Datensätze enthalten Fehler. Hmm...
- ...

Fehlerklassifikation



Fehlerarten

1. Datenintegrität
 - Primärschlüssel, Fremdschlüssel, Integritätsbedingungen, ...
2. Fehlende Daten
 - fehlende Attributwerte, fehlende Tupel, ...
3. Falsche Daten
 - Objektiv falsch: Negative Preise, 32.13.2002, Buchstaben statt Ziffern, unbekannte Codes, ...
 - Widersprüchliche Werte aus unterschiedlichen Quellen: Schreibweisen, Adressen, ...
4. Formatfehler
 - Verstoß gegen Formatvorgaben (Datum), falsche Genauigkeit,...

Quelle: Prof. Ulf Leser (VL Data Warehouses)

23



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Fehlerarten

5. Unplausible Daten
 - Ausreißer, ...
6. Semantik von NULL Werten
 - Wert möglich, aber unbekannt:
 - Ist er Professor ?
 - Wert möglich, existiert aber nicht:
 - Er ist kein Professor !
 - Wert unmöglich:
 - Kinder unter 18 haben keinen Titel
7. Duplikate
 - Kunden, Lieferanten, Produkte, etc.
8. Eingabefehler
9. Schlechte / fehlende Dokumentation

Quelle: Prof. Ulf Leser (VL Data Warehouses)

24



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Data Cleansing Operationen

- Ziel: Verbesserung der Datenqualität
 - Ergänzung fehlender Werte
 - Korrektur durch Lookup, Neuberechnen, Runden, ...
 - Erkennen und Löschen „unrettbarer“ Daten
 - Optimum kaum erreichbar: 80/20 Regel
- DQ muss gemessen werden
 - DQ Metriken **notwendig**
 - Verbesserung quantifizierbar
- Data Cleansing
 - Immer ein domänenabhängiger Prozess
 - Produkte gibt es nur für Adresdaten

Quelle: Prof. Ulf Leser (VL Data Warehouses)

25



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Nachvollziehbarkeit

- Änderungen durch DC müssen nachvollziehbar sein
- Auch das Fehlen von Daten
- Unprotokolliertes, nicht nachvollziehbares DC
 - Ad-Hoc DC führt aus Sicht der Anwender zu „fehlenden und falschen“ Daten, nicht zu besseren Daten
 - Daten im DWH müssen erklärbar sein
 - „Da fehlt ein Produkt im Report“
 - Analysewerkzeug fehlerhaft ?
 - Report falsch ?
 - Data Mart Definition falsch ?
 - Basisdatenbank unvollständig ?
 - ETL Prozeduren fehlerhaft ?
 - Übertragungsfehler ?
 - ...

Quelle: Prof. Ulf Leser (VL Data Warehouses)

26



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Data Cleansing

- In integrierten Systemen: Wo und wann?
- Föderierte Systeme
 - Online Cleansing (schwierig da teuer)
 - Beim Mediator
- Data Warehouses
 - Offline (Loading)
 - Eventl. auch bei den Quellen selbst (z.B. Dateneingabe)

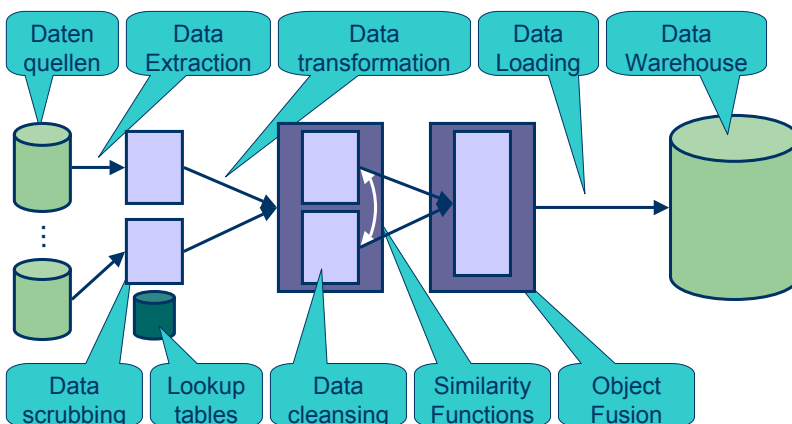
27



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Data Cleansing Schritte



28



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Beispiel Data Cleansing



```
<pub>
<Title> MAC: Merging Autonomous Content </Titel>
<First> Felix </First>
<Last> Naumann </Last>
<Year> 2002 </Year>
</pub>
```



```
<pub>
<Title> MAC: Merging Autonomous Content </Titel>
<First> Dr. Felix</First>
<Last> Naumann </Last>
```

Title	First	Last	Year
MAC: Merging Autonomous Content	Felix	Naumann	2002
Merging Autonomous Content	Dr. Felix	Naumann	2001

Seminarthemen

1. Ähnlichkeitsmaße
 2. Data Warehouse Duplicates
 3. Merge/Purge Algorithmus
 4. Domain-independent Dup. Detection
 5. AJAX framework
 6. IntelliClean
 7. Potter's Wheel ABC
 8. Data Cleansing in Genome Databases
 9. Data Lineage
 10. Object Fusion
- } Techniken und Algorithmen
- } Architekturen und Systeme
- } Anwendung und Postprocessing

Data Cleansing Themen

• Seminarthema 1: Ähnlichkeitsmaße

- Autonomous \approx Autonomus?
- Tupel 1 \approx Tupel 2?
- Similarity(Autonomous, Autonomus) = ?
- N-Gram
 - Tri-grams(„Autonomous“):
{(Aut), (uto), (ton), (ono), (nom), (omo), (mou), (ous)}
 - Tri-grams(„Autonomus“):
{(Aut), (uto), (ton), (ono), (nom), (omu), (mus)}
- Edit-distance
 - Autonomus + insert(„o“,7) = Autonomous
 - EditDistance(Autonomous, Autonomus) = 1
 - Smith-Waterman-Algorithm

31

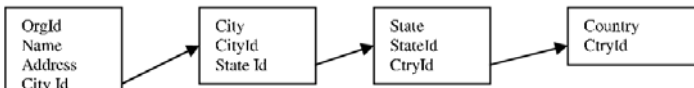


Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Data Cleansing Themen

• Thema 2: Data Warehouse Duplicates



OrgId	Name	Address	CityId	CityId	City	StateId	State	State	StateId	State	CtryId	Country
01	Clintstone Assoc.	#1, Lake View Blvd.	C1	C1	Joplin	S1	MO	Missouri	S1	MO	1	United States of America
02	Compuware	#20, Main Street	C2	C2	Joplin	S2	MO	Missouri	S2	MO	2	United States
03	Compuwar	#20, Main Street	C3	C3	Joplin	S3	MO	Missouri	S3	MO	3	USA
04	Clintstone Associates	#1, Lake View	C4	C4	Joplin	S3	MO	Missouri	S3	MO	3	USA
05	Ideology Corp.	#10, Vancouver Pl.	C5	C5	Victoria	S5	BC	British Columbia	S5	BC	4	United States
06	Victoria Films	#5, Victoria Av.	C6	C6	Victoria	S6	BC	British Columbia	S6	BC	4	United States
07	Ideology Corporation	#10, Vanc. Pl.	C7	C7	Vancouver	S5	BC	British Columbia	S6	British Columbia	4	USA
08	Clark Consultants Ltd.	#8, Cherry Street	C8	C8	Aberdeen	S7	Aberdeen shire	Aberdeen shire	S7	Aberdeen shire	5	Canada
09	Clark Consultants	#8, Cherr St.	C9	C9	Aberdeen	S8	Aberdeen	Aberdeen	S8	Aberdeen	5	UK

Organization (at Level 1)

City (at Level 2)

State (at Level 3)

Country (at Level 4)

32



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Data Cleansing Themen

- Thema 3: Merge/Purge Algorithmus
- Idee
 - Daten geschickt partitionieren.
 - Nur innerhalb dieser Partitionen Duplikate suchen.
- Algorithmus
 1. Create Key
 2. Sort
 3. Merge

Data Cleansing Themen

- Thema 4: Domain-independent Duplicate Detection
- Drei Ideen zusammengesetzt
 1. Smith-Waterman-Algorithmus (Domänen-unabhängig)
 2. UNION/FIND Algorithmus
 - Transitive Hülle aller Duplikate
 3. Clustering ähnlicher Datensätze
 - 2 Phasen (Domänen-unabhängig)
 - Priority Queue

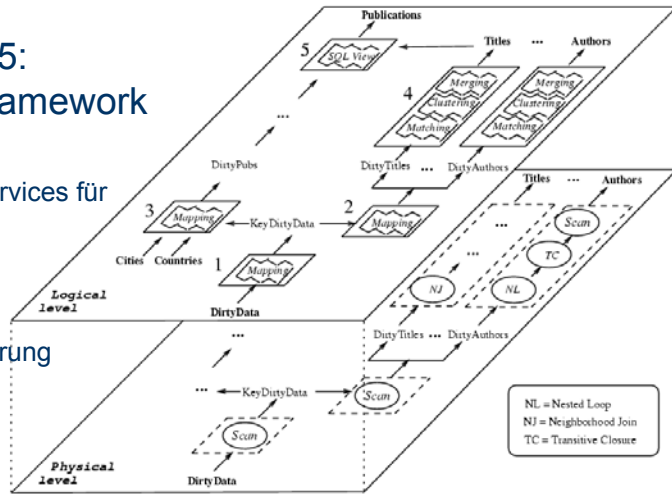
Data Cleansing Themen

- Thema 5:
AJAX framework

Operatoren/Services für

- Duplicates
- Errors
- Conflicts

+ SQL-Erweiterung
+ Optimierung



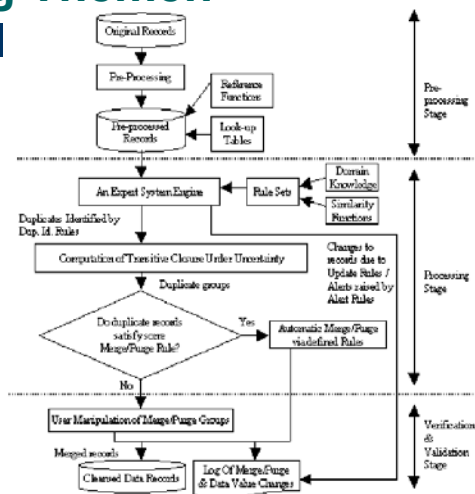
35



Felix I

Data Cleansing Themen

- Thema 6:
IntelliClean
- Recall/Precision
- Management von Domänen-Wissen
– Regeln



36



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Data Cleansing Themen

- Thema 7: Potter's Wheel ABC
- Interaktives System
- Cleansing-Schritte schnell und in GUI
- Cleansing-By-Example

The screenshot shows a software window with a menu bar (File, Cluster, Transform, Discrepancies, Sort) and a toolbar. Below is a table with columns: Delay, Carrier, Number, DMA, Day, Depart_Sch, Depart_Act, Arr_Sch, Arr_Act, Status, Random, (S)UID, (MergeID). The table contains flight data for carriers like DELTA, AMERIC, and UNITED. A dialog box titled 'Discrepancies found so far' is open, showing a list of discrepancies with columns for Value, Column, and Comments. Comments include 'compare 268 with mean 6.557, std dev. 27.894' and 'structure must be IspellWord'.

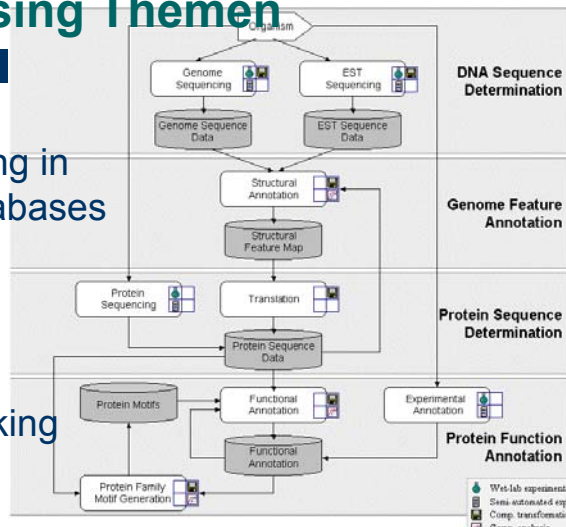
37



Felix Nauman

Data Cleansing Themen

- Thema 8: Data Cleansing in Genome Databases
- Fehlerquellen
- Fehlerarten
- Quality-Checking



38



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Data Cleansing Themen

- Thema 9: Data Lineage & Provenance
- Komplexe Transformationen in Data Warehouses
 - Aggregationen
 - Data Cleansing
- Problem: Zurückverfolgung der Werte zu Ihrer ursprünglichen Quelle
- Wie Speichern/Zugreifen?

39



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Data Cleansing Themen

- Thema 10: Object Fusion

Quelle 1 → $\begin{bmatrix} \underline{a}, b \\ \underline{a}, b \end{bmatrix} \Rightarrow a, b$

Quelle 2 → $\begin{bmatrix} \underline{a}, b \\ \underline{a}, b \end{bmatrix}$

$\begin{bmatrix} \underline{a}, b \\ \underline{a}, - \end{bmatrix} \Rightarrow a, b$

$\begin{bmatrix} \underline{a}, b \\ \underline{a}, c \end{bmatrix} \Rightarrow a, r(b,c)$

$\begin{bmatrix} \underline{a}, -, d \\ \underline{a}, c, - \end{bmatrix} \Rightarrow a, c, d$

40



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Literatur

- Data Cleaning: Problems and Current Approaches, Rahm & Do, IEEE Bulletin 23(4), 2000.
- Problems, Methods, and Challenges in Comprehensive Data Cleansing Heiko Müller, Johann-Christoph Freytag, Technical Report HUB-IB-164, Humboldt University Berlin, 2003

Sprechstunden

- Prof. Felix Naumann
 - Raum: IV.105
 - Sprechstunden: jederzeit, am besten n.V.
 - Email: naumann@informatik
 - Telefon: 2093 3905
- Mitarbeiter
 - Jens Bleiholder bleiho@informatik
 - Melanie Weis

Ankündigung für ein Seminar im WS03/04

Advanced Data Warehousing

Ulf Leser, Jörg Hakenberg

Wissensmanagement in der Bioinformatik

Termin: freitags, 13 – 15 Uhr, RUD26, 1'305

- Ausgewählte Themen aus dem Gebiet des Data Warehousing, außerdem Arbeiten aus den Gebieten Data Mining und Text Mining
- Fortgeschrittene Algorithmen zur effizienten Modellierung, Analyse und Bearbeitung großer Datenmengen
- Implementierung von SQL Operatoren CUBE/GROUPING. Caching und Chunking. Index Selection. Materialized View Selection. Association Rule Mining. Efficient k-Nearest Neighbor Searching. Searching text in relational databases: Containment Queries
- Text Classification. Text Clustering. Identification and Extraction of Biological Terms. Knowledge Discovery: Protein-Protein Interactions. Automatic Database Annotation

43



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003

Themenvergabe

Thema	Betreuer	Studenten
Ähnlichkeitsmaße	Felix Naumann	
Data Warehouse Duplicates	Felix Naumann	
Merge/Purge Algorithmus	Felix Naumann	
Domain-independent Dup. Detection	Felix Naumann	
AJAX framework	Felix Naumann	
IntelliClean	Felix Naumann	
Potter's Wheel ABC	Jens Bleiholder	
Data Cleansing in Genome Databases	Heiko Müller	
Data Lineage	Jens Bleiholder	
Object Fusion	Felix Naumann	

44



Felix Naumann, SE Data Cleansing, WS 03/04

22.10.2003