HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Informationsqualität

Antrittsvorlesung am 26.4.2007

Felix Naumann
Hasso-Plattner-Institut
Fachgebiet Informationssysteme

Überblick

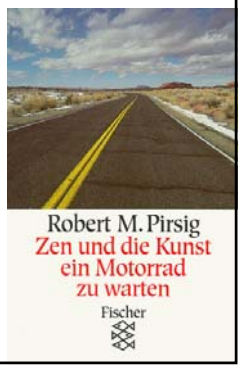
- Informationsqualität
- Informationsintegration
- Duplikaterkennung
  - Ähnlichkeit
  - Algorithmen
- Ausblick

Quality

3

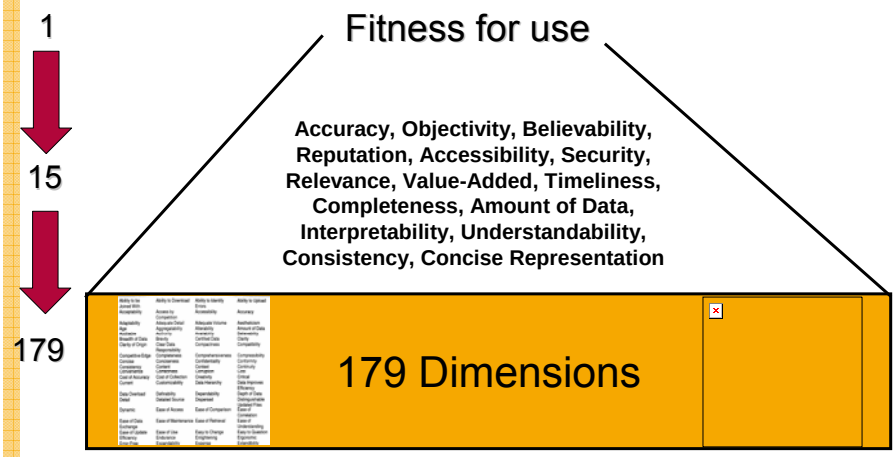
"Even though quality cannot be defined, you know what it is."

Robert Pirsig



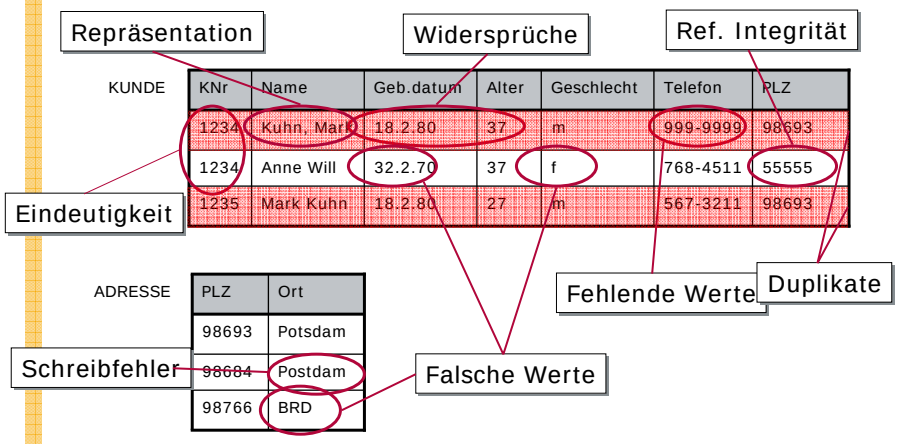
Zoom in die Informationsqualität

4



Datenqualität: Probleme

5



DQ-Probleme: Auswirkungen

6

- Fehlerhafte Warenpreise in Artikel-DB des US-Einzelhandels [English 1999]
  - Kosten für Konsumenten 2.5 Mrd $
  - 80% der Barcode-Scan-Fehler zulasten der Konsumenten
- US-Finanzbehörde 1992: knapp 100.000 Steuererstattungsbescheide nicht zustellbar [English 1999]
- 50-80% der Einträge im US-Vorstrafenregister ungenau, unvollständig oder fehlerhaft [Strong et al. 1997a]
- US-Post: von 100.000 Massen-Postsendungen bis zu 7.000 aufgrund vo Adressfehlern nicht zustellbar [Pierce 2004]

Goodyear reveals $100 million error 10/23/03 USA

Goodyear said late Wednesday that it will restate earnings for the past five years, decreasing income by as much as $100 million because an accounting system caused billing errors. The tiremaker is delaying the release of its third-quarter earnings, expected this morning, until mid-November. Shares closed up 2 cents to $6.83 before the announcement; in after-hours trading, shares plummeted 27%, or $1.83, to $5.

DQ-Probleme: Auswirkungen

- 2006 werden die Fortune 1000 Unternehmen mehr Geld wegen DQ-Problemen ausgeben als für ERP, CRM und BI zusammen [Gartner]
- Mehr als 35% aller IT-Projekte schlagen fehl aufgrund von DQ-Problemen; die jährlichen Kosten betragen allein in der USA 2-4 Mrd. $ [Meta Group]
- DQ ist einer der wichtigsten Erfolgsfaktoren in DWH- und CRM-Projekten [PriceWaterhouseCoopers]
- Datensammlungen in der Folge des Tsunami 2004
  - Todesfälle und Verletzungen
  - Zerstörung von Häusern und Eigentum
  - Zuweisung von Hilfsmitteln und Unterstützung

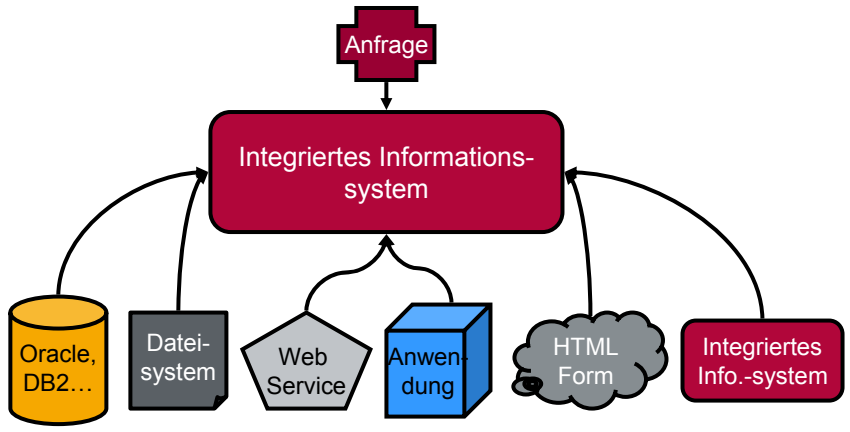
Quelle: http://www.informationquality.org/publiclyexposediqproblems.cfm

Überblick

- Informationsqualität
- Informationsintegration
- Duplikaterkennung
  - Ähnlichkeit
  - Algorithmen
- Ausblick

Integrierte Informationssysteme

9



Informationsintegration

10

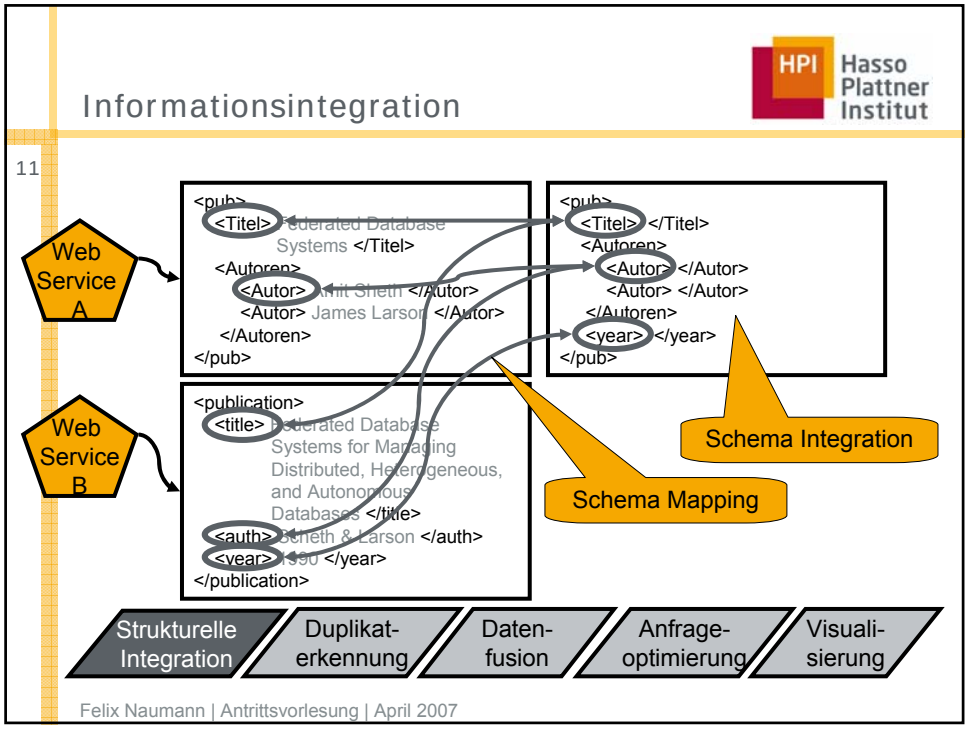
Web Service A

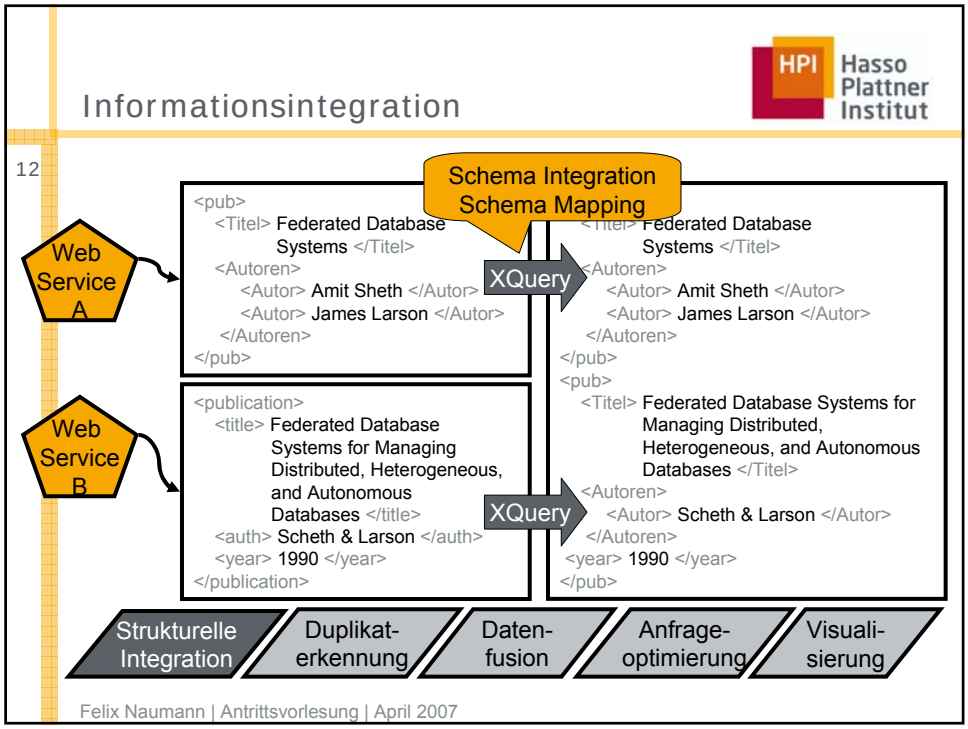
```
<pub>
   <Titel> Federated Database
           Systems </Titel>
   <Autoren>
      <Autor> Amit Sheth </Autor>
      <Autor> James Larson </Autor>
   </Autoren>
</pub>
```

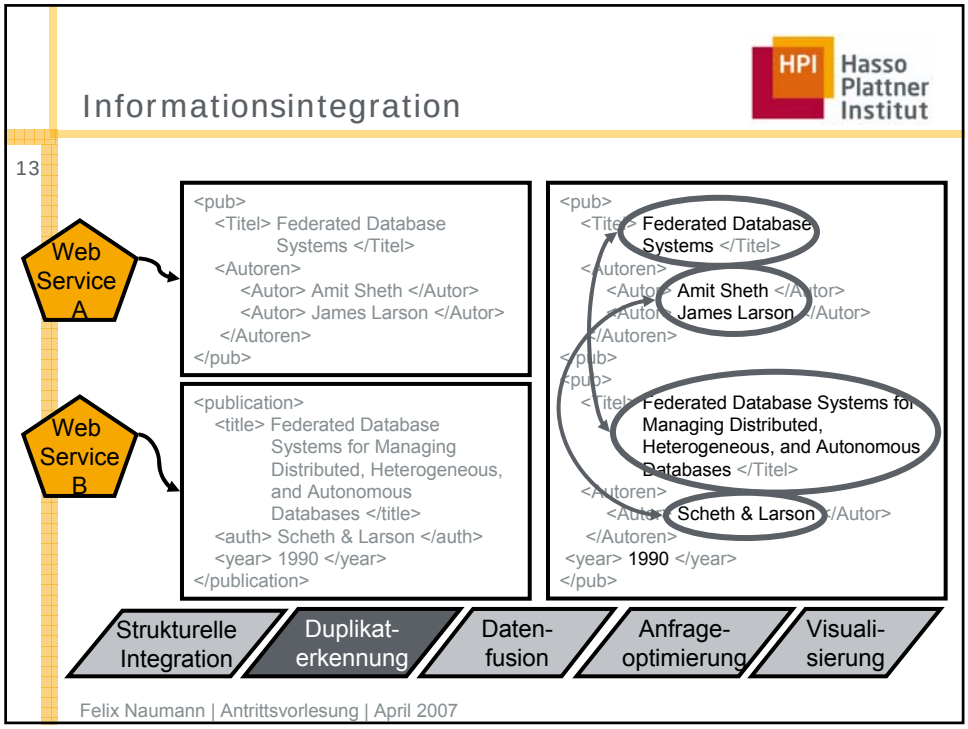
Web Service B

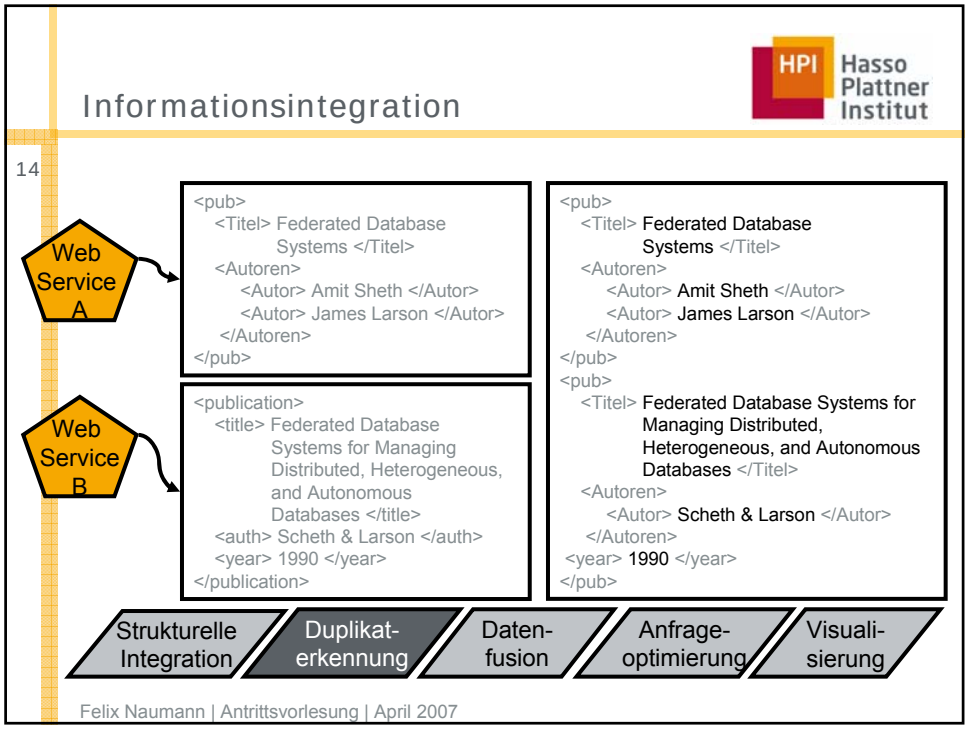
```
<publication>
   <title> Federated Database
           Systems for Managing
           Distributed, Heterogeneous,
           and Autonomous
           Databases </title>
   <auth> Scheth & Larson </auth>
   <year> 1990 </year>
</publication>
```

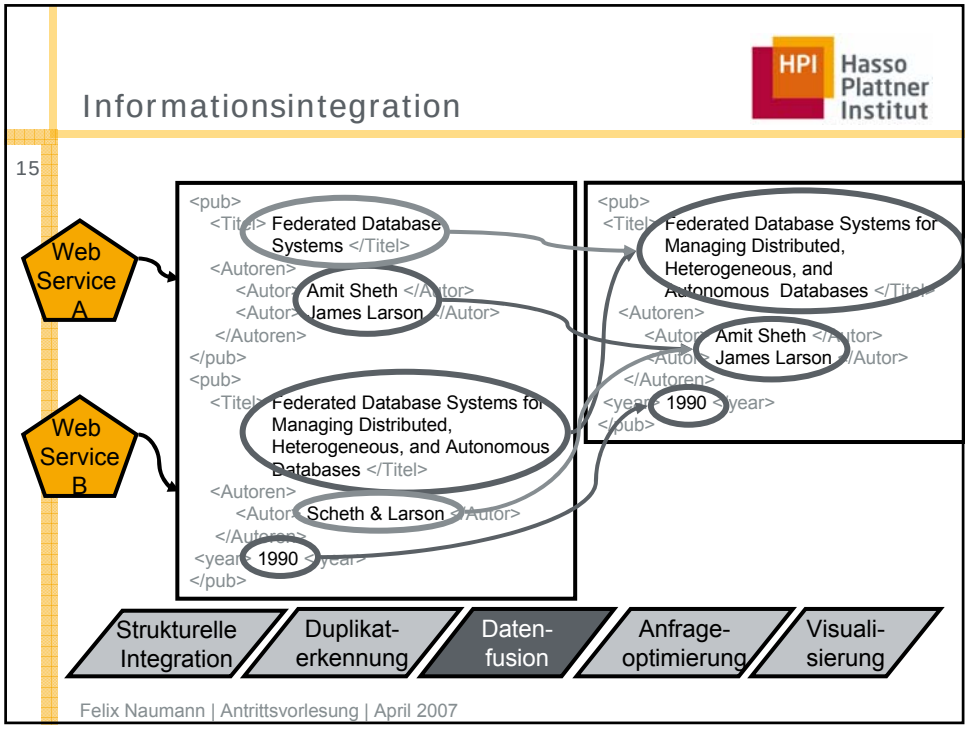
Strukturelle Integration | Duplikat-erkennung | Daten-fusion | Anfrage-optimierung | Visuali-sierung

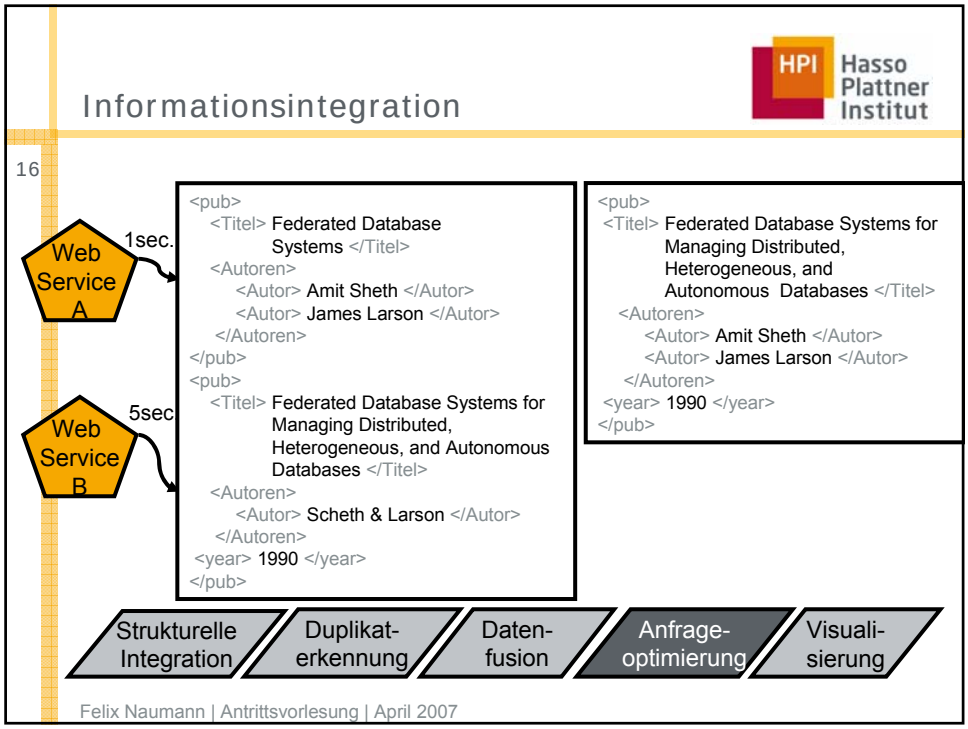

Informationsintegration
HPI Hasso Plattner Institut
11
Web Service A
Web Service B
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
<publication>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
<auth> Scheth & Larson </auth>
<year> 1990 </year>
</publication>
<pub>
<Titel> </Titel>
<Autoren>
<Autor> </Autor>
<Autor> </Autor>
</Autoren>
<year> </year>
</pub>
Schema Integration
Schema Mapping
Strukturelle Integration
Duplikat-erkennung
Daten-fusion
Anfrage-optimierung
Visuali-sierung
Felix Naumann | Antrittsvorlesung | April 2007


Informationsintegration
HPI Hasso Plattner Institut
12
Schema Integration
Schema Mapping
Web Service A
Web Service B
XQuery
XQuery
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
<publication>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
<auth> Scheth & Larson </auth>
<year> 1990 </year>
</publication>
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
<pub>
<Titel> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </Titel>
<Autoren>
<Autor> Scheth & Larson </Autor>
</Autoren>
<year> 1990 </year>
</pub>
Strukturelle Integration
Duplikat-erkennung
Daten-fusion
Anfrage-optimierung
Visuali-sierung
Felix Naumann | Antrittsvorlesung | April 2007


Informationsintegration
HPI Hasso Plattner Institut
13
Web Service A
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
Web Service B
<publication>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
<auth> Scheth & Larson </auth>
<year> 1990 </year>
</publication>
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
<pub>
<Titel> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </Titel>
<Autoren>
<Autor> Scheth & Larson </Autor>
</Autoren>
<year> 1990 </year>
</pub>
Strukturelle Integration
Duplikat-erkennung
Daten-fusion
Anfrage-optimierung
Visuali-sierung
Felix Naumann | Antrittsvorlesung | April 2007

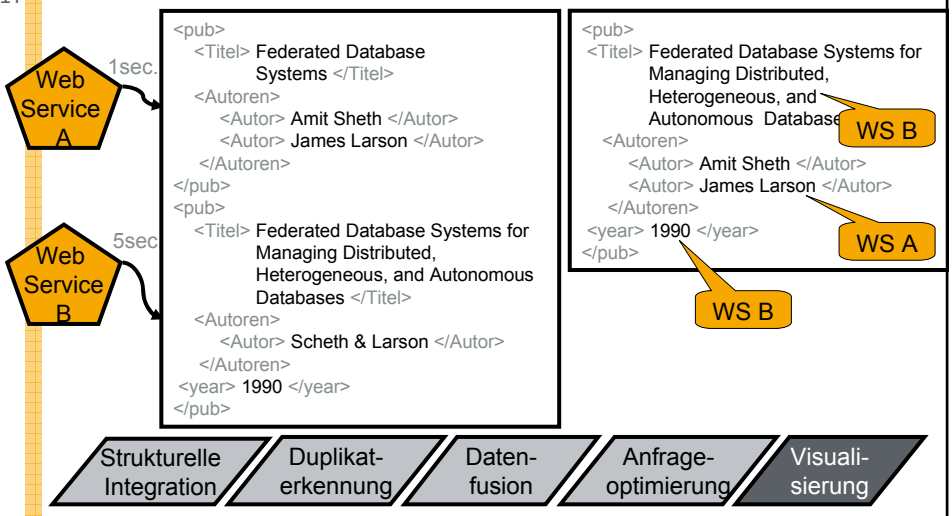

Informationsintegration
HPI Hasso Plattner Institut
14
Web Service A
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
Web Service B
<publication>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
<auth> Scheth & Larson </auth>
<year> 1990 </year>
</publication>
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
<pub>
<Titel> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </Titel>
<Autoren>
<Autor> Scheth & Larson </Autor>
</Autoren>
<year> 1990 </year>
</pub>
Strukturelle Integration
Duplikat-erkennung
Daten-fusion
Anfrage-optimierung
Visuali-sierung
Felix Naumann | Antrittsvorlesung | April 2007


Informationsintegration
HPI Hasso Plattner Institut
15
Web Service A
Web Service B
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
<pub>
<Titel> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </Titel>
<Autoren>
<Autor> Scheth & Larson </Autor>
</Autoren>
<year> 1990 </year>
</pub>
<pub>
<Titel> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
<year> 1990 </year>
</pub>
Strukturelle Integration
Duplikat-erkennung
Daten-fusion
Anfrage-optimierung
Visuali-sierung
Felix Naumann | Antrittsvorlesung | April 2007


Informationsintegration
HPI Hasso Plattner Institut
16
Web Service A
1sec.
Web Service B
5sec
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
<pub>
<Titel> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </Titel>
<Autoren>
<Autor> Scheth & Larson </Autor>
</Autoren>
<year> 1990 </year>
</pub>
<pub>
<Titel> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
<year> 1990 </year>
</pub>
Strukturelle Integration
Duplikat-erkennung
Daten-fusion
Anfrage-optimierung
Visuali-sierung
Felix Naumann | Antrittsvorlesung | April 2007

Informationsintegration

17



Überblick

18

- Informationsqualität
- Informationsintegration
- Duplikaterkennung
  - Ähnlichkeit
  - Algorithmen
- Ausblick

Duplikaterkennung

19

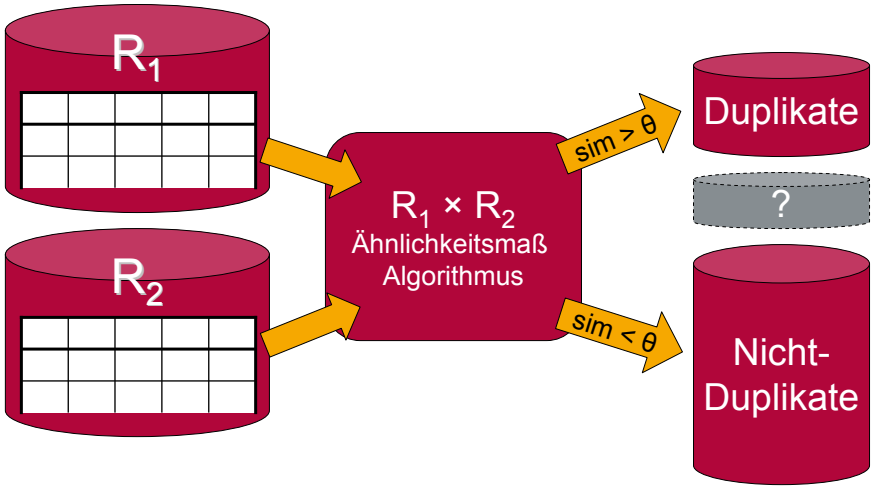
Duplikaterkennung ist das Finden mehrerer Repräsentationen desselben Realweltobjekts.

- Problem 1: Repräsentationen sind nicht identisch.
  - *Fuzzy duplicates*
- Lösung: Ähnlichkeitsmaße
  - Wert- und Datensatzvergleiche
  - Domänenunabhängig oder -abhängig

- Problem 2: Die Datenmenge ist groß.
  - Quadratischer Aufwand: Jedes Paar muss verglichen werden.
- Lösung: Algorithmen
  - Z.B. Vergleiche durch Partitionierung vermeiden

Duplikaterkennung

20



Wirkungen von Duplikaten

21

- Mehrfache Zusendung von Katalogen
- Rechnungen werden doppelt bezahlt
- Banken
  - Überschreiten des Kreditlimits wird nicht erkannt
- Lagerhaltung / Einkauf
  - Zu niedriger Lagerbestand einzelner Waren wird ausgewiesen.
  - Kein Ausnutzen von Mengenrabatten bei Bestellungen
- Gesamtumsatz eines Kunden bleibt unbekannt.
- Mehraufwand in der IT
- Sinkende Kundenzufriedenheit
- Potenziale und Gefahren nicht erkannt
- Inkorrekte Kennzahlen

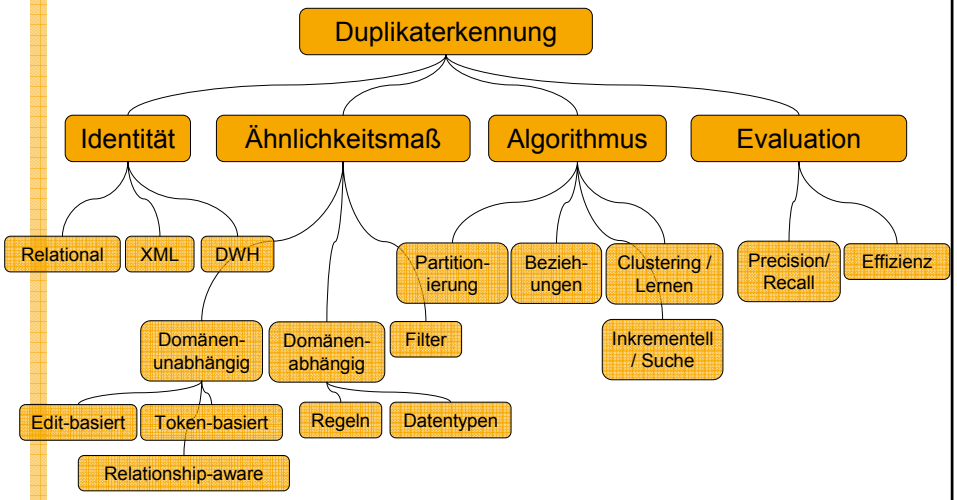
Kunde	Umsatz
BMW	20.000
BaMoWe	5.000.000
Bayerische Motorenwerke	300.000
...	...

"Duplikaterkennung" hat viele Duplikate

22

Household matching
Dubletten
Duplicate detection
Mixed and split citation problem
Record linkage
Object identification
Match
Deduplication
Object consolidation
Fuzzy match
Entity resolution
Entity clustering
Approximate match
Identity uncertainty
Reference reconciliation
Merge/purge
Hardening soft databases
Householding
Reference matching

Duplikaterkennung



Überblick

- Informationsqualität
- Informationsintegration
- Duplikaterkennung
  - Ähnlichkeit
  - Algorithmen
- Ausblick

Token-basierte Ähnlichkeitsmaße

25

- Tokens
  - Words / Terms
  - n-grams
- Jaccard
  - |{gemeinsame token}| / |{alle token}|
- TFIDF
  - *Term frequency* (tf)
  - *Inverse document frequency* (idf)
  - TFIDF: log (tf+1) x log idf
  - Häufige Wörter haben niedriges Gewicht.
  - Ähnlichkeit ist Kosinus der Termvektoren, gewichtet durch TFIDF.
- ...

Edit-basierte Ähnlichkeitsmaße

26

- Edit-Distanz / Levenshtein-Distanz [Levenshtein 1965]
  - Minimale Anzahl von Edit-Operationen um den einen String in den anderen umzuwandeln.
  - Domänenspezifische Kosten
- Jaro [Jaro 1989] / Jaro-Winkler [Winkler 1999]
  - Common letters within ½ string length
  - Transposed letters
- Soundex
  - 4-stelliger Code für jedes Wort
  - SOUNDEX('Farwick ') = F620
- ...

Frass, Fricke, Fahruschi, Feuerhake

Domänenabhängige Ähnlichkeitsmaße

Datentypen

- Spezielle Ähnlichkeitsmaße für Datumsangaben
- Spezielle Ähnlichkeitsmaße für numerische Attribute
- ...

Regeln

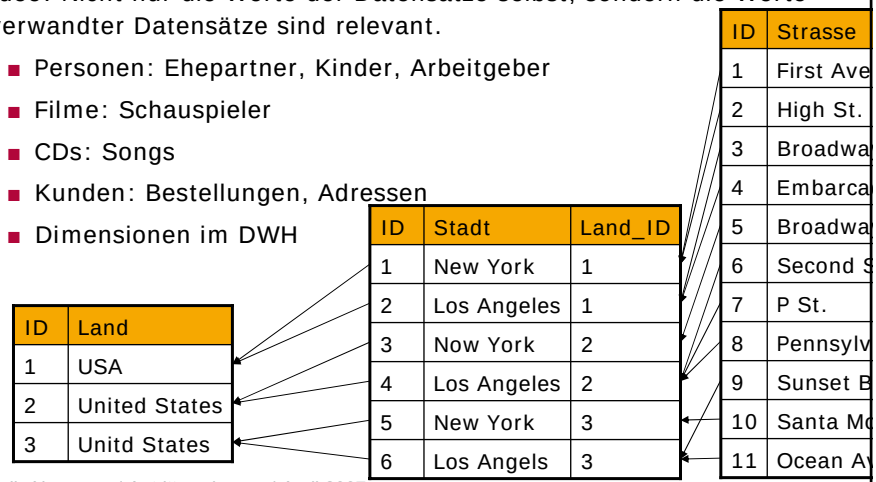
- [Hernandez Stolfo 1998], [Lee et al. 2000]
- Gegeben zwei Datensätze r1 und r2:

```
IF  last name of r1 = last name of r2,
AND first names differ slightly,
AND address of r1 = address of r2
THEN r1 is equivalent to r2
```


Beziehungs-basierte Ähnlichkeitsmaße

Idee: Nicht nur die Werte der Datensätze selbst, sondern die Werte verwandter Datensätze sind relevant.

- Personen: Ehepartner, Kinder, Arbeitgeber
- Filme: Schauspieler
- CDs: Songs
- Kunden: Bestellungen, Adressen
- Dimensionen im DWH



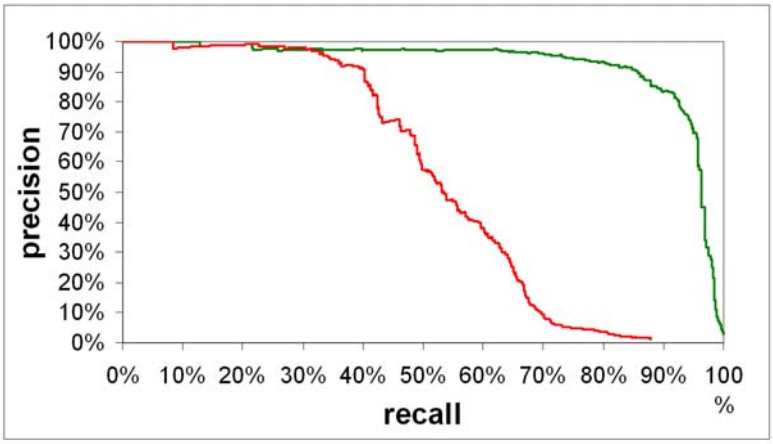
ID	Land
1	USA
2	United States
3	Unitd States

ID	Stadt	Land_ID
1	New York	1
2	Los Angeles	1
3	Now York	2
4	Los Angeles	2
5	New York	3
6	Los Angels	3

ID	Strasse
1	First Ave
2	High St.
3	Broadwa
4	Embarca
5	Broadwa
6	Second S
7	P St.
8	Pennsylv
9	Sunset B
10	Santa Mo
11	Ocean Av

Beziehungs-basierte Ähnlichkeitsmaße – Evaluation

29



Überblick

30

- Informationsqualität
- Informationsintegration
- Duplikaterkennung
  - Ähnlichkeit
  - Algorithmen
- Ausblick

Komplexität

31

Problem: Zu viele Vergleiche

- 10.000 Kunden => 49.995.000 Vergleiche
  - $(n^2 - n) / 2$
  - Jeder Vergleich ist teuer.

Idee: Vergleiche vermeiden

- Durch Herausfiltern einzelner Datensätze
- Durch Partitionierung (Heuristiken)

Partitionierung

32

Idee: Partitioniere Datensätze und vergleiche Paare nur innerhalb der Partition

- Partitionierung nach ersten zwei Ziffern der PLZ
  - Ca. 100 Partitionen
  - Ca. 100 Kunden / Partition
  - Insgesamt 495.000 Vergleiche
- Partitionierung nach erstem Buchstaben des Nachnamens
- ...

Quelle: wikipedia.de

Idee: Partitioniere mehrfach nach unterschiedlichen Kriterien

- Transitive Hülle

Sorted Neighborhood [Hernandez Stolfo 1998]

33

Sortierte Nachbarschaft

- Sortiere Datensätze so, dass ähnliche Datensätze nah beieinander liegen.
- Vergleiche nur Datensätze, die innerhalb einer Nachbarschaft liegen.

Algorithmus

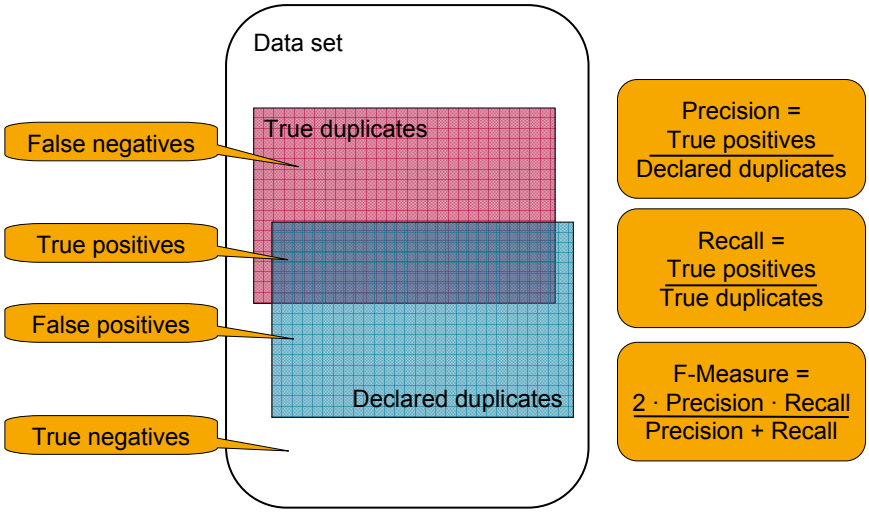
1. Sortierschlüssel erzeugen
   - Wahl des Schlüssels
2. Sortieren
3. Fenster über sortierte Liste schieben und nur innerhalb des Fensters vergleichen
   - Wahl der Fenstergröße
4. Transitive Hülle bilden

Komplexität ist I/O-bound

- 3 Durchläufe durch die Daten

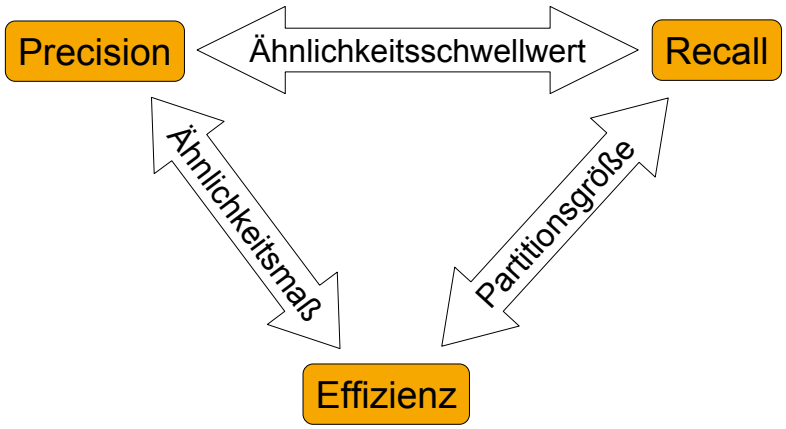
Precision & Recall

38



Duplikaterkennung – Zielkonflikte

39



Zurück zur Informationsqualität

40

Duplikatfreiheit ist nur eine Dimension der Informationsqualität.

- Weitere „harte“ Qualitätsdimensionen
  - Vollständigkeit (*completeness*)
  - Genauigkeit (*accuracy*)
  - Aktualität (*timeliness*)
- Herausforderungen
  - Messung
  - Aggregation
  - Verwendung

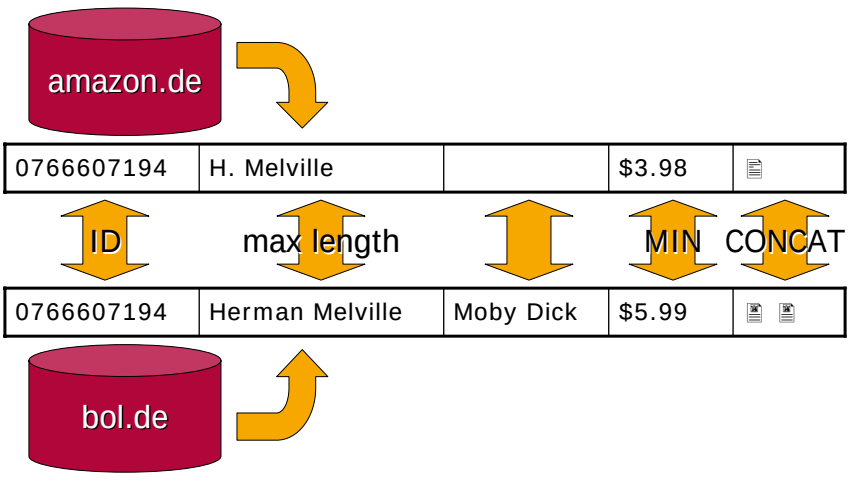
Überblick

41

- Informationsqualität
- Informationsintegration
- Duplikaterkennung
  - Ähnlichkeit
  - Algorithmen
- Ausblick

Datenfusion

42



Die Arbeitsgruppe Informationssysteme

- Wissenschaftliche Mitarbeiter / Doktoranden
  - Alexander Albrecht: Personal Information Management
  - Jana Bauckmann: Data Profiling, Aladin
  - Jens Bleiholder: Data Fusion, HumMer & FuSem
  - Frank Kaufer: Matching, Forschungskolleg
  - Armin Roth: Peer-Daten-Management, System P
  - Melanie Weis: Duplicate Detection
- Studentische Hilfskräfte
  - Karsten Draba: HumMer
  - Christoph Böhm: Ranking, SPRINT
  - NN: Aladin Projekt
  - Matthias Weidlich: System P
- http://www.hpi.uni-potsdam.de/~naumann/