HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Duplicate Detection

Keynote presentation at AusIQ Conference 2007

Felix Naumann
Hasso Plattner Institute, Potsdam, Germany
Information Systems
naumann@hpi.uni-potsdam.de

The HPI – Hasso Plattner Institut

HPI Hasso Plattner Institut

- Founded in 1998 as a Public Private Partnership
- Hasso Plattner, co-founder of SAP, endowed over 200 Mio. Euro.
- Adjoined with the University of Potsdam
  - Capital of Brandenburg, near Berlin

The Information Systems Group

3

- Research assistants / PhD students
  - Alexander Albrecht: ETL, Schema Mapping, PIM
  - Jana Bauckmann: Data Profiling, Aladin
  - Jens Bleiholder: Data Fusion, HumMer & FuSem
  - Paul Führing: DQ Assessment, Viqtor
  - Frank Kaufer: Schema and Ontology Matching
  - Armin Roth: Peer-Daten-Management, System P
  - Melanie Weis: Duplicate Detection
- Student assistants
  - Karsten Draba: HumMer
  - Christoph Böhm: Ranking, SPRINT
  - Tobias Flach: Aladin Projekt
  - Matthias Weidlich: System P
- http://www.hpi.uni-potsdam.de/naumann/

Overview

4

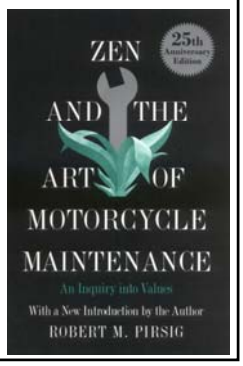
- Information Quality
- Information Integration
- Duplicate Detection
  - Similarity
  - Algorithms
  - Goals
- Data Fusion
- Outlook

Quality

5

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

Robert Pirsig



Zooming into Information Quality

6

1

15

179

Fitness for use

Accuracy, Objectivity, Believability, Reputation, Accessibility, Security, Relevance, Value-Added, Timeliness, Completeness, Amount of Data, Interpretability, Understandability, Consistency, Concise Representation

179 Dimensions

Data Quality: Problems

7

Representation

Contradictions

Ref. integrity

CUST

KNr	Name	Birthday	Age	Sex	Phone	ZIP
1234	Kuhn, Mark	18.2.80	37	m	999-9999	98693
1234	Anne Will	32.2.70	37	f	768-4511	55555
1235	Mark Kuhn	18.2.80	27	m	567-3211	98693

Uniqueness

ADDRESS

ZIP	Place
98693	Potsdam
98684	Postdam
98766	Germany

Missing values

Duplicates

Typos

Incorrect values

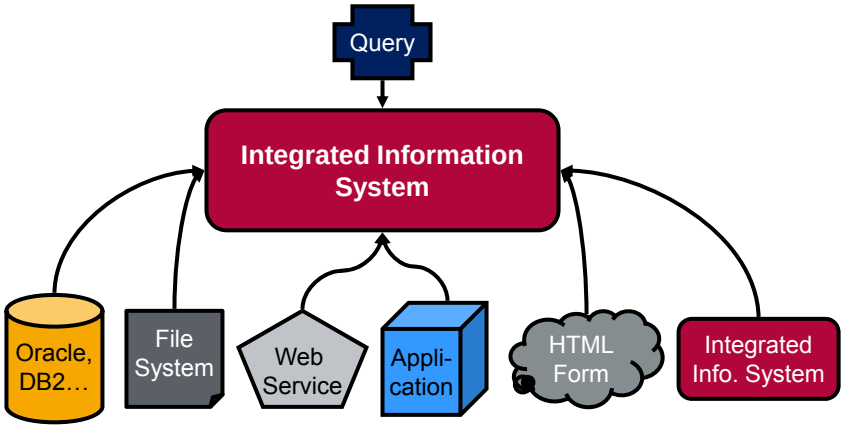
Overview

8

- Information Quality
- Information Integration
- Duplicate Detection
  - Similarity
  - Algorithms
  - Goals
- Data Fusion
- Outlook

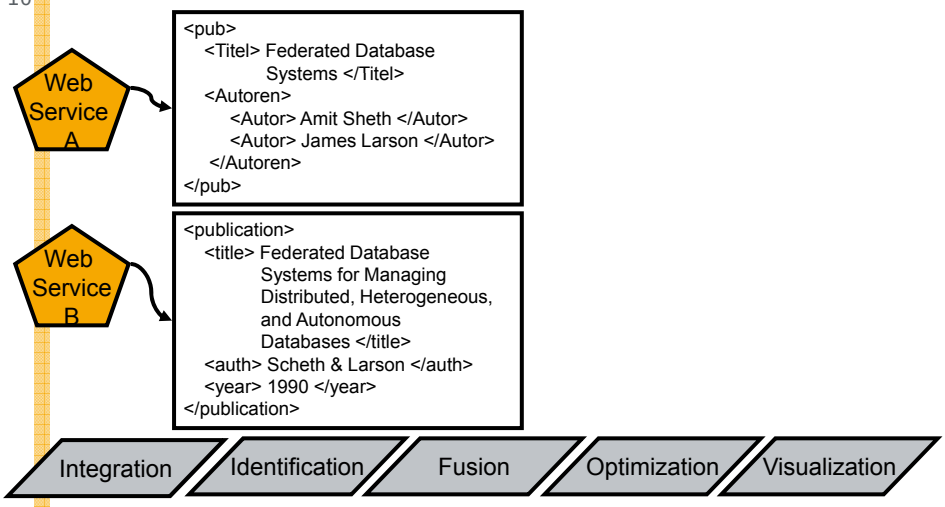
Integrated Information Systems

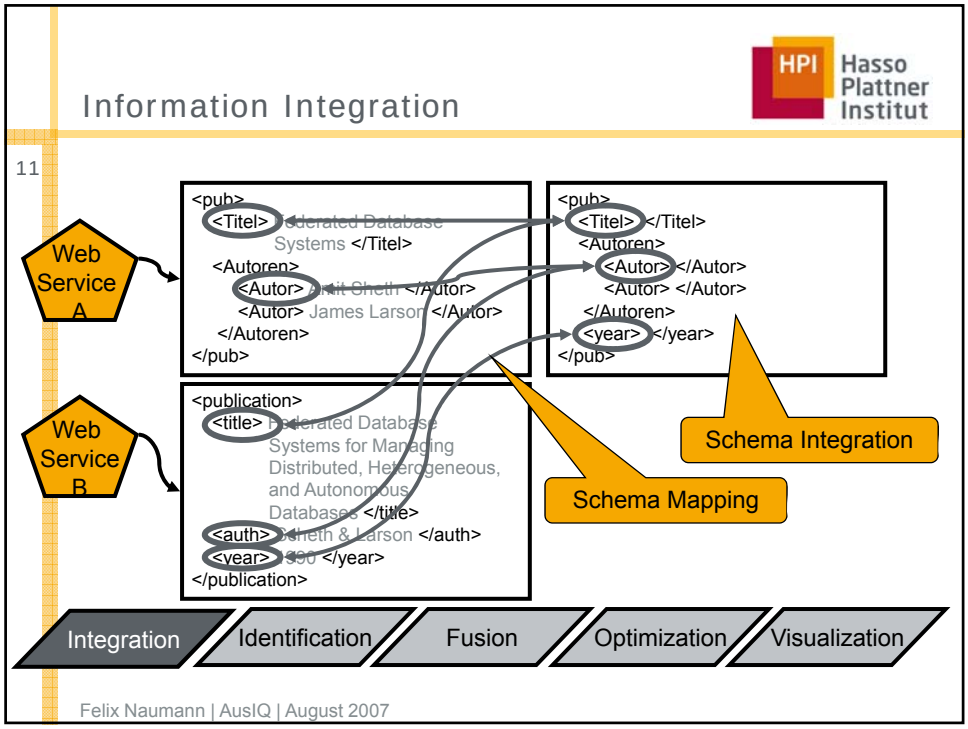
9

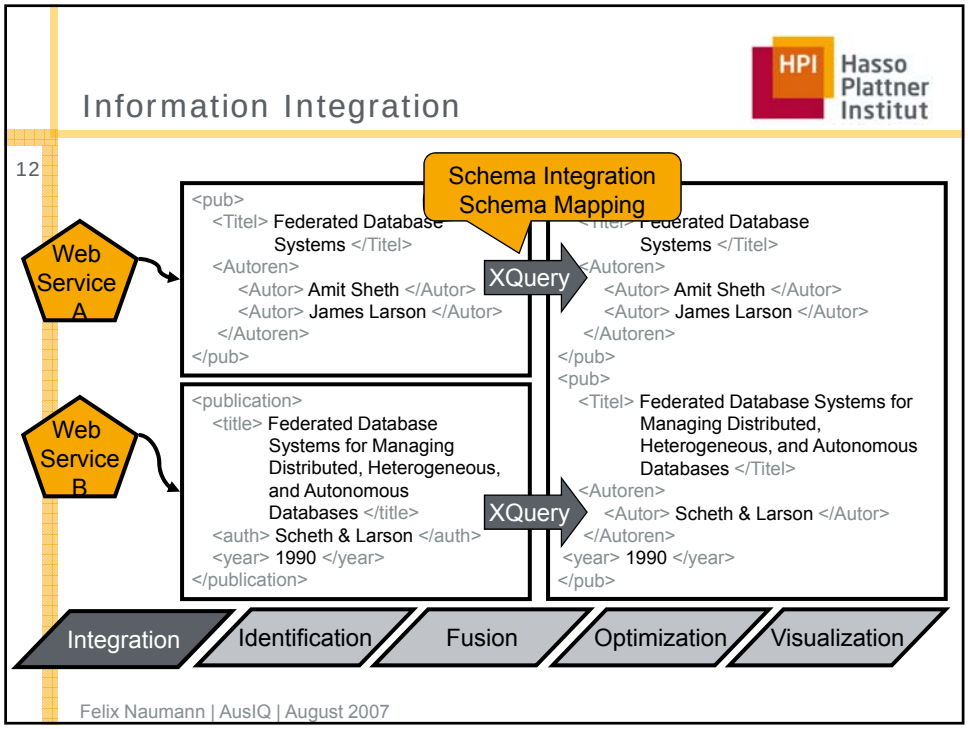


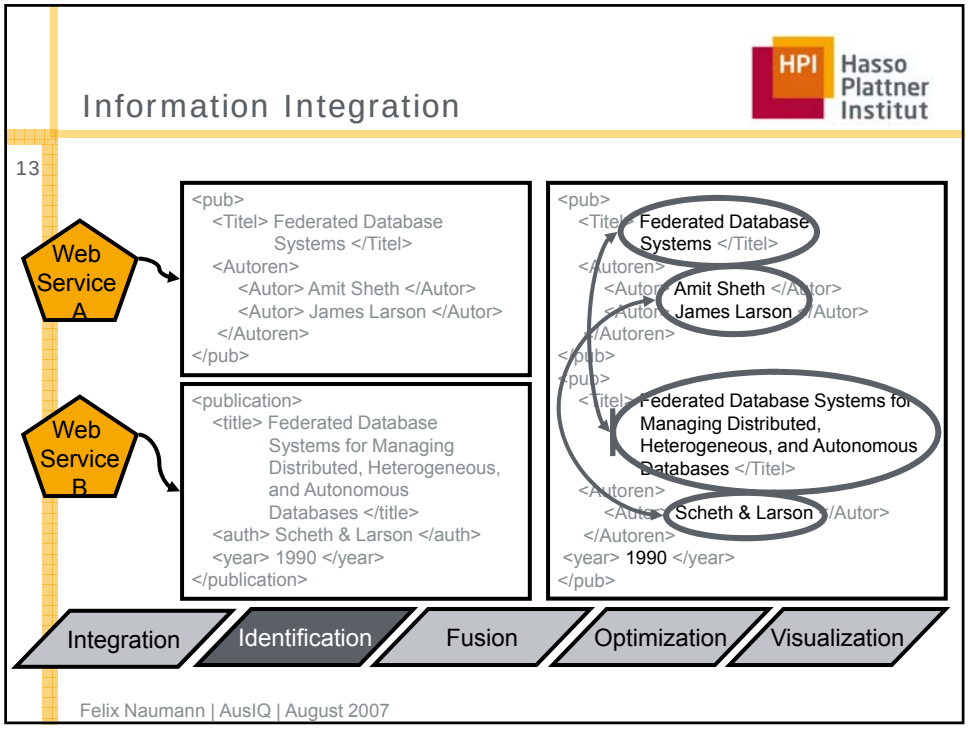
Information Integration

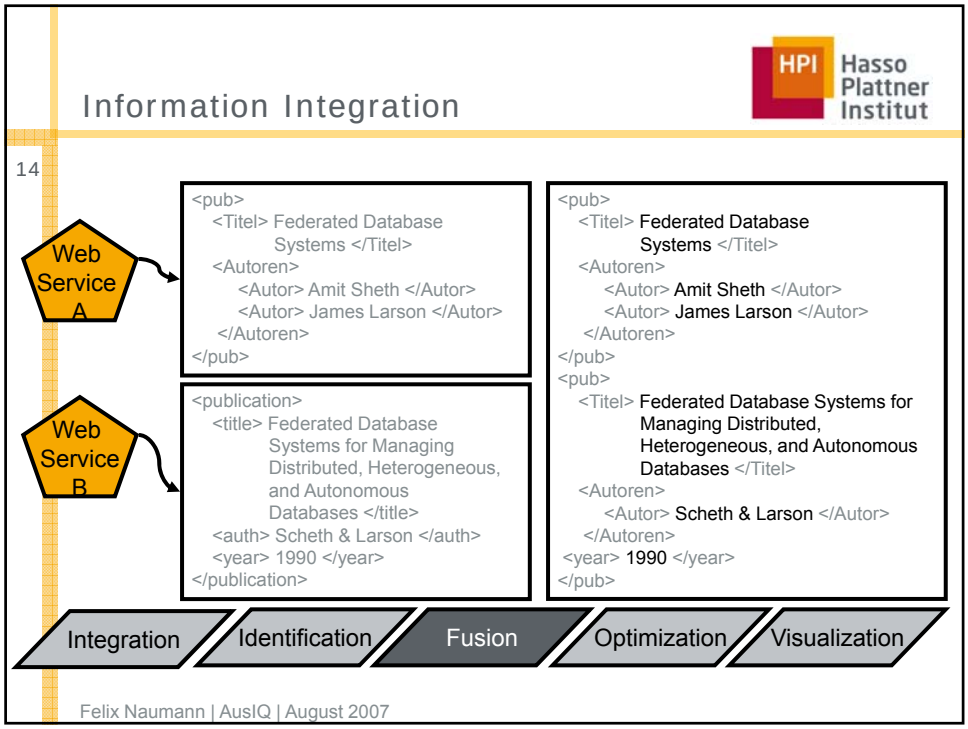
10

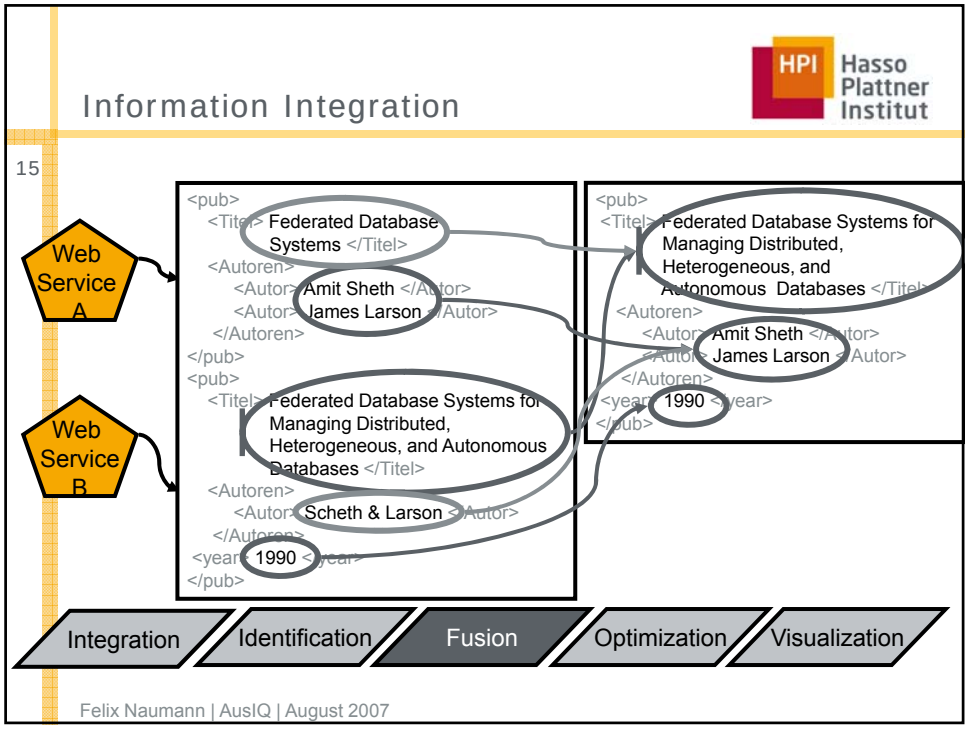


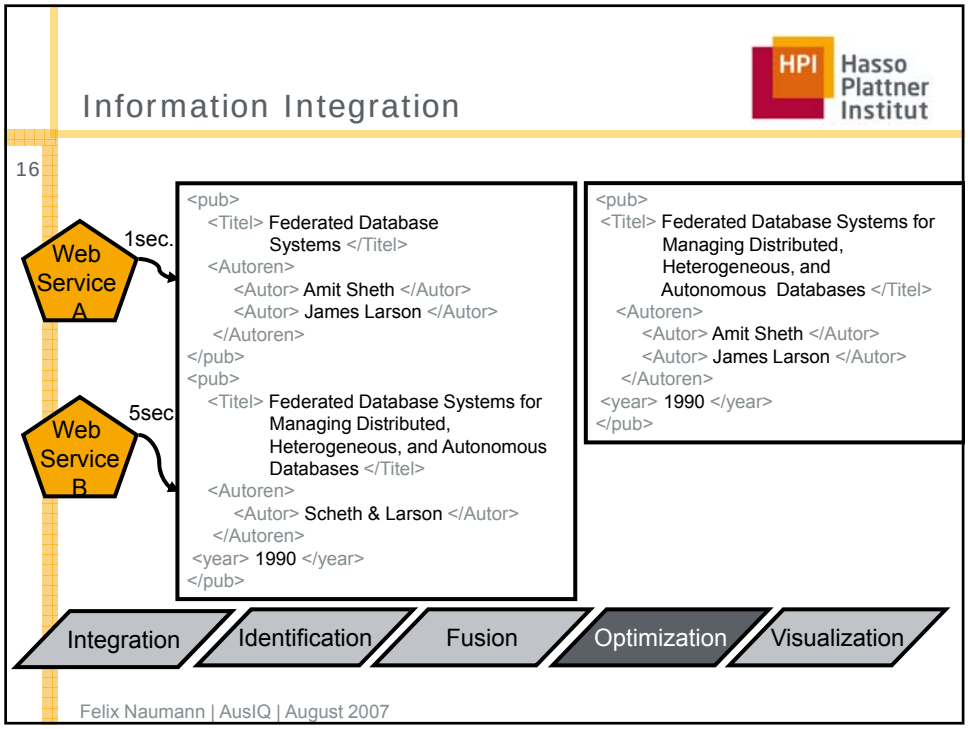

Information Integration
HPI Hasso Plattner Institut
11
Web Service A
Web Service B
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
<publication>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, 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
Integration
Identification
Fusion
Optimization
Visualization
Felix Naumann | AusIQ | August 2007


Information Integration
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>
Integration
Identification
Fusion
Optimization
Visualization
Felix Naumann | AusIQ | August 2007


Information Integration
HPI Hasso Plattner Institut
13
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>
Federated Database Systems
Amit Sheth
James Larson
Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases
Scheth & Larson
<year> 1990 </year>
Integration
Identification
Fusion
Optimization
Visualization
Felix Naumann | AusIQ | August 2007

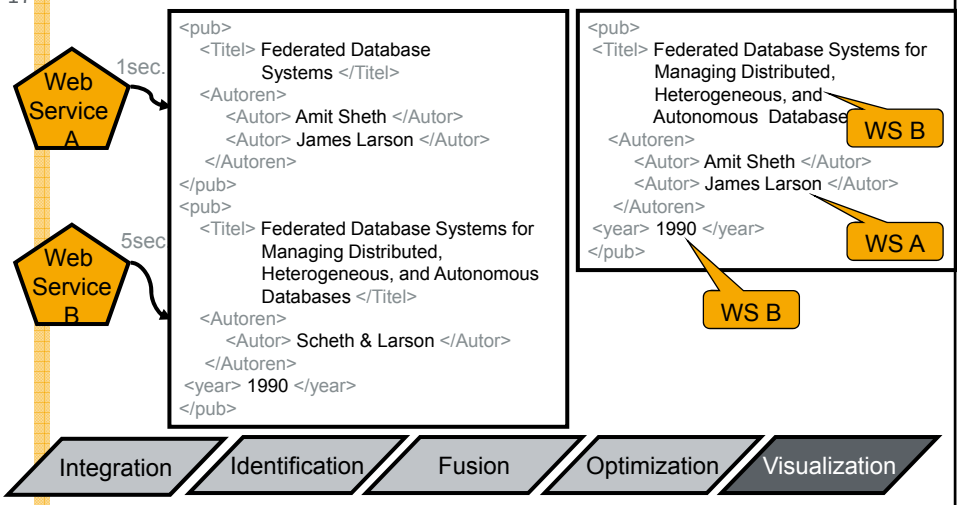

Information Integration
HPI Hasso Plattner Institut
14
Web Service A
Web Service B
<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>
<publication>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
<auth> Scheth & Larson </auth>
<year> 1990 </year>
</publication>
<Titel> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </Titel>
<Autor> Scheth & Larson </Autor>
<year> 1990 </year>
Integration
Identification
Fusion
Optimization
Visualization
Felix Naumann | AusIQ | August 2007


Information Integration
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>
Integration
Identification
Fusion
Optimization
Visualization
Felix Naumann | AusIQ | August 2007


Information Integration
HPI Hasso Plattner Institut
16
1sec.
Web Service A
5sec
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>
Integration
Identification
Fusion
Optimization
Visualization
Felix Naumann | AusIQ | August 2007

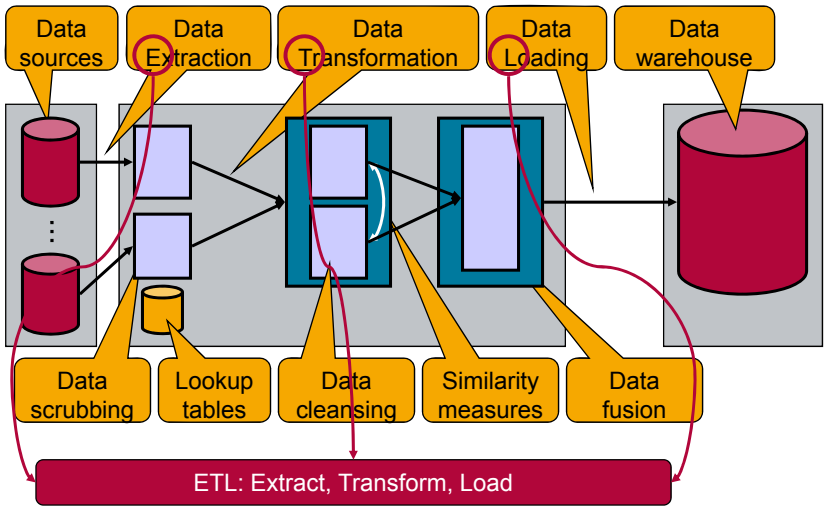
Information Integration

17



Integration and Cleansing: ETL

18



Overview

19

- Information Quality
- Information Integration
- Duplicate Detection
  - Similarity
  - Algorithms
  - Goals
- Data Fusion
- Outlook

Duplicate Detection

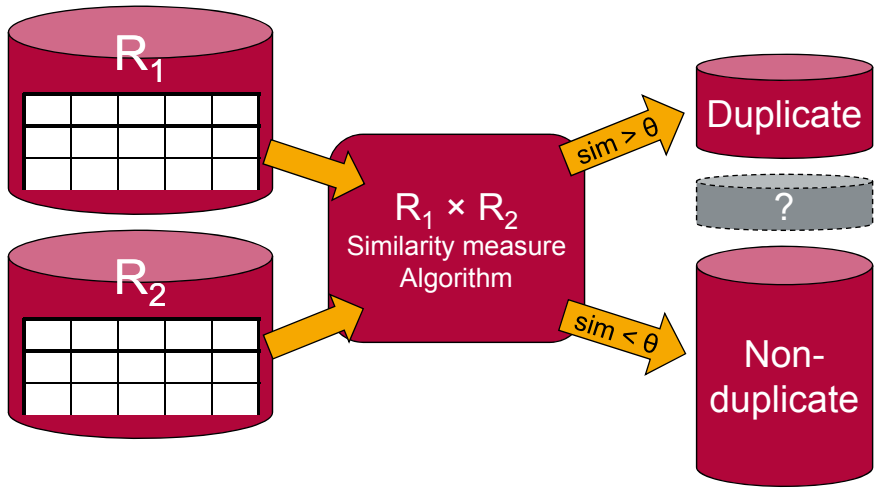
20

Duplicate detection is the discovery of multiple representations of the same real-world object.

- Problem 1: Representations are not identical.
  - *Fuzzy duplicates*
- Solution: Similarity measures
  - Value- and record-comparisons
  - Domain-dependent or domain-independent
- Problem 2: Data sets are large.
  - Quadratic complexity: Comparison of every pair of records.
- Solution: Algorithms
  - E.g., avoid comparisons by partitioning.

Duplicate Detection

21



Motivation

22

Customer	Revenue
BMW	20.000
BaMoWe	5.000.000
Bayerische Motorenwerke	300.000
...	...

- Possible effects
  - Example: Portfolio Management Offers
  - Credit maximum not detected
  - Too low inventory levels
  - No quantity discount for multiple orders
  - Total revenue of preferred customers unknown
  - Multiple mailings of same catalog to same household
- General problems
  - Additional, unnecessary IT expenses
  - Low customer satisfaction
  - Potentials and dangers not detected
  - Poor quality financial data

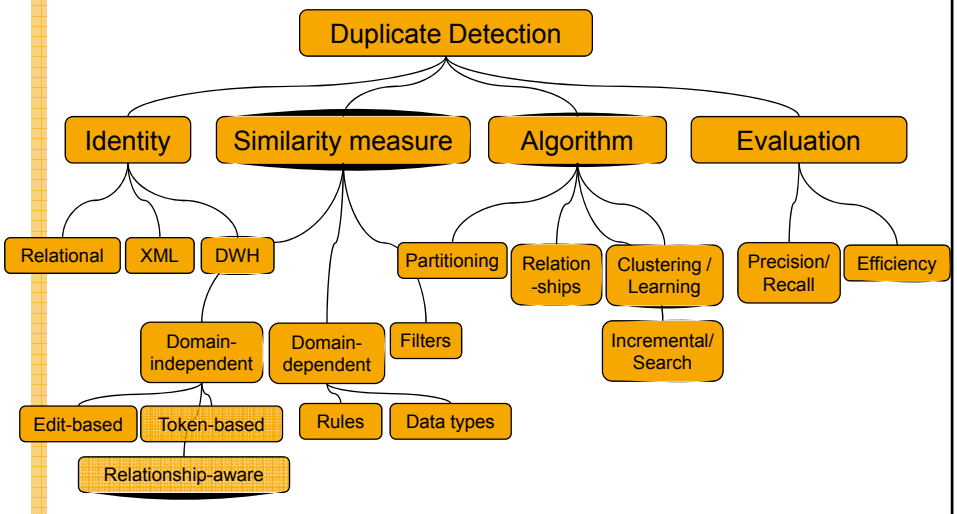
Ironically, “Duplicate Detection” has many Duplicates

23

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

Duplicate Detection – Research

24



Overview

26

- Information Quality
- Information Integration
- Duplicate Detection
  - Similarity
  - Algorithms
  - Goals
- Data Fusion
- Outlook

Token-based Similarity Measures

27

- Tokens
  - Words / Terms
  - n-grams
- Jaccard
  - |{common tokens}| / |{all tokens}|
- TFIDF [Cohen et al. 2003]
  - Term frequency: tf
  - Inverse document frequency: idf
  - TFIDF: log (tf+1) x log (idf)
  - Common words have low weight
  - Cosine similarity of term vectors weighted by tfidf
- And many more
  [Koudas Srivastavasa 2005]

Edit-based Similarity Measures

28

- Jaro [Jaro 1989] / Jaro-Winkler [Winkler 1999]
  - Common letters within ½ string length
  - Transposed letters
- Edit-distance / Levenshtein-distance [Levenshtein 1965]
  - Minimum number of edits from one word to the other
  - Domain-specific costing
  - Dynamic Programming
- Soundex
  - 4-letter code for each word
  - SOUNDEX('Farwick ') = F620
- ...

Frass, Fricke, Fahruschi, Feuerhake

Domain-dependent Similarity Measures

29

- Data Types
  - Special similarity for dates
  - Special similarity for numerical attributes
  - ...
- Rules
  - [Hernandez Stolfo 1998], [Lee et al. 2000]
  - **Given two records, r1 and 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.

Relationship-aware Similarity Measures

30

Idea: Not only values of the records, but values of related records are relevant for similarity.

- Persons: spouse, children, employer
- Movies: actors
- CDs: songs
- Customers: orders, addresses
- Dimensions in a DWH [Ananthakrishna et al. 2002]

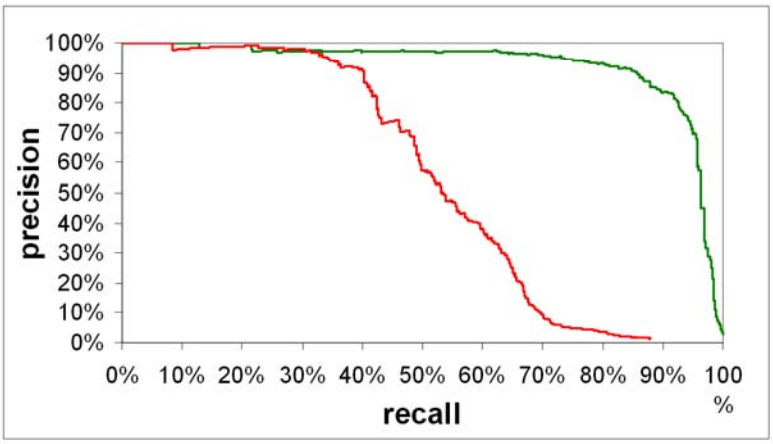
ID	Country
1	USA
2	United States
3	Unitd States

ID	City	Country
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	Street
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

Relationship-aware Similarity Measures – Evaluation

31



Overview

32

- Information Quality
- Information Integration
- Duplicate Detection
  - Similarity
  - Algorithms
  - Goals
- Data Fusion
- Outlook

Complexity

33

- Problem: Too many comparisons!
  - 10.000 customers => 49.995.000 comparisons
    - $(n^2 - n) / 2$
    - Each comparison is already expensive.
- Idea: Avoid comparisons…
  - … by filtering out individual records.
  - … by partitioning the records and comparing only within a partition.

Partitioning / Blocking

34

- Partition the records (horizontally) and compare pairs of records only within a partition.
  - Partitioning by first two zip-digits
    - Ca. 100 partitions in Germany
    - Ca. 100 customers per partition
    - => 495.000 comparisons
  - Partition by first letter of surname
  - ...

Source: wikipedia.de

- Idea: Partition multiple times by different criteria.
  - Then apply transitive closure on discovered duplicates.

Sorted Neighborhood
[Hernandez Stolfo 1998]

35

- Idea
  - Sort tuples so that similar tuples are close to each other.
  - Only compare tuples within a small neighborhood (window).

1. Generate key
   - E.g.: SSN+“first 3 letters of name“ + ...
2. Sort by key
   - Similar tuples end up close to each other.
3. Slide window over sorted tuples
   - Compare all pairs of tuples within window.

- Problems
  - Choice of key
  - Choice of window size
- Complexity: At least 3 passes over data
  - Sorting!

Sorted Neighborhood – Key Generation

FNAME	LNAME	ADDRESS	ID	Key
Sal	Stolpho	123 First St.	456780	STOSAL123FRST456
Mauricio	Hernandez	321 Second Ave	123456	HERMAU321SCND123
Frank	Meier	Hauptstr. 11	987654	MEIFRA11HPTSTR987
Sal	Stolfo	123 First Street	456789	STOSAL123FRST456

Sorted Neighborhood – Sort

FNAME	LNAME	ADDRESS	ID	Key
Mauricio	Hernandez	321 Second Ave	123456	HERMAU321SCND123
Frank	Meier	Hauptstr. 11	987654	MEIFRA11HPTSTR987
Sal	Stolpho	123 First St.	456780	STOSAL123FRST456
Sal	Stolfo	123 First Street	456789	STOSAL123FRST456

Sorted Neighborhood – Merge

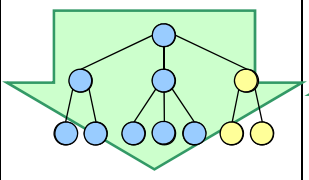
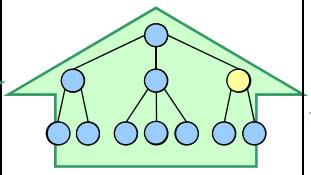
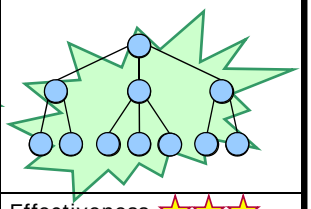
FNAME	LNAME	ADDRESS	ID	Key
Mauricio	Hernandez	321 Second Ave	123456	HERMAU321SCND123
Frank	Meier	Hauptstr. 11	987654	MEIFRA11HPTSTR987
Sal	Stolpho	123 First St.	456780	STOSAL123FRST456
Sal	Stolfo	123 First Street	456789	STOSAL123FRST456

Sorted Neighborhood Method – Variations

- Multi-Pass [Hernandez Stolfo 1998]
  - Several runs with different keys
  - Smaller window
  - Transitive closure over different runs
- Domain-independent [Monge Elkan 1997]
  - 1st pass: Key = tuple
  - 2nd pass: Key = reversed tuple
  - Similarity: Smith-Waterman
- Prime representatives [Monge Elkan 1997]
  - Clusters of duplicates
  - Compare new tuple only with some prime representative of a cluster.
- SNM for Trees [Puhlmann et al. 2006]
  - Idea: Apply SNM bottom up to each hierarchy level.
  - Intuition: Only elements with many duplicate children need to be compared.
  - Increased efficiency

Relationship-aware Algorithms

40

Top-down [Weis, Naumann 04]	Bottom-up [Weis, Naumann 06]	From-the-middle [Weis, Naumann 06b]
		
Effectiveness ★ Efficiency ★★★	Effectiveness ★★ Efficiency ★★	Effectiveness ★★★ Efficiency ★
Further improvements: object-filter; edit-distance-filter; transitivity		

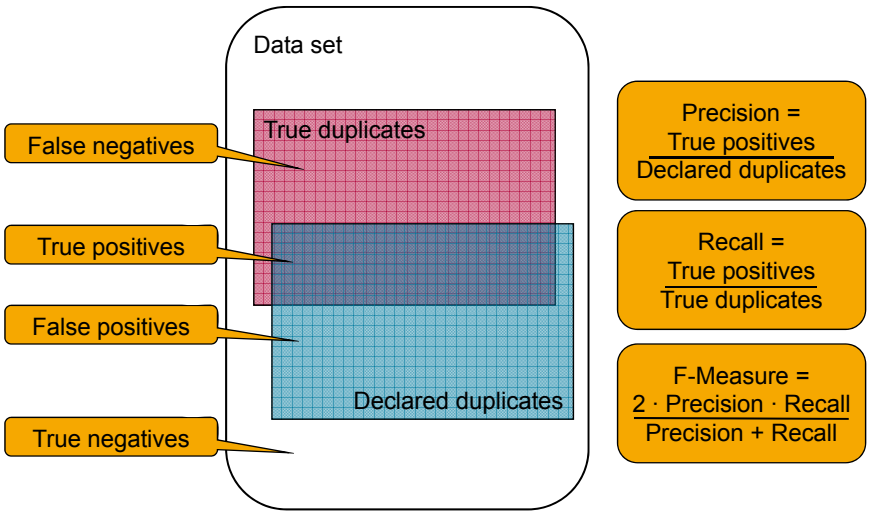
Overview

41

- Information Quality
- Information Integration
- Duplicate Detection
  - Similarity
  - Algorithms
  - Goals
- Data Fusion
- Outlook

Precision & Recall

42



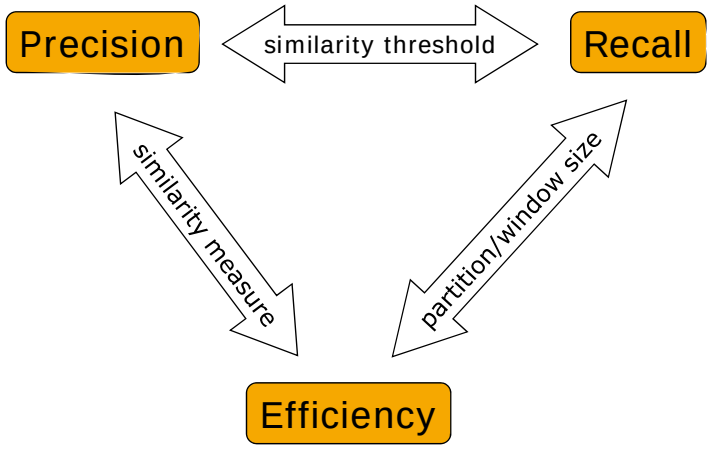
$$\text{Precision} = \frac{\text{True positives}}{\text{Declared duplicates}}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True duplicates}}$$

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Evaluating Duplicate Detection

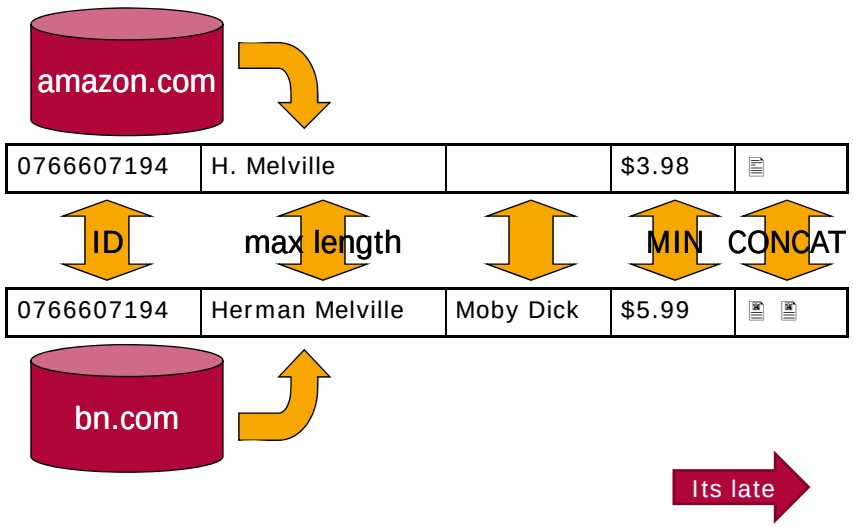
43



Overview

- Information Quality
- Information Integration
- Duplicate Detection
  - Similarity
  - Algorithms
  - Goals
- Data Fusion
- Outlook

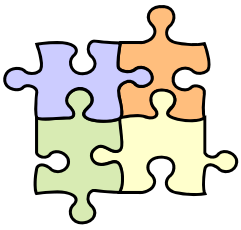
Data Fusion



Relational Data Fusion

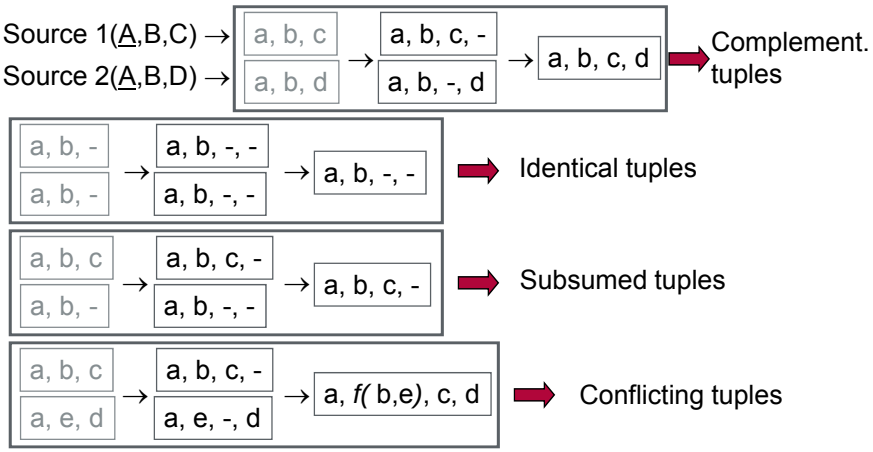
46

- Union ∪
  - Exact duplicate elimination
- Outer union ⟗ (Codd 1979)
  - Union under heterogeneous schemata
- Minimum union ⊕ (Ullmann 1989, Galindo-Legaria 1994)
  - Elimination of subsumed tuples
- (Proper) Data fusion
  - Integrates duplicates
  - Solves conflicts



"Proper" Data Fusion

47



Conflict Resolution Functions

48

Min, Max, Sum, Count, Avg, StdDev	Standard aggregation
Random	Random choice
First, Last	Choose first/last value; depends on order
Longest, Shortest	Choose longest/shortest value
Choose(*source*)	Choose value froma particular source
ChooseDepending(*col, val*)	Choose depending on *val* in other column *col*
Vote	Majority decision
Coalesce	Choose first non-null value
Group, Concat	Group or concatenate all values
MostRecent	Choose most recent (up-to-date) value
MostAbstract, MostSpecific	Use a taxonomy / ontology
….	….

Overview

49

- Information Quality
- Information Integration
- Duplicate Detection
  - Similarity
  - Algorithms
  - Goals
- Data Fusion
- Outlook

Outlook

- Applications beyond CRM
  - Life Sciences and other scientific databases
  - Accounting / billing
  - Products / Catalogs
- Duplicate detection
  - New types of data (XML, multimedia, ontologies, text, …)
    - Similarity measures
  - Algorithms & scalability
- Data fusion
  - Tooling / Automation
  - Efficiency
- Tooling
  - See workshop this afternoon
- CS aspect vs. IS aspect

Summary and thanks

- Data Quality and Data Cleansing
  - In particular in the information integration context
- Duplicate Detection
  - Similarity of records
  - Avoid comparisons to improve efficiency (SNM, etc)
- Data Fusion is the next (and much ignored) step

- Particular thanks to Melanie Weis and Jens Bleiholder!
- Questions and comments anytime: naumann@hpi.uni-potsdam.de

References

- [Ananthakrishna et al. 2002] R. Ananthakrishna, S. Chaudhuri, V. Ganti: Eliminating Fuzzy Duplicates in Data Warehouses, *Proc. of the 28th VLDB Conference*, Hong Kong, China, pages 586-597, 2002.
- [Dey et al. 2002] D. Dey, S. Sarkar, P. De: A Distance-based Approach to Entity Reconciliation in Heterogeneous Databases, *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 14(3): 567-582, 2002.
- [Gravano et al. 2001] L. Gravano, P. Ipeirotis, H. Jagadish, N. Koudas, S. Muthukrishnan, D. Srivastava: Approximate String Joins in a Database (Almost) for Free, *Proc. of the 27th VLDB Conference*, Roma, Italy, pages 491-500, 2001.
- [Hernandez Stolfo 1995] M. Hernandez, S. Stolfo: The Merge/Purge Problem for Large Databases, *Proc. ACM SIGMOD Conference 1995*, San Jose, USA, pages 127-138, 1995.
- [Hernandez Stolfo 1998] M. Hernandez, S. Stolfo: Real-world Data is Dirty: Data Cleansing and the Merge/Purge, *Journal of Data Mining and Knowledge Discovery*, 2(1):9-37, 1998.
- [Jaro 1989] M. Jaro: Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association* 84(406):414-420, 1989.
- [Low et al. 2001] W. Low, M. Lee, T. Ling: A Knowledge-based Approach for Duplicate Elimination in Data Cleaning, *Information Systems* 26(8):585-606, 2001.
- [Monge 2000] A. Monge: Matching Algorithms within a Duplicate Detection System, *IEEE Data Engineering Bulletin*, 23(4):14-20, 2000.
- [Navarro 2001] G. Navarro: A Guided Tour of Approximate String Matching, *ACM Computing Surveys* 31(1):31-88, 2001.
- [Weis Naumann 2005] Melanie Weis, Felix Naumann: DogmatiX Tracks down Duplicates in XML. In Proc. of the ACM Conference on Management of Data (SIGMOD) 2005.
- [GL94] Outerjoins as Disjunctions, Cesar A. Galindo-Legaria, SIGMOD 1994 conference
- [Cod79] E. F. Codd: Extending the Database Relational Model to Capture More Meaning. TODS 4(4): 397-434 (1979)
- [Ull89] Jeffrey D. Ullman: Principles of Database and Knowledge-Base Systems, Volume II. Computer Science Press 1989