600 JAHRE UNIVERSITÄT LEIPZIG

Object Matching for Improving Information Quality

Erhard Rahm

http://dbs.uni-leipzig.de
http://dbs.uni-leipzig.de/wdi-lab

November 25, 2009

WDI-Lab

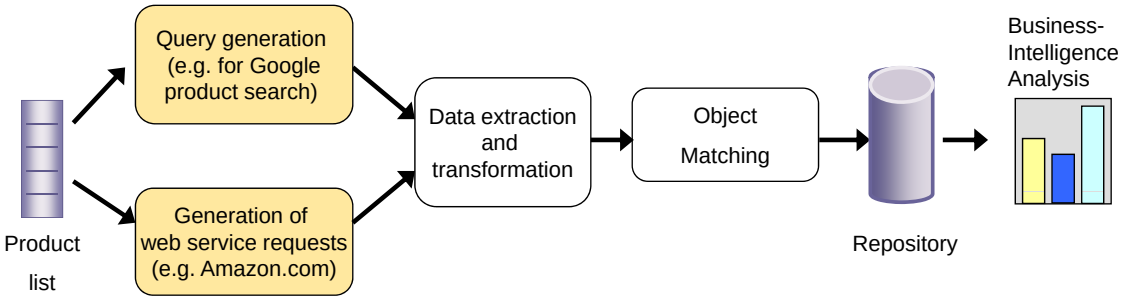
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

ForMaT UNTERNEHMEN REGION Die BMBF-Innovationsinitiative Neue Länder

- Innovation Lab at Univ. of Leipzig on semantic **W**eb **D**ata **I**ntegration
- Funded by BMBF (German ministry for research and education)
  - 2009: initial phase
  - Full funding starts in Jan. 2010 (10 full-time employees + students)
- **Goals**
  - Semi-automatic, high quality data integration of heterogeneous (web) data
  - Faster development of data integration solutions than with traditional integration approaches, e.g. data warehouses
  - Make research approaches ready for the market

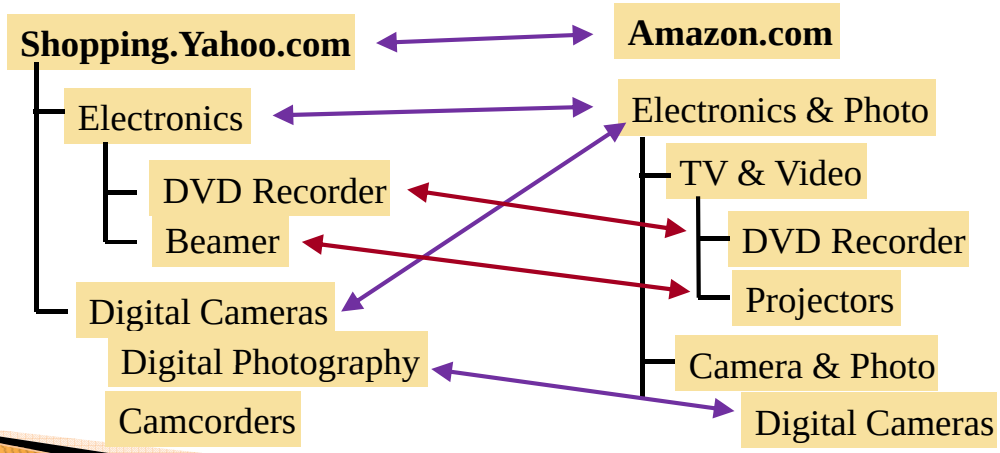
WDI-Lab: Working Groups (1)

- Mashup/Workflow-like data integration
  - Framework to specify and execute workflows for data acquisition from web sources, data transformation, integration and analysis
  - Support for dynamic (runtime) data integration
  - Research prototype: iFuice + extensions



WDI-Lab: Working Groups (2)

- Ontology and Schema matching
  - Semi-automatic generation of mappings between related schemas (e.g., XML business schemas) or ontologies (e.g., product catalogs)
  - Support for large schemas/ontologies
  - Research prototype: COMA++



WDI-Lab: Working Groups (3)

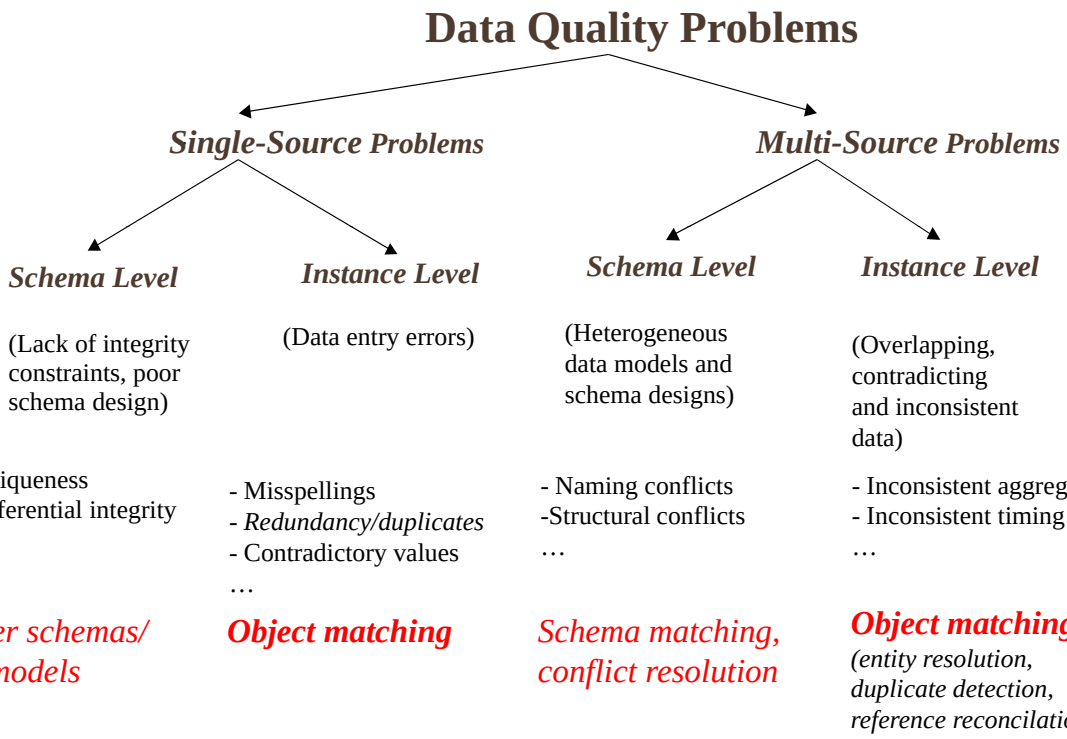
- Object Matching (Entity resolution, Deduplication)
  - Effective strategies for matching related objects (entities, instances) from one or several sources
  - Offline matching (e.g. with data warehouse) and online matching (e.g., within mashup applications)
  - Research prototypes: MOMA, FEVER

Optimizing Result Prefetching in Web Search Engines with Segr	1	1	Optimizing Result Prefetching in Web Search Engines with Segmen
AQuery: Query Language for Ordered Data, Optimization Techr	1	1	AQuery Query Language for Ordered Data Optimization Technique
A Query Language and Optimization Techniques for Unstructur	0.583	0	AQuery Query Language for Ordered Data Optimization Technique
Querying Heterogeneous Information Sources Using Source De	0.545	1	J J Ordille Quering Heterogeneous Information Sources Using Sour
Indexing and Querying XML Data for Regular Path Expressions	0.727	1	Indexing and queryifig XML data for regular path expressions C
RP*: A Family of Order Preserving Scalable Distributed Data Str	1	1	RP A Family of Order Preserving Scalable Distributed Data Structu
NeuroRule: A Connectionist Approach to Data Mining	0.545	1	Rudy setiiono Huan liu A connectionist approach to data mining
NetCube: A Scalable Tool for Fast Data Mining and Compressic	1	1	NetCube A Scalable Tool for Fast Data Mining and Compression
NetCube: A Scalable Tool for Fast Data Mining and Compressic	1	1	NetCube A Scalable Tool for Fast Data Mining and Compression
Change-Centric Management of Versions in an XML Warehous	0.800	1	Changecentric management of versions in an XML warehouse Sep
A Single Pass Computing Engine for Interactive Analysis of VLD	1	1	A Single Pass Computing Engine for Interactive Analysis of VLDBs

Agenda

- Introduction (Object Matching)
- FEVER platform for object matching strategies
  - Architecture
  - Manually specified match strategies (operator trees)
  - Training-based learning of match strategies
  - Evaluation
- Dynamic object matching in mashups
  - OCS (Online Citation Service)
- Instance-based ontology matching
  - Approaches
  - Support in COMA++
- Conclusions

Classification of data quality problems*



* E. Rahm, H. H. Do: *Data Cleaning: Problems and Current Approaches.* IEEE Techn. Bull. Data Eng., Dec. 2000

Object matching problem

- Identify semantically equivalent (matching) objects
  - within one data source or between different sources
  - to integrate (merge) them, compare them, improve data quality, etc.
- Most previous work for structured (relational) data

Source1: Customer

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Source2: Client

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Duplicates in (integrated) web sources

Canon VIXIA HF S10 Camcorder - 1080p - 8.59 MP - 10 x optical zoom
Flash card, 32 GB, 1y warranty, F/1.8-3.0
The VIXIA HF S10 delivers brilliant video and photos through a Canon exclusive 8.59 megapixel CMOS image sensor and the latest version of Canon's advanced image processor, ...
★★★★½ 12 reviews - Add to Shopping List

$975 new
from 52 sellers
Compare prices

Canon (VIXIA) HF S10 VIS Dual Flash Memory Camcorder
Canon HF S10 iVIS Dual Flash Memory CamcorderSPECIAL SALE PRICE: $899 Display both English/Japanese + we supplu all English manuals in English as PDF. ....
Add to Shopping List

$899.00 new
Made in Japan Online

Canon VIXIA HF S10
Dual Flash Memory High Definition Camcorder The Next Step Forward in HD Video Canon has a well-known and highly-regarded reputation for optical excellence, ...
Add to Shopping List

$999.00 new
Performance Audio
2 seller ratings

Canon VIXIA HF S100 Flash Memory Camcorder
***Canon Video HF S100 Instant Rebate Receive $200 with your purchase of a new Canon VIXIA HF S100 Flash Memory Camcorder. (Price above includes $200 ....
Add to Shopping List

$899.95 new
Arlingtoncamera.com
5 seller ratings

Canon Vixia Hf S10 Care & Cleaning
Care & Cleaning Digital Camera/Camcorder Deluxe Cleaning Kit with LCD Screen Guard Canon VIXIA HF S10 Camcorders Care & Cleaning.
Add to Shopping List

$2.99 new
shop.com
★★½☆☆ 38 seller ratings

Duplicates in web sources (2)

A **survey** of approaches to automatic **schema matching** - psu.edu [PDF]
E Rahm, PA **Bernstein** - the VLDB Journal, 2001 - Springer
The VLDB Journal 10: 334–350 (2001) / Digital Object Identifier (DOI) 10.1007/s007780100057 ... A **survey** of approaches to automatic **schema matching** ... Erhard Rahm 1 , Philip A. **Bernstein** 2 ... 1 Universitat Leipzig, ...
Cited by 1818 - Related articles - All 58 versions

[CITATION] A **survey** of **approaches** to automatic **schema matching**
PA **Bernstein**, E Rahm - VLDB Journal, 2001
Cited by 19 - Related articles

[CITATION] Asurveyof approaches to automatic schema **matching**
E **Rahm**, PA **Bernstein** - VLDB Journal, 2001
Cited by 2 - Web Search

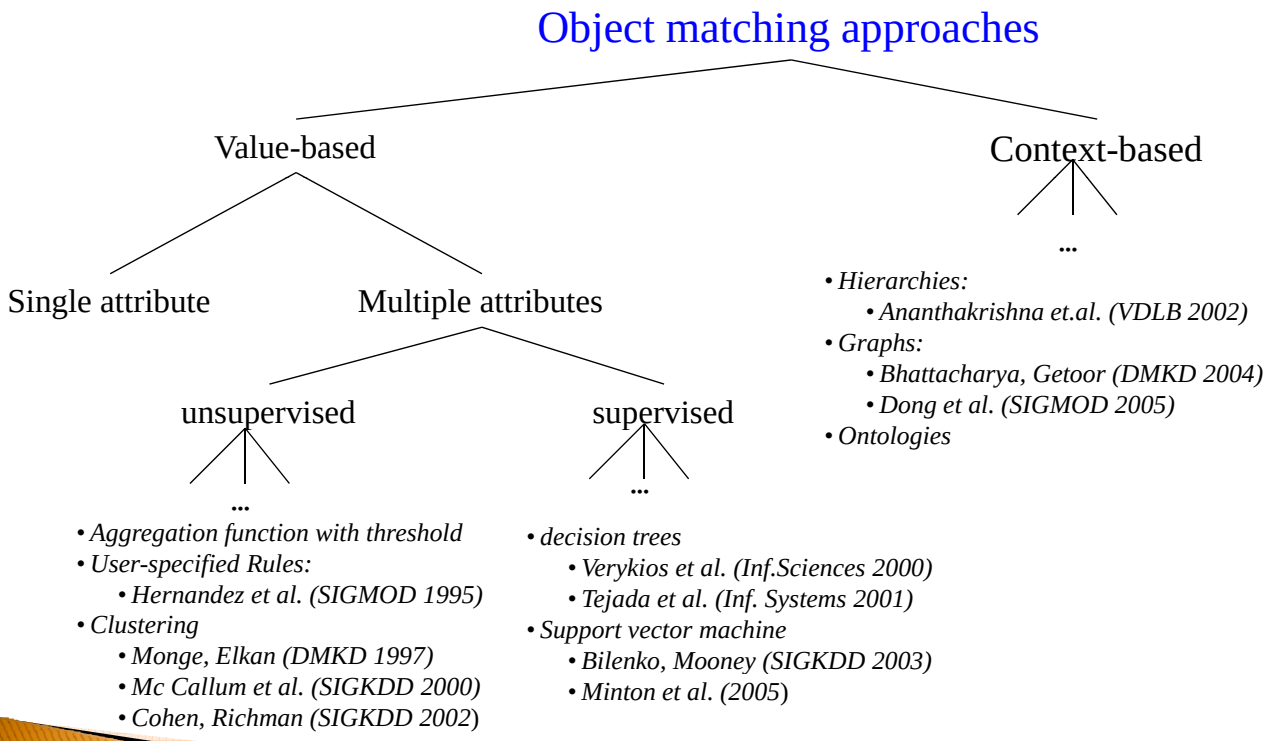
[PDF] On **matching** schemas automatically
E **Rahm**, PA Bernstein - VLDB Journal, 2001 - db15.informatik.uni-leipzig.de
... Erhard **Rahm**, University of Leipzig, Germany Philip A. Bernstein, Microsoft Rese... Redmond, WA, USA Abstract **Schema matching** is a basic problem in many ...
Zitiert durch: 149 - Ähnliche Artikel - HTML-Version - Alle 2 Versionen

[CITATION] A **survey** of approaches to automatic **schema matching**
R Erhard, AB Philip - VLDB Journal, 2001
Cited by 19 - Related articles

Duplicates due to
- Order of authors
- Extraction error (title, author)
- Different titles!
- Typos (author name)

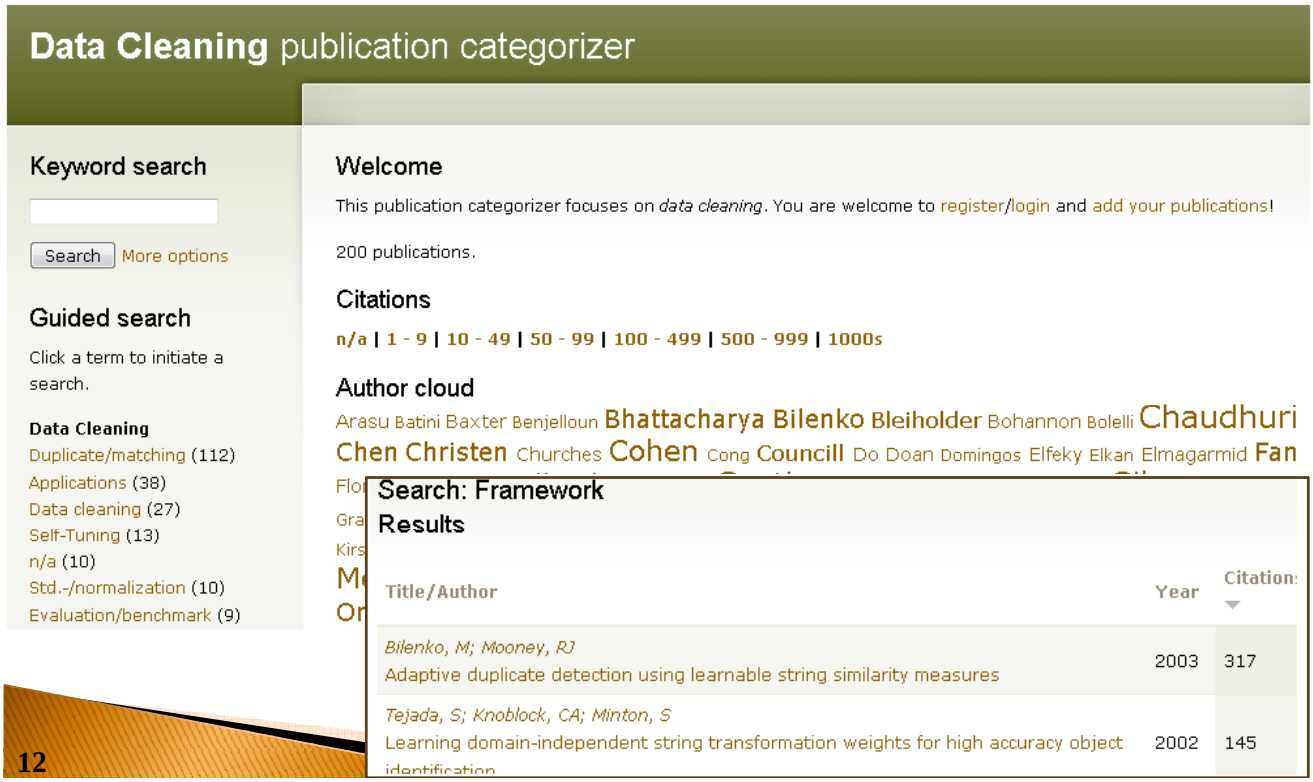
etc.

Object matching approaches



Online Bibliography

http://dc-pubs.dbs.uni-leipzig.de



Object matching frameworks*

- Support combination of several match techniques
- Manual construction of combined strategies
  - BN, MOMA, SERF …
- Learning-based frameworks
  - FEBRL, MARLIN, TAILOR, Active Atlas …
- Problems
  - Evaluation results not conclusive
  - High tuning effort needed
  - Dependency on training data for learning-based approaches

* H. Köpcke, E. Rahm: *Frameworks for Entity Matching: An Overview.* Data and Knowledge Engineering, 2009

Agenda

- Introduction (Object Matching)
- FEVER platform for object matching strategies
  - Architecture
  - Manually specified match strategies (operator trees)
  - Training-based learning of match strategies
  - Evaluation
- Dynamic object matching in mashups
  - OCS (Online Citation Service)
- Instance-based ontology matching
  - Approaches
  - Support in COMA++
- Conclusions

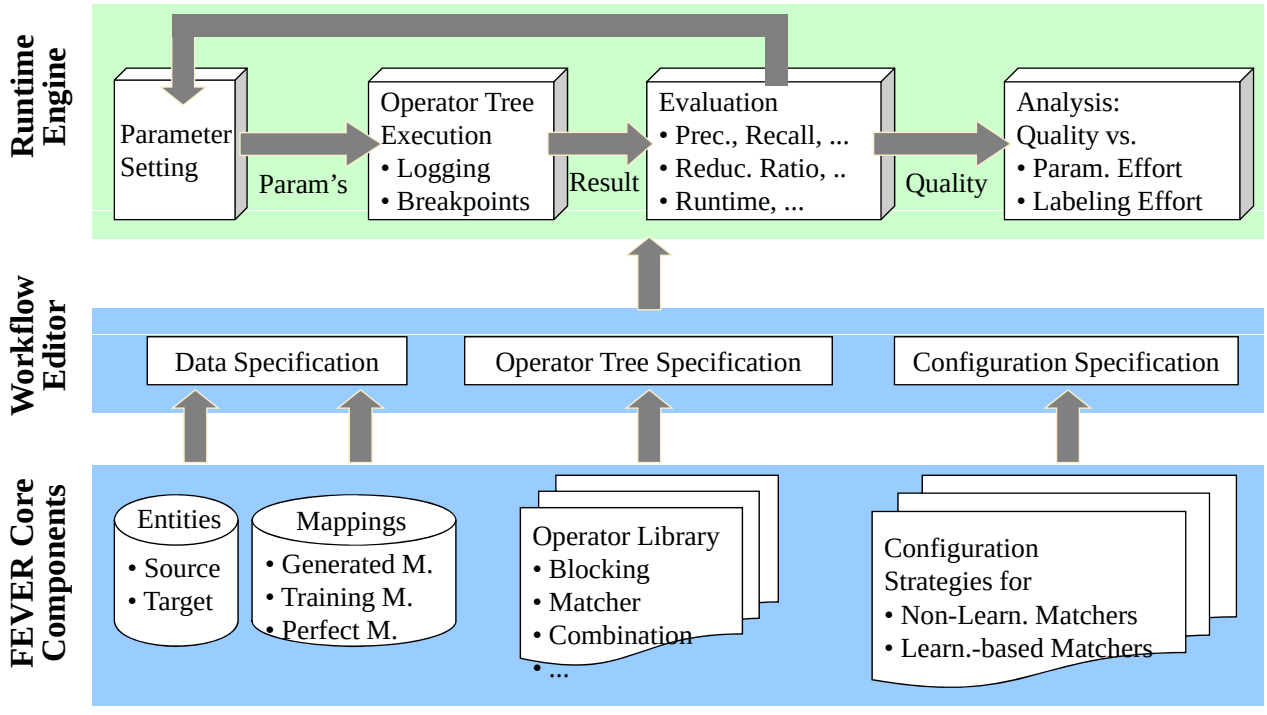
FEVER Framework

Framework for EValuating Entity Resolution

- FEVER = Framework for EValuating Entity Resolution
- Platform for configuration and evaluation of entity resolution (object matching) algorithms and strategies
- Key features:
  - Flexible specification of object matching workflows
  - Semi-automatic parameter configuration (e.g., similarity thresholds)
  - Support for training-based matching to reduce manual tuning effort
  - Comparative evaluations of different match approaches

Köpcke, H.; Thor, A.; Rahm, E.: *Comparative evaluation of entity resolution approaches with FEVER.* Demo, Proc. VLDB, 2009

Architecture of FEVER



Match results

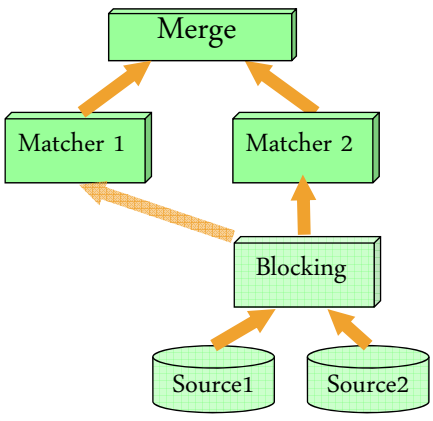
- Match results are represented as instance mappings (correspondences) between 2 sources
  - Mappings can be stored for re-use

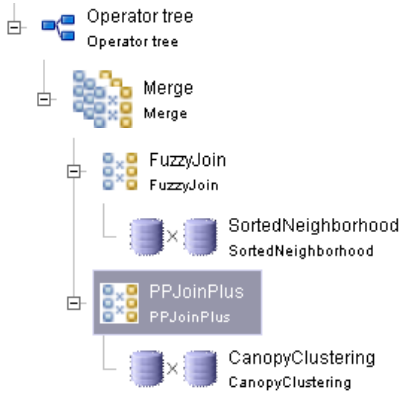
Source1	Source2	Sim
p_1	p'_1	1
p_2	p'_1	0.9
p_3	p'_3	0.8

- Matchers also operate on mappings
  - Cartesian product between input sources
  - Output of previously executed matchers/operators

Operator tree

- Describe workflows implementing a match strategy
  - Leaves: data sources
  - inner nodes: operators (for blocking, matching etc.)
- Execution in post-order traversal sequence
- Match result = Result of root operator



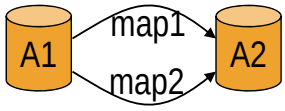


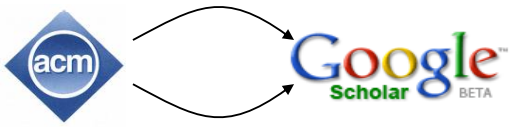
Manual match strategies: operators

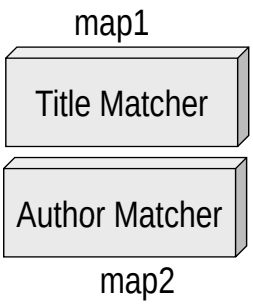
- **Blocking**: Sorted Neighborhood, Canopy Clustering, …
  - ✓ Necessary to reduce search space from Cartesian product to more likely matching object pairs
- Attribute matchers (on preselected pair of attributes):
  - string similarity (TFIDF, Jaccard, Cosine, Trigram, …)
  - PPJoinPlus, EdJoin
  - External implementations, e.g., Fuzzy Lookup (MS SQL Server)
- Context matchers (e.g., Neighborhood matcher)
- Combination of match results: Merge, Compose

Match Strategies: Merge & Compose

1. Merge

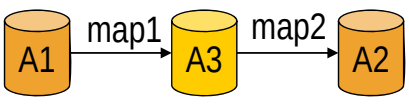


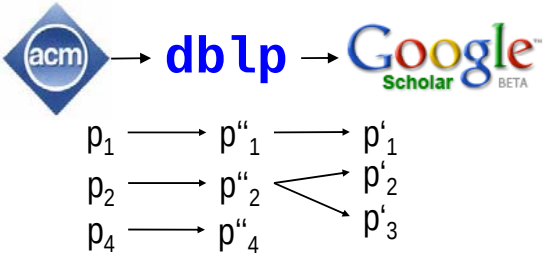




- Overcome short-comings (e.g., precision or recall)

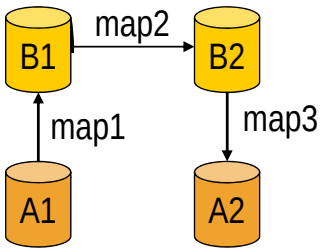
2. Compose

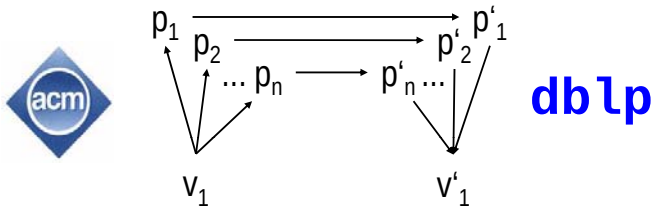




- Efficient re-use of mappings

Match Strategies: Neighborhood

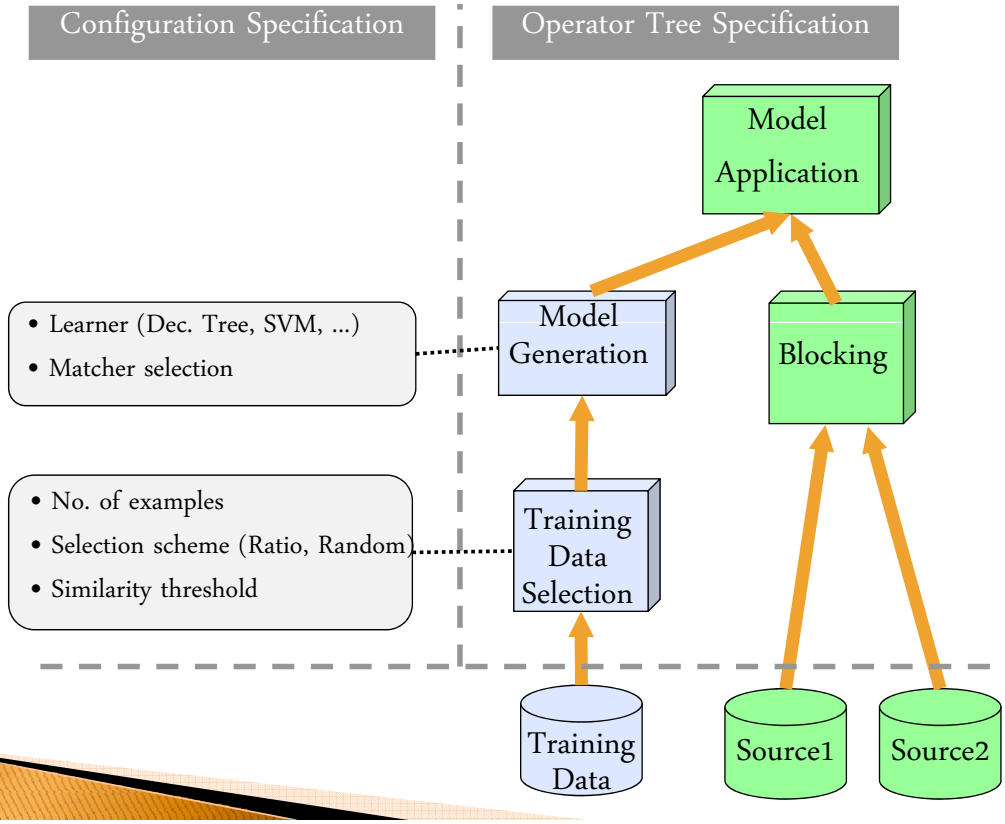




- Combine match mappings with general object relationships
- Bibliographic example: Conference@DBLP - Conference@ACM
  - Attribute matching suffers from highly different values
  - *„Two conferences are the same if they share a significant number of publications.“*
  - Reuse of match result for publications
- Very effective in experiments

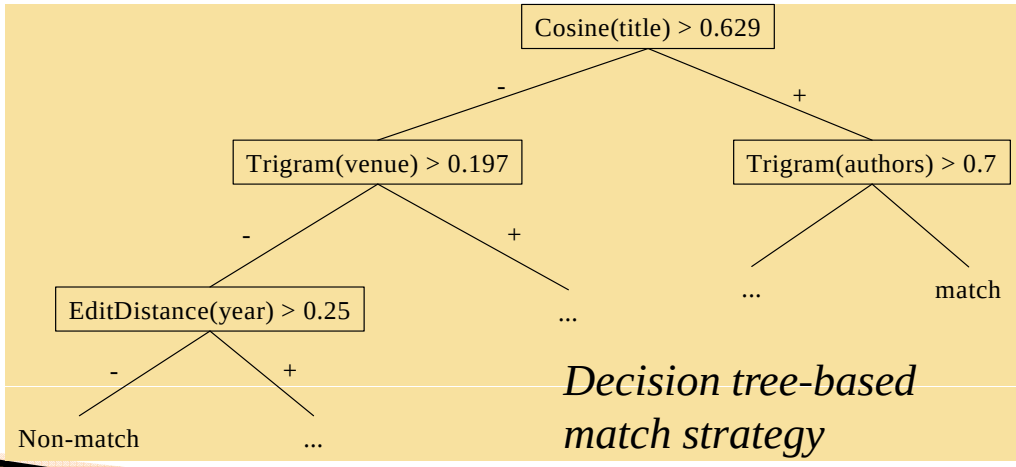
Thor, A.; Rahm, E.: *MOMA - A Mapping-based Object Matching System*. Proc. CIDR, 2007

Learning-based match strategies



Learning-based match strategies (2)

- Use of training data to find effective matcher combination and configuration (supervised learning)
- Learners for model generation in FEVER:
  - Decision Tree, Logistic Regression, SVM
  - Multiple learning approach



Decision tree-based match strategy

Training Selection

- Training data: set of object pairs with manually labelled match/mon-match decisions
  - # training pairs should be low (limit manual effort)
- Training pairs should be non-trivial
  - Similarity above a certain threshold

- Selection approaches in FEVER
  - RANDOM: randomly select *n* object pairs above a similarity threshold *t* for labeling
  - RATIO: reduce *n* randomly selected pairs (above sim. threshold *t*) so that at least a fraction *ratio* (<=0.5) of matching or non-matching pairs are in the training set
    - ✓ ratio 0.4: 40%/60% matches/non-matches (or vice versa)
    - ✓ Balances positive and negative training

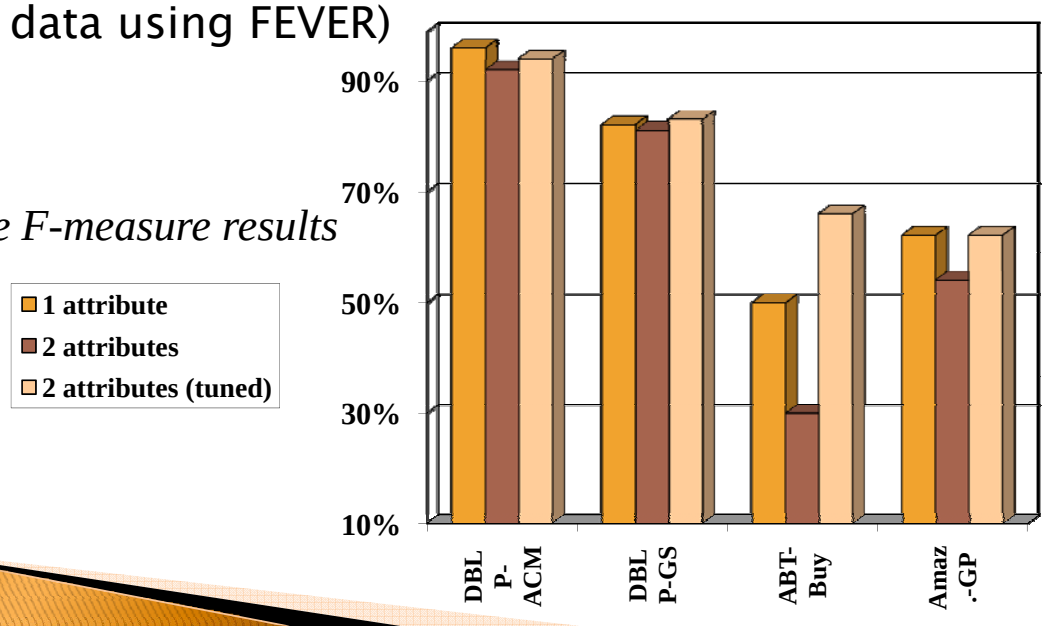
Evaluation

- 7 real data sources:
  - Bibliographic: DBLP, ACM Digital library, GoogleScholar (GS)
  - E-commerce: Abt.com, Buy.com, Amazon.com, Google Product Search (GP),
  - from 1,100 to 64,000 objects per source
- 4 match tasks
  - publications: DBLP-ACM
    DBLP-GS
  - E-Commerce: Abt-Buy
    Amazon - GP
- Perfect mapping:
  - manually determined for bibliographic tasks
  - use of UPCs for E-commerce data

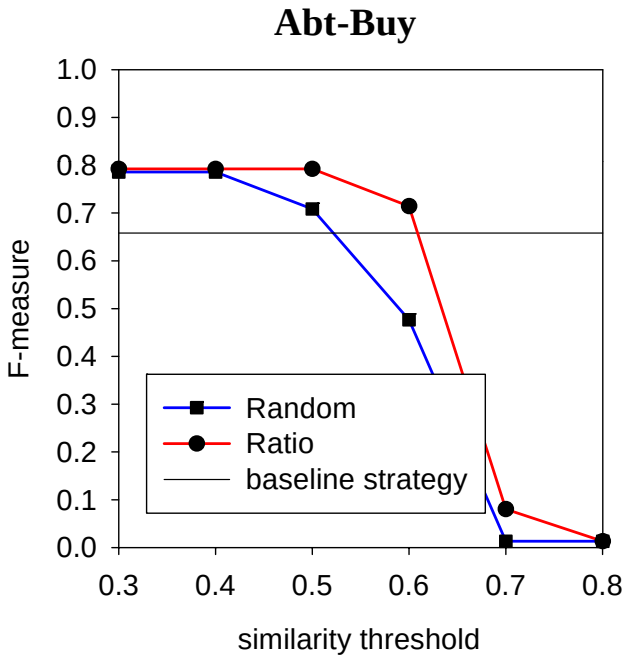
Baseline results

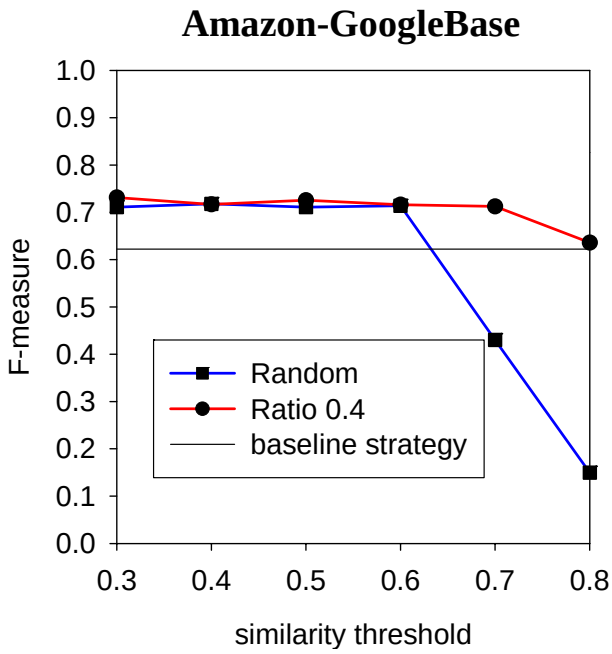
- Use of MS Fuzzy Lookup for comparison
- Similarity on 1-2 attributes
- Default setting for similarity threshold vs. manually tuned settings (varying more than 1000 settings on test data using FEVER)

Baseline F-measure results



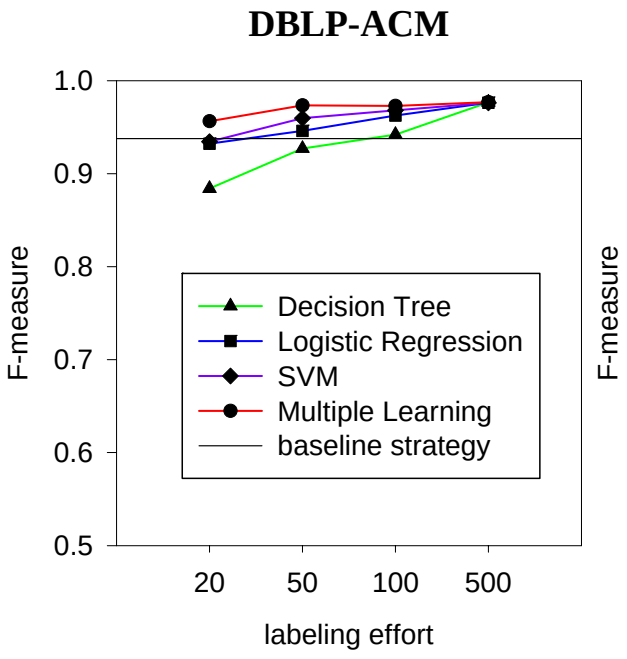
Random vs. Ratio training selection

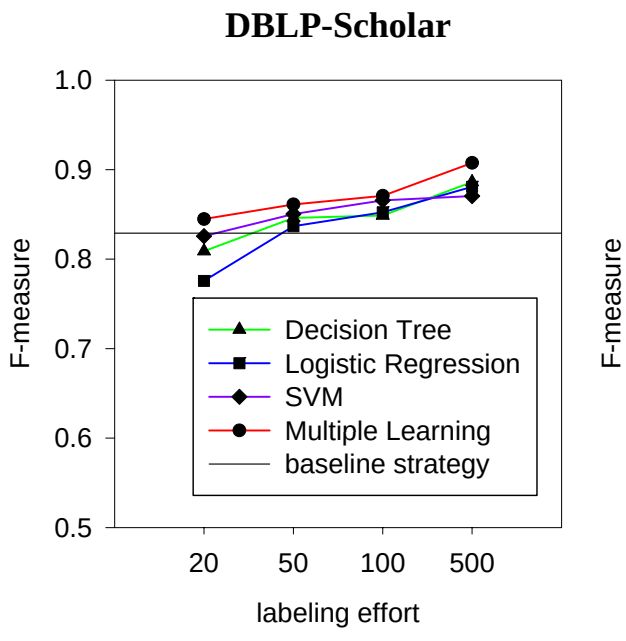




E-commerce tasks, labeling effort 50

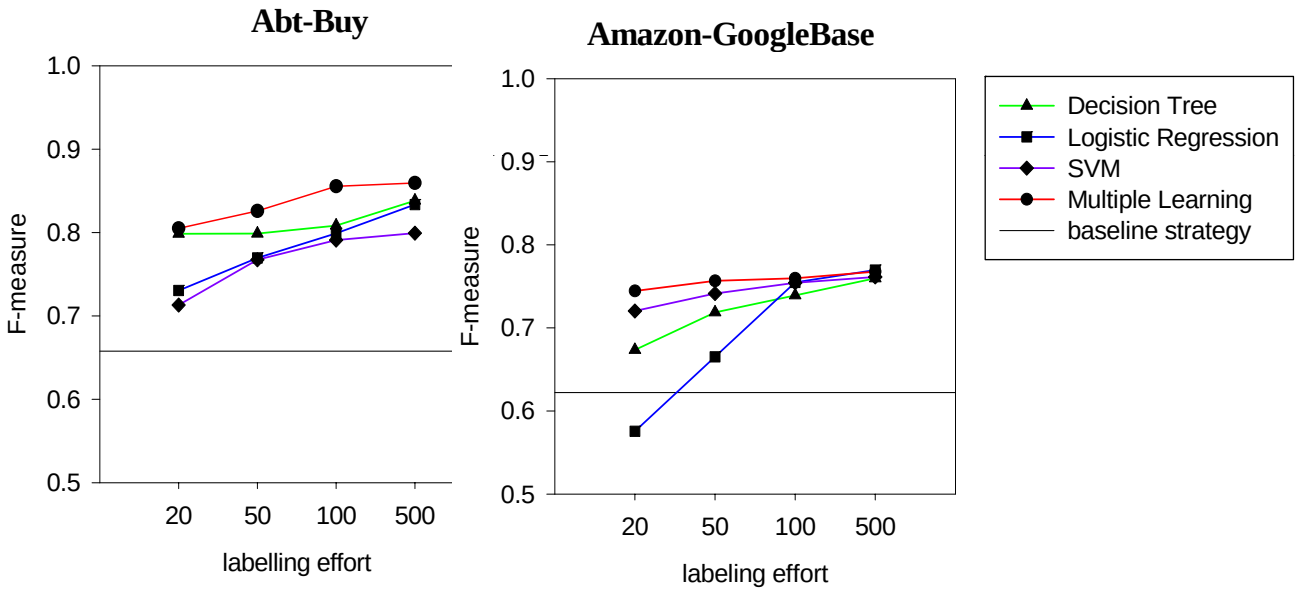
Learner comparison (1)





Bibliographic tasks, Ratio training selection

Learner comparison (2)



E-commerce tasks, Ratio training selection

Evaluation observations

- Match configurations with several matchers are difficult to tune manually with current implementations, e.g. MS Fuzzy Lookup
- Learning-based match strategies can clearly outperform manual match strategies even with small training data, especially for challenging tasks
- *Ratio* is a simple and effective approach for training selection providing a balanced number of matching and non-matching object pairs
- *Multiple learning approach* effectively combines several basic learners

Agenda

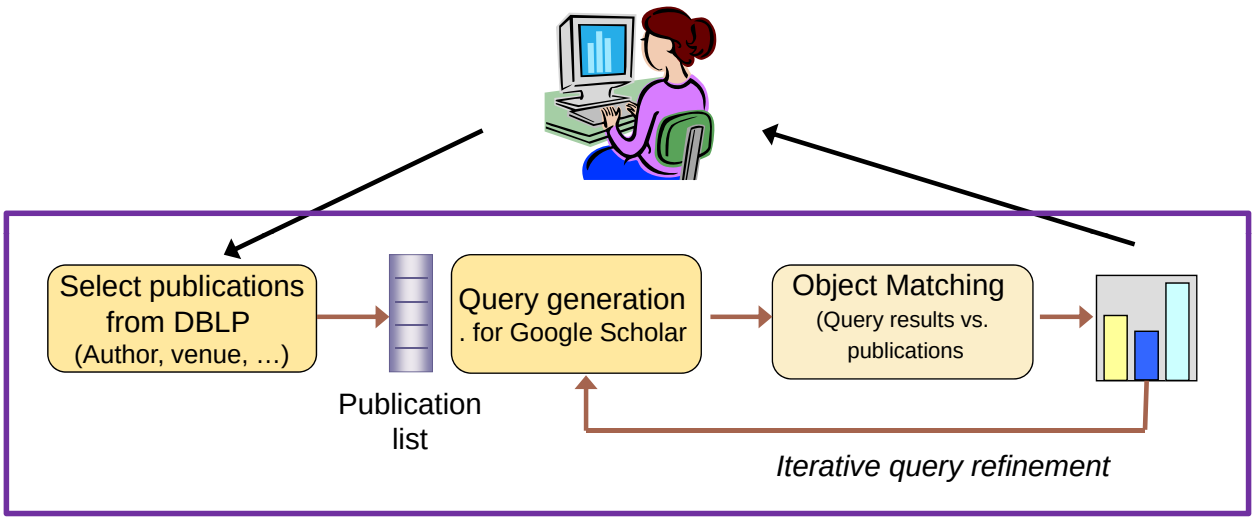
- Introduction (Object Matching)
- FEVER platform for object matching strategies
  - Architecture
  - Manually specified match strategies (operator trees)
  - Training-based learning of match strategies
  - Evaluation
- Dynamic object matching in mashups
  - OCS (Online Citation Service)
- Instance-based ontology matching
  - Approaches
  - Support in COMA++
- Conclusions

OCS Mashup Example

- On-demand citation service (OCS)*
  - What are the most cited papers of conference X or author Y?
  - Frequent changes, i.e., new publications & new citations
- Idea: Combine publication lists, e.g. from DBLP or Pubmed, with citation counts, e.g from Google Scholar, Citeseer or Scopus
  - DBLP, Pubmed: high bibliographic data quality
  - GS: large coverage of citations counts
- Query and match problem: Given a set of DBLP publications → How to effectively find corresponding GS publications?

* **http://labs.dbs.uni-leipzig.de/ocs**

OCS Workflow



- Automatic generation of search queries, e.g. on author, venue, title (pattern)
- Dynamic object matching for search results

Online Citation Service: Result overview

	Title	Authors	Venue	Year	Citation
✥	Knowledge Engineering: Principles and Methods.	Rudi Studer, V. Richard Benjamins, Dieter Fensel	Data Knowl. Eng.	1998	1081
✥	Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information.	Stefan Decker, Michael Erdmann, Dieter Fensel, Rudi Studer	DS-8	1999	540
−	Knowledge Processes and Ontologies. 4 S Staab, R Studer, HP Schnurr, Y Sure: *Knowledge processes and ontologies. Intelligent systems* (2001) 446 S Staab, R Studer, HP Schnurr, Y Sure: *Knowledge processes and ontologies* (2001) 4 R Studer, S Staab, H Schnurr, Y Sure: *Knowledge processes and ontologies* (2001) 3 S Staab, H Schnurr, R Studer, Y Sure: *Y.(2001). Knowledge Processes and Ontologies* 1 S Staab, HP Schnurr, R Studer: *Sure; Y.?Knowledge Processes and Ontologies?* (2001) 1 S Staab, R Studer, HP Schnurr: *Y. Sure, 2001. Knowledge processes and ontologies* 1 S Staab, HP Schnurr, R Studer: *² Sure, Y.(2001). Knowledge processes and ontologies* 1 S STAAB?: *Knowledge processes and ontologies. Intelligent systems, Institute of Electrical and*	Steffen Staab, Rudi Studer, Hans-Peter Schnurr, York Sure	IEEE Intelligent Systems	2001	461

Bibliographic data from DBLP

Corresponding GS publications

Sum of GS citations

OCS example: Top conference papers

OCS result for venue WWW 2007

- Found 247 GS publications for 211 DBLP publications.
- No GS publications found for 19 DBLP publications.
- Overall: 230 DBLP publications having 4561 citations.
- Average: 19,8 citations per publication.
- H-Index: 38
- Match configuration: 80% title similarity, max. 1 year(s) difference, 50% author similarity.

	Title	Authors	Venue	Year	Citation ▼
✥	Yago: a core of semantic knowledge.	Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum	WWW	2007	196
✥	The complex dynamics of collaborative tagging.	Harry Halpin, Valentin Robu, Hana Shepherd	WWW	2007	164
✥	Optimizing web search using social annotations.	Shenghua Bao, Gui-Rong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, Zhong Su	WWW	2007	134
✥	Google news personalization: scalable online collaborative filtering.	Abhinandan Das, Mayur Datar, Ashutosh Garg, ShyamSundar Rajaram	WWW	2007	107
✥	Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography.	Lars Backstrom, Cynthia Dwork, Jon M. Kleinberg	WWW	2007	106
✥	Analysis of topological characteristics of huge online social networking services.	Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, Hawoong Jeong	WWW	2007	104
✥	The two cultures: mashing up web 2.0 and the semantic web.	Anupriya Ankolekar, Markus Krötzsch, Thanh Tran, Denny Vrandecic	WWW	2007	101

OCS Match Strategy

- Interactive approach, i.e., user selects match thresholds

Title	Year	Authors
80%	+/- two years	50%
85%	+/- one year	60%
90%	equal year	70%
95%		80%
100%		90%
		100%

relaxed

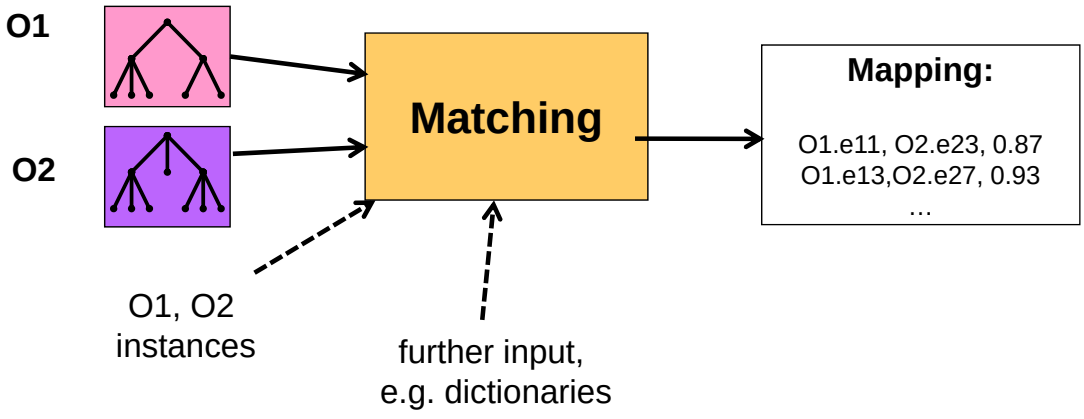
restrictive

- Aggregated result is adjusted automatically based on match definition

Agenda

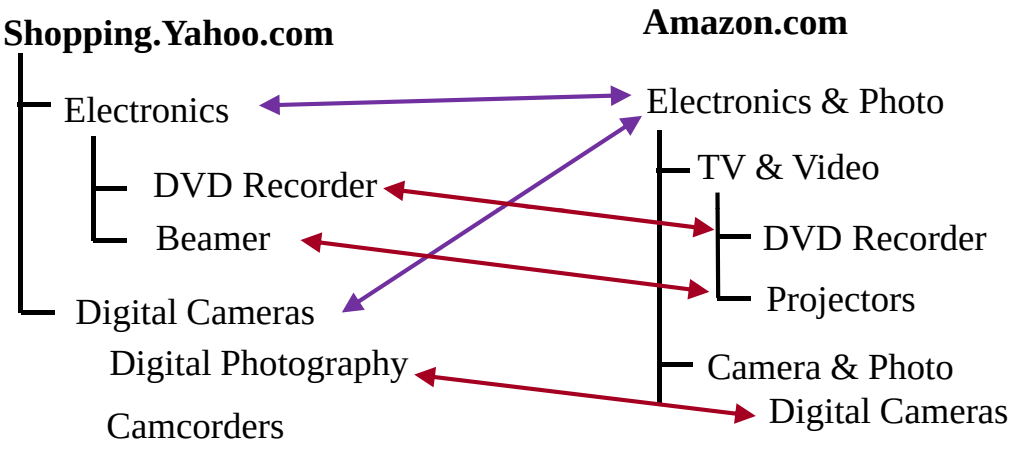
- Introduction (Object Matching)
- FEVER platform for object matching strategies
  - Architecture
  - Manually specified match strategies (operator trees)
  - Training-based learning of match strategies
  - Evaluation
- Dynamic object matching in mashups
  - OCS (Online Citation Service)
- Instance-based ontology matching
  - Approaches
  - Support in COMA++
- Conclusions

Ontology Matching / Alignment



- Process of identifying semantic correspondences between 2 ontologies
  - Result: ontology mapping
  - Mostly equivalence mappings: correspondences specify equivalent ontology concepts
- Variation of schema matching problem

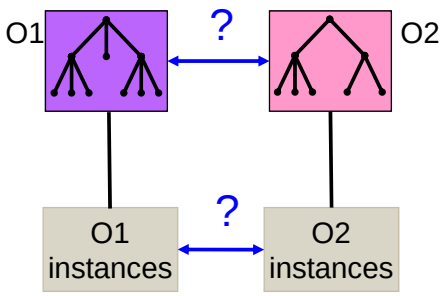
Matching of Product Catalogs



- Ontology mappings useful for
  - Improving query results, e.g. to find specific products
  - Automatic categorization of products in different catalogs
  - Merging catalogs

Instance-based matching

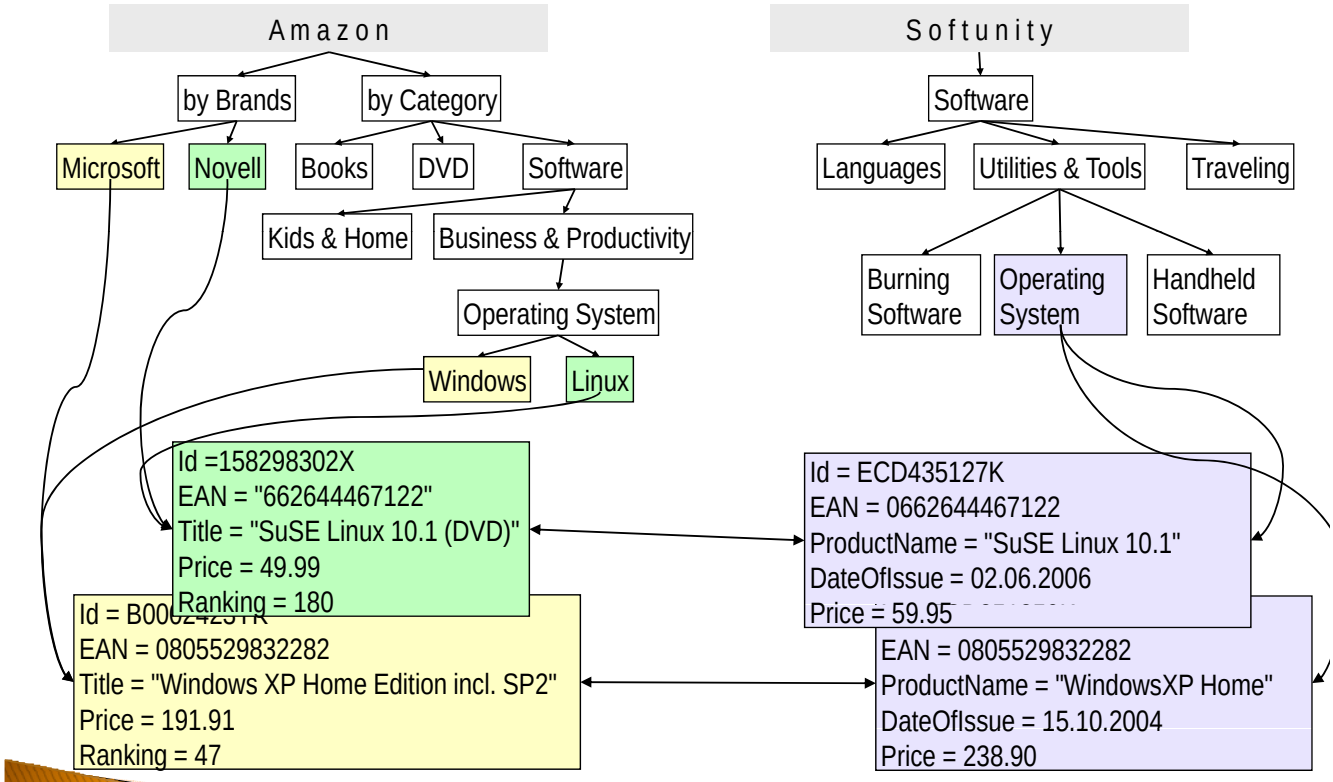
- semantics of a concept/category may be better expressed by the instances associated to category than by metadata (e.g. concept name, description)
- Categories with most similar instances should match
  - Requires shared or similar instances for most/all concepts



Two cases

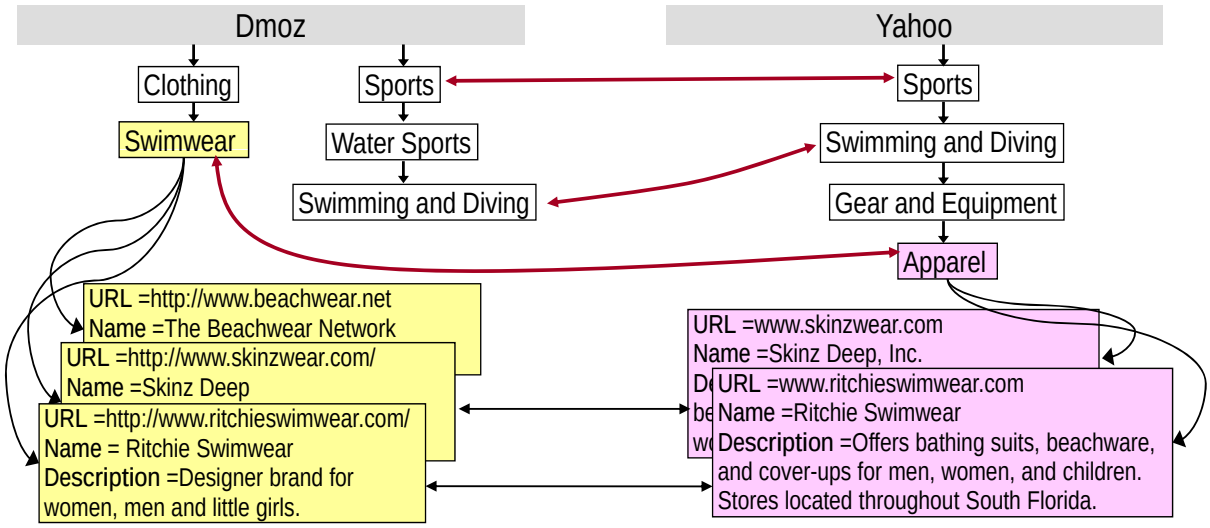
- ontologies share instances
- ontologies do not share but have similar instances

Use case 1: Product Catalogs



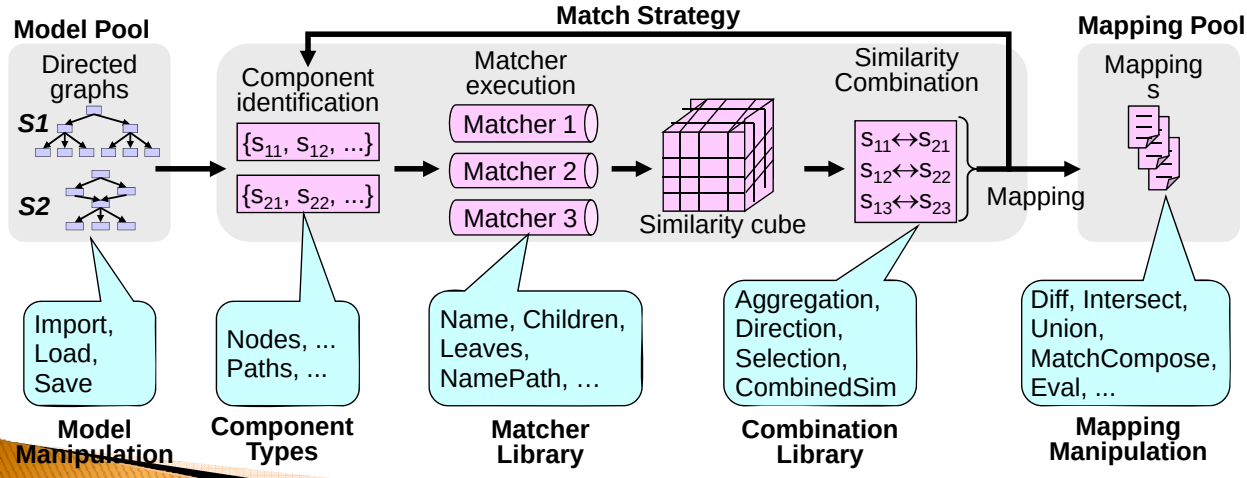
Thor, A., Kirsten, T., Rahm, E.: *Instance-based matching of hierarchical ontologies.* Proc. 12th BTW Conf., 2007

Use case 2: Web Directory Matching *



* Massmann, S., Rahm, E.: *Evaluating Instance-based Matching of Web Directories.* Proc. WebDB 2008

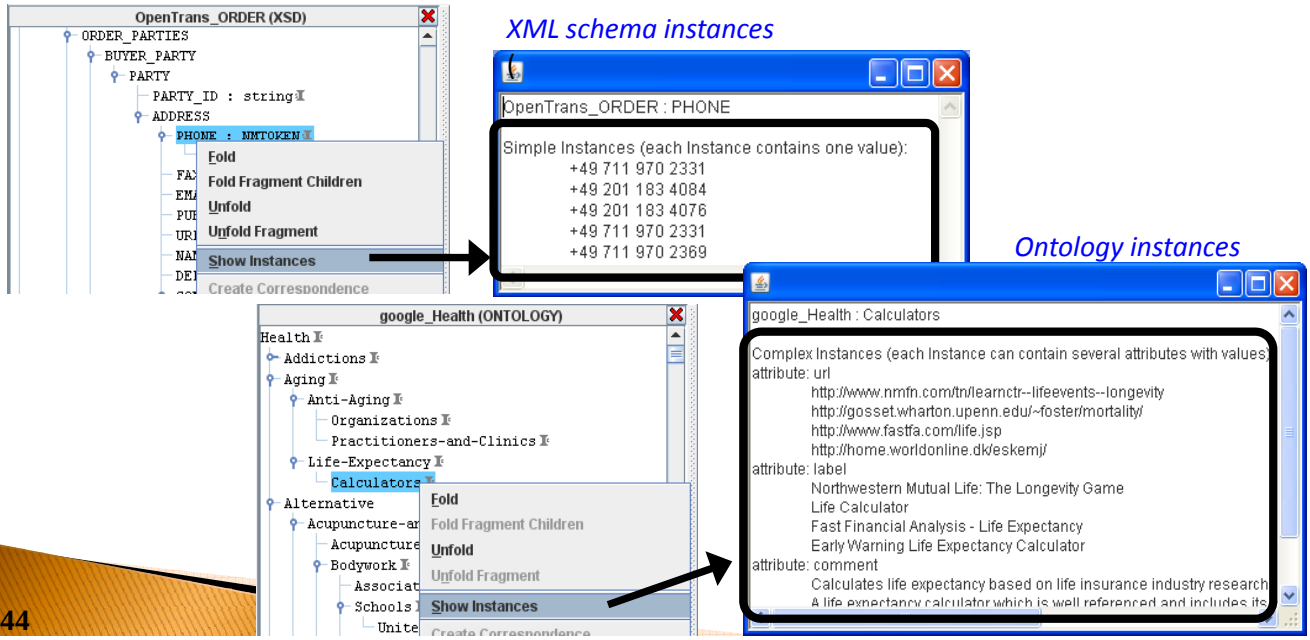
- Extends previous COMA prototype (VLDB2002)
- Matching of XML & rel. Schemas and OWL ontologies
- Several match strategies: Parallel (composite) and sequential matching; Instance-based matching; Fragment-based matching for large schemas; Reuse of previous match results



*Schema and Ontology Matching with COMA++. Proc. SIGMOD 2005

Instance-based Matching in COMA++

- Instance matchers introduced in 2006
  - Constraint-based matching
  - Content-based matching: 2 variations
- Coma++ maintains *instance value set* per element

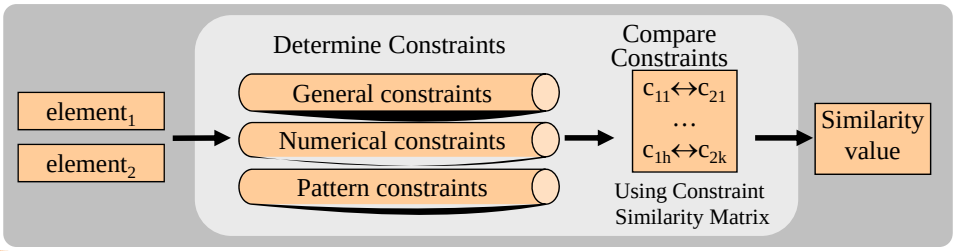


Constraint-based Matching

- Instance constraints are assigned to schema elements
  - **General constraints**: always applicable
    Example: average length and used characters (letters, numeral, special char.)
  - **Numerical constraints**: for numerical instance values
    Example: positive or negative, integer or float
  - **Pattern constraints**:
    Example: Email and URL

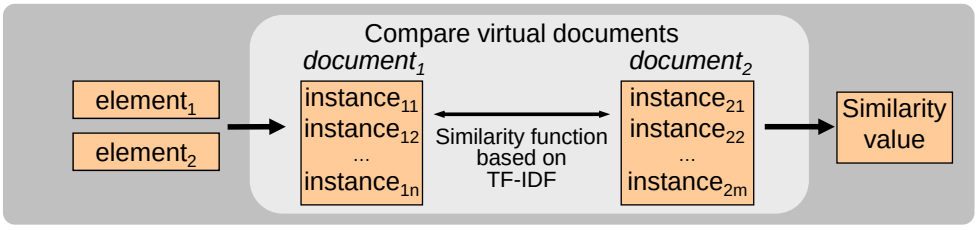
"My@email.com" vs. "Your@email.org"

- Use of constraint similarity matrix to determine element similarity (like data type matching)
- Simple and efficient approach
  - Effectiveness depends on availability of constrained value ranges / pattern
- Approach does not require shared instances



Content-based Matching

- 2 variations
  - *Value Matching*: pairwise similarity comparison of instance values
  - *Document (value set) matching*: combine all instances into a virtual document and compare documents
  - Both approaches do not require shared instances
- Document matching
  - 1 instance document per category or selected string category attribute (e.g. description)
  - Document comparison based on TF-IDF to focus on most significant terms



Agenda

- Introduction (Object Matching)
- FEVER platform for object matching strategies
  - Architecture
  - Manually specified match strategies (operator trees)
  - Training-based learning of match strategies
  - Evaluation
- Dynamic object matching in mashups
  - OCS (Online Citation Service)
- Instance-based ontology matching
  - Approaches
  - Support in COMA++
- Conclusions

Conclusions

- Object matching is a critical step for data quality and data integration
  - Offline and online data integration
- Effective match strategies combining several matchers are hard to find and tune
  - Very large number of possible combinations and configurations
  - High quality vs. efficiency tradeoff
  - Utilization of domain knowledge
- Learning-based approaches support semi-automatic generation of suitable match strategies
  - Requires suitable training selection (e.g. Ratio approach)
  - Multiple Learning approach is robust and effective (but slow)

Conclusions (2)

- Instance-based matching of ontologies facilitated by object matching
  - Instances can reflect well semantics of categories
  - Same/similar instances required in both ontologies
- Instance-based matching in COMA++
  - 3 basic instance matchers (constraint-based, content-based) not requiring shared instances
  - Flexible combination with many metadata-based approaches
- Correct ontology mappings NOT limited to 1:1 correspondences

Some Areas for Further Work

- Support for high efficiency and high effectiveness
  - Performance techniques, e.g. parallel object matching
- Evaluation and validation for larger datasets
- Self-Tuning of context matchers
- Scalable instance-based ontology match approaches

References

- Köpcke, H.; Thor, A.; Rahm, E.: *Comparative evaluation of entity resolution approaches with FEVER.* Proc. 35th Intl. Conference on Very Large Databases (VLDB), Demo, 2009
- Köpcke, H., Rahm, E.: *Frameworks for Entity Matching An Overview.* Data and Knowledge Engineering, 2009
- Köpcke, H.; Rahm, E.: *Training Selection for Tuning Entity Matching.* Proc. 6th Int. Workshop on Quality in Databases and Management of Uncertain Data (QDB/MUD), 2008
- Massmann, S.; Rahm, E.: *Evaluating Instance-based matching of web directories.* Proc. 11th Int. Workshop on the Web and Databases (WebDB), 2008
- Thor, A., Kirsten, T., Rahm, E.: *Instance-based matching of hierarchical ontologies.* Proc. 12th German Database Conf. (BTW), 2007
- Thor, A.; Rahm, E.: *MOMA - A Mapping-based Object Matching System*. Proc. of the 3rd Biennial Conf. on Innovative Data Systems Research (CIDR), 2007