Data Warehouses and Web Data Integration

Andreas Thor

July 2nd, 2008

Two worlds: Data warehouse & web data

- Data warehouse
  - integrated, centralized data
  - closed world (company-internal information)
  - high data quality
  - stable
- Web data
  - non integrated
  - open world (“data about everything”)
  - diverse data quality
  - volatile
- Web data can enhance DWH data
  - additional dimensions for fact table(s)
  - additional characteristics for existent dimension(s)

→ Increase coverage by preserving high data quality

Example scenarios

- E-Commerce: product sales
  - DWH: products (article information, supplier, purchases, ...)
  - Web: reviews, competitor's prices, ...
- E-Commerce: customer relationship management
  - DWH: customers (address, channels, …)
  - Web: socio-economic data, territory risk analysis, …
- Bibliographic domain
  - DWH: bibliographic information (title, venue, authors, ...)
  - Web: citation counts, institutions, keywords, ...
- Geographic domain
  - DWH: geographic places (name, inhabitants, region, ...)
  - Web: hotel ratings, point of interest, …
- Bioinformatics: GeWare (?)
- …

Agenda

- Motivation
- Web entity search architecture
- OCS 2.0: Lessons learned
- Search strategy generation
  - by query relaxation
  - by page analysis
- Conclusion

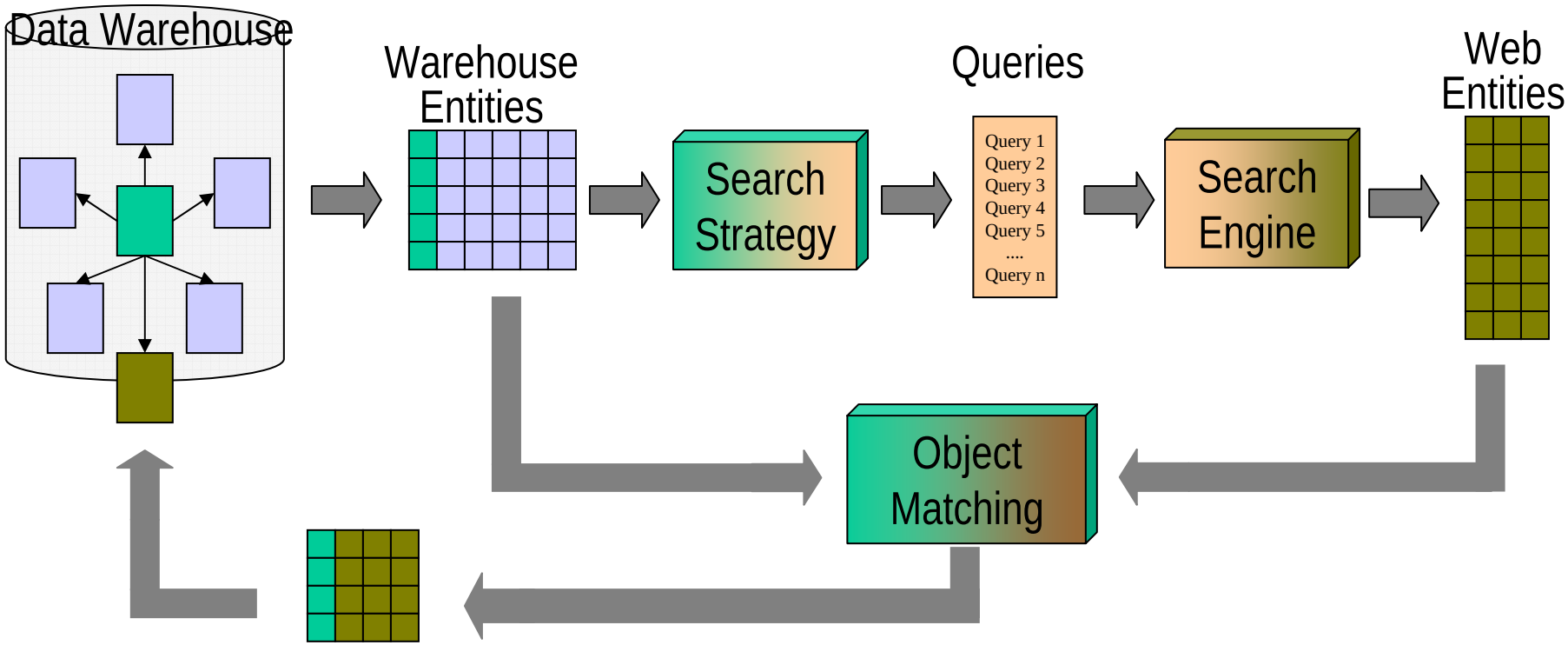
Finding web entities

- Problem: Find all relevant (web) entities for a given subset of the data warehouse

- „Google approach“
  - crawl the WWW, extract entities, import into warehouse
  - too expensive, even for focused crawling
- Search-based approach (“Use Google approach”)
  - use existing (and powerful) search engine technology
  - efficient access to millions of web sources

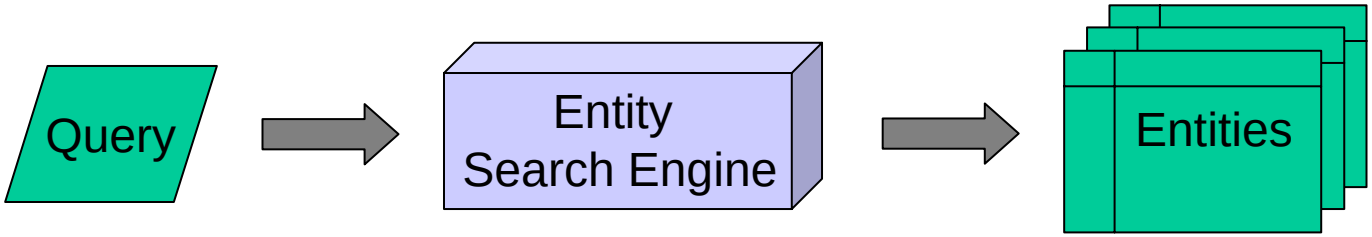
Architecture / workflow

- Find all relevant (web) entities for a given DWH subset
  - set of fact table instances + associated dimensions
  - set of dimension instances

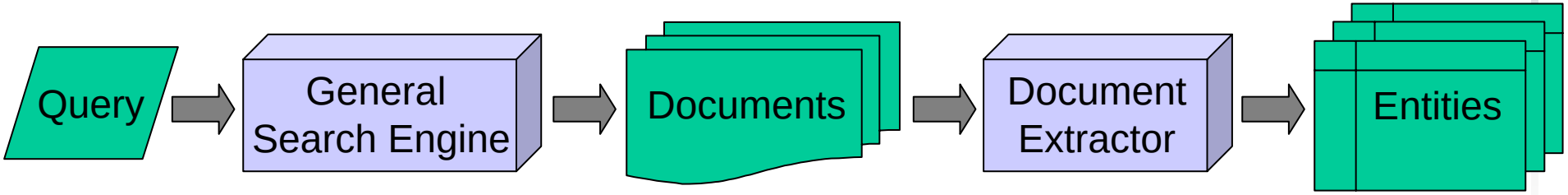


Entity search

- Entity search engines: Google Scholar, Google Base, ...



- General search engines: Google, Yahoo, ...
  - document extractor is site-specific (e.g., created with Dapper)
  - restrict search to website, e.g., "<query> site:portal.acm.org" searches for <query> within ACM portal
  - „interesting" web sites, e.g., ACM portal, Wikipedia, ...



Search strategies

- Input: n input entities
- Output: m queries (for a given search engine)
- Intention: query execution yields corresponding web entities

- Effectiveness
  - find all relevant entities (recall = 100%)
  - find only relevant entities (precision = 100%) → object matching
- Efficiency
  - minimize the user-/administrator effort for query generation
  - minimize the number of queries (run-time, accessibility, ...)

- Problem: Automatic or manual strategy definition / query generation?

Search strategies: Lessons learned

- Prototype: Online Citation Service (OCS) 2.0
  - user selects set of DBLP publications
  - OCS searches for corresponding publications in other sources (and summarizes the citation counts per source)
- Sources
  - Google Scholar, MS Libra → entity search engine
  - ACM, Citeseer → general search engine
- Application of multiple search strategies
  - OCS allows a flexible integration of new strategies

Online Citation Service 2.0

	All	DBLP	GS	ACM	CS	Libra
# Publications	81	81	163	82	28	86
∑ Citations			5861	1087	932	1850
∅ Citations			72.4	13.4	11.5	22.8
H-Index			35	20	16	24

Title 50 | Auth 50 | Year -2

GS Citations (DESC)

Google Scholar (GS)

Author | Author+CoAuthors | Title Pattern | Keyword

Unmark All | Mark All | Mark Selected

A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces (VLDB 1998) *Roger Weber, Hans-Jörg Schek, Stephen Blott*	687	144	147	223
A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces				
A quantitative analysis and performance study for similarity-search methods in high-dimensional	678			
-¬A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional	3			
S. Blott. A quantitative analysis and performance study for similarity-search methods in high-	2			
Schek, Stephen Blott. A Quantitative Analysis and Performance Study for Similarity-Search Methods in	2			
August 24-27). A quantitative analysis and performance study for similarity-search methods in	1			
quantitative analysis and performance study for similarity-search methods in high-dimensional spaces	1			
A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces		144		
A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces			147	
A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces				222
Stephen Blott: A Quantitative Analysis and Performance Study for Similarity - Search Methods in High - Dimensional Spaces				1
Algorithms for Mining Distance-Based Outliers in Large Datasets (VLDB 1998) *Edwin M. Knorr, Raymond T. Ng*	480	62	60	90
MindReader: Querying Databases Through Multiple Examples (VLDB 1998) *Yoshiharu Ishikawa, Ravishankar Subramanya, Christos Faloutsos*	386	57	61	125

A **survey of approaches to automatic schema** matching - Alle 35 Versionen »
E **Rahm**, PA Bernstein - The VLDB Journal The International Journal on Very Large ..., 2001 - Springer
Page 1. The VLDB Journal 10: 334–350 (2001) / Digital Object Identifier (DOI)
10.1007/s007780100057 A **survey of approaches to automatic schema** matching ...
Zitiert durch: 1222 - Ähnliche Artikel - Websuche

[ZITATION] A **survey of approaches to automatic schema** matching
PA Bernstein, E **Rahm** - VLDB Journal, 2001
Zitiert durch: 17 - Ähnliche Artikel - Websuche

[ZITATION] A **survey of approaches to automatic schema** mapping
E **Rahm**, PA Bernstein - The VLDB Journal, 2001
Zitiert durch: 7 - Ähnliche Artikel - Websuche

[ZITATION] A **survey of approaches to semantic schema** matching
E **Rahm**, PA Bernstein - The VLDB Journal 10: 334, 2001
Zitiert durch: 6 - Ähnliche Artikel - Websuche

[ZITATION] A **survey of approaches to automatic schema** mapping" the VLDB Journal
E **Rahm**, PA Bernstein - Vol
Zitiert durch: 3 - Ähnliche Artikel - Websuche

[ZITATION] A.(2001). A **survey of approaches to automatic schema** matching
E **Rahm**, PA Bernstein - The International Journal on Very Large Data Bases (VLDB), 2001
Zitiert durch: 2 - Ähnliche Artikel - Websuche

[ZITATION] A **survey of approaches to automatic schema** matching. 2001
E **Rahm**, P Bernstein - VLDB Journal
Zitiert durch: 1 - Ähnliche Artikel - Websuche

+ 2 additional

Entity search engine: Example (2)

- Publication: *Erhard Rahm, Philip A. Bernstein: A survey of approaches to automatic schema matching. The VLDB Journal 10(4): 334-350 (2001)*

Search strategy	Query	#Results
title (simple)	a survey of approaches to automatic schema matching	18.2000
title (simple phrase)	„a survey of approaches to automatic schema matching”	1.040
title	intitle:a survey of approaches to automatic schema matching	3.530
title (phrase)	intitle:”a survey of approaches to automatic schema matching”	7
title pattern	intitle:"survey * approaches * * schema“	13
author (1)	author:rahm	2.230
author (2)	author:bernstein author:rahm	44
author (1) + title pattern	author:rahm intitle:"survey * approaches * * schema“	9
author (1) + keyword	author:rahm intitle:survey intitle:schema	9

Lessons learned: Entity search engine

- Search strategies can be generated by answering …
  - What attributes should be queried for?
  - What attribute value transformation should be applied?
- Domain 1: Computer science
  - small number of authors
  - title contains common words + acronyms
  - heterogeneous venue names

  → good strategies: „author + keyword“, „title pattern“
- Domain 2: Chemistry
  - high number of authors
  - title contains string representations of chemical formulas
  - homogenous venue names

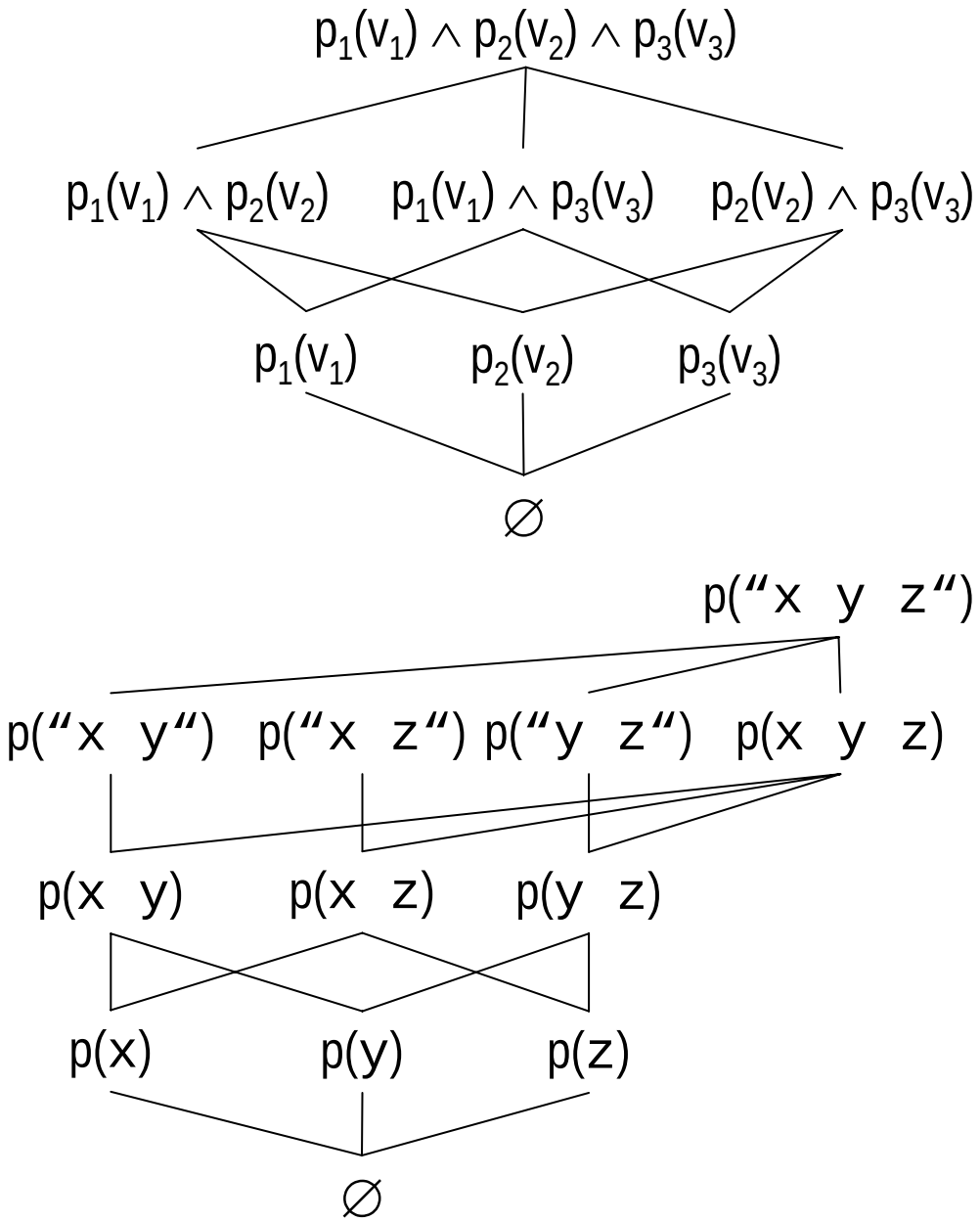
  → good strategies: „venue (+year)“, „title keywords“, „author group“

Entity search: Model

- Query = $p_1(v_1) \wedge p_2(v_2) \wedge \ldots \wedge p_n(v_n)$
- p_i = predicates
  - realized as search form fields and/or „[predicate name]:“
  - p_0 = free text
- v_i = search values for predicates
  - derived from attribute values
  - may be transformed (keyword selection, pattern, reduction, …)

Query containment

- $A \subseteq B$: all resulting entities of query A appear in B‘s search result
- Containment based on
  - predicates
  - attribute values



Search strategy generation by query relaxation

- Example-based query learning
  - input: n DWH entities + m corresponding web entities
  - output: algorithm for query generation
  - intention: algorithm also works for other DWH entities
- Start with fully specified query, e.g., all predicates with exact attribute value
- Relax query by
  - eliminating predicates
  - generalizing / transforming attribute values
- Determine precision and recall
  - for example pages (training data)
  - for test data
- Find optimum (w.r.t. to minimal number of query size)
  - overfitting

General search engine: Example

URL: http://portal.acm.org/citation.cfm?id=767149.767154
Title: A survey of approaches to automatic schema matching

A survey of approaches to automatic schema matching

Full text Pdf (196 KB)

Source The VLDB Journal — The International Journal on Very Large Data Bases archive
Volume 10 , Issue 4 (December 2001) table of contents
Pages: 334 - 350
Year of Publication: 2001
ISSN:1066-8888

Authors Erhard Rahm Universität Leipzig, Institut für Informatik, 04109 Leipzig, Germany; (e-mail: rahm@informatik.uni-leipzig.de)
Philip A. Bernstein Microsoft Research, Redmond, WA 98052-6399, USA; (e-mail: philbe@microsoft.com)

Publisher Springer-Verlag New York, Inc. Secaucus, NJ, USA

Bibliometrics Downloads (6 Weeks): 14, Downloads (12 Months): 181, Citation Count: 172

Query	#Results
"a survey of approaches to automatic schema matching"	13.200
intitle:"a survey of approaches to automatic schema matching"	63
intitle:"a survey of approaches to automatic schema …" inurl:citation.cfm site:portal.acm.org	1
"Erhard Rahm" inurl:citation.cfm site:portal.acm.org	2.530
inanchor:"Erhard Rahm" inurl:citation.cfm site:portal.acm.org	40
intitle:schema inanchor:"Erhard Rahm" inurl:citation.cfm site:portal.acm.org	4
intitle:survey inanchor:"Erhard Rahm" inurl:citation.cfm site:portal.acm.org	1

Lessons learned: General search engine

- Search strategies can be generated by (manually) analyzing web pages
- Website: ACM Digital Library
  - publication title ≈ page title (in some cases: substring)
  - author names with hyperlinks (→ inanchor:)
  - publication Id (URL parameter) contains venue Id (→ inurl:)
- Website: Citeseer
  - publication title = page title
  - author names are concatenated (→ pattern), no hyperlinks
  - publication Id ”useless”

Search strategy generation by page analysis

- Example-based query learning
  - input: n DWH entities + m corresponding web pages (entities)
  - output: algorithm for query generation
  - intention: algorithm also works for other DWH entities
- Analyze example pages regarding “searchable“ elements
  - content: keywords, phrases (““), pattern (“a * b“)
  - structure: URL (site:), linked pages (link:)
  - content + structure: page title (intitle:), anchor text (inanchor:)
- Determine precision and recall
  - for example pages (training data)
  - for test data
- Find optimum (w.r.t. to minimal number of query size)
  - overfitting

Conclusion and future work

- Web data integration can be crucial for data warehouses
- Automatic and efficient querying for web entities
- Two approaches
  - entity search engine $\rightarrow$ query relaxation
  - general search engine $\rightarrow$ page analysis
- Algorithms
  - elaboration (bulk search)
  - implementation
- Evaluation for different scenarios