UNIVERSITÄT LEIPZIG

Comparing the Scientific Impact of Conference and Journal Publications in Computer Science

Erhard Rahm

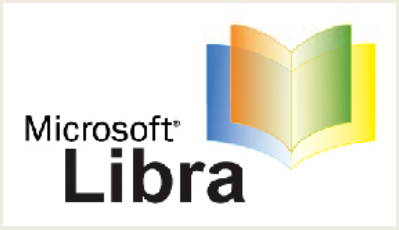
http://dbs.uni-leipzig.de

Academic Publishing in Europe (APE) Conf., 2008, Berlin
Jan. 23, 2008

Citation analysis

- Citation analysis is increasingly used to measure scientific impact of
  - Journals (impact factor)
  - Authors
  - Institutions
- JCR impact factors limited to journals
- Much computer science research is published only in conferences
  - Need to consider citations from / to (refereed) conference publications
- Citation analysis is a huge data integration problem
  - Need to automate as much as possible with good data quality

MS Libra statistics (Dec. 2007)



http://libra.msra.cn

Computer Science Directory (Overall)

	#venues	#papers (all)	#cited (all)	#papers (top 100 venues)	#cited (top100 venues)
journals	471	321.000	1.655.000	190.000	1.434.000
Conference / workshop series	2.297	585.000	1.752.000	167.000	1.216.000

Agenda

- Motivation
- In-depth comparison for CS publications on databases
  - Data sources
  - Conference vs. journal impact factors
  - Citation skew, rankings (nation, institution)
- Data integration of bibliographic web data
  - MOMA framework for record matching
  - Online citation service (OCS)
- Summary

Citation analysis of database publications*

- 10 years: 1994 – 2003
- 5 venues:
  - 2 conference series (ACM SIGMOD, VLDB),
  - 3 journals (ACM TODS, VLDB Journal, Sigmod Record)
- Evaluation using 2005 and 2007 citation data

dblp.uni-trier.de
COMPUTER SCIENCE BIBLIOGRAPHY

- good coverage of CS venues
- manually curated, good quality
- no citation counts

Google Scholar BETA

- many citations
- very good coverage of computer science research
- data quality problems (duplicates, …) due to automatic information extraction

* Rahm, E., A. Thor: *Citation analysis of database publications*, ACM Sigmod Record, Dec. 2005

Further Citation Sources

SCOPUS

ISI Web of Knowledge℠

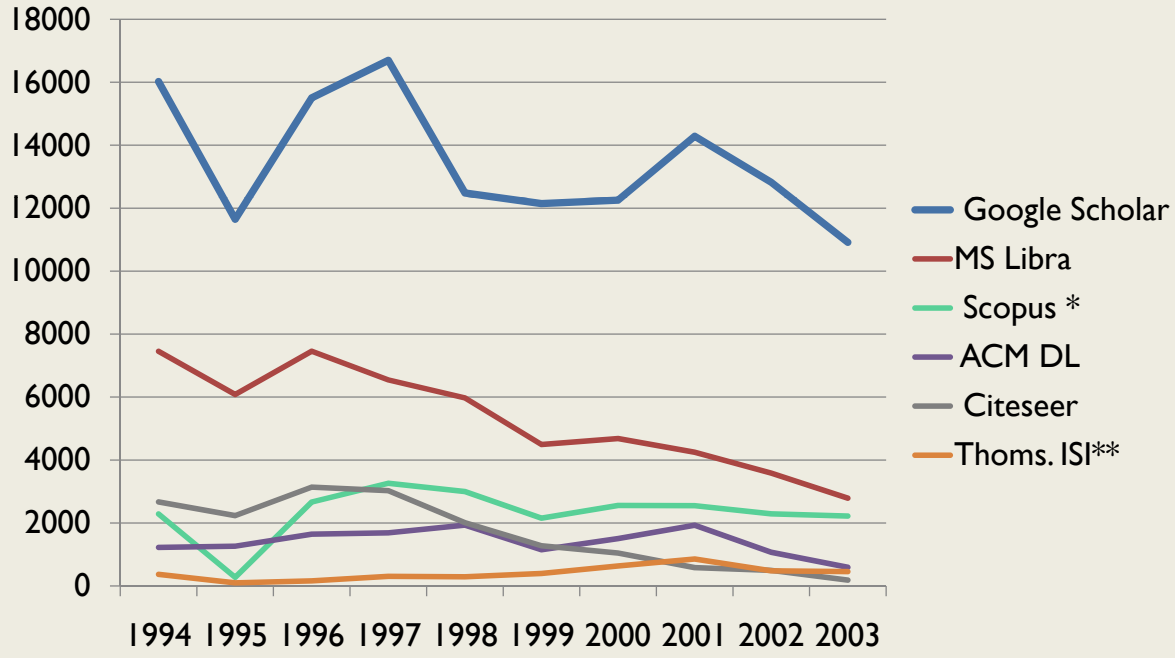
Microsoft Libra

CiteSeer.IST
Scientific Literature Digital Library

ACM Digital Library

PORTAL
HeBIS - Consortium
THE ACM DIGITAL LIBRARY

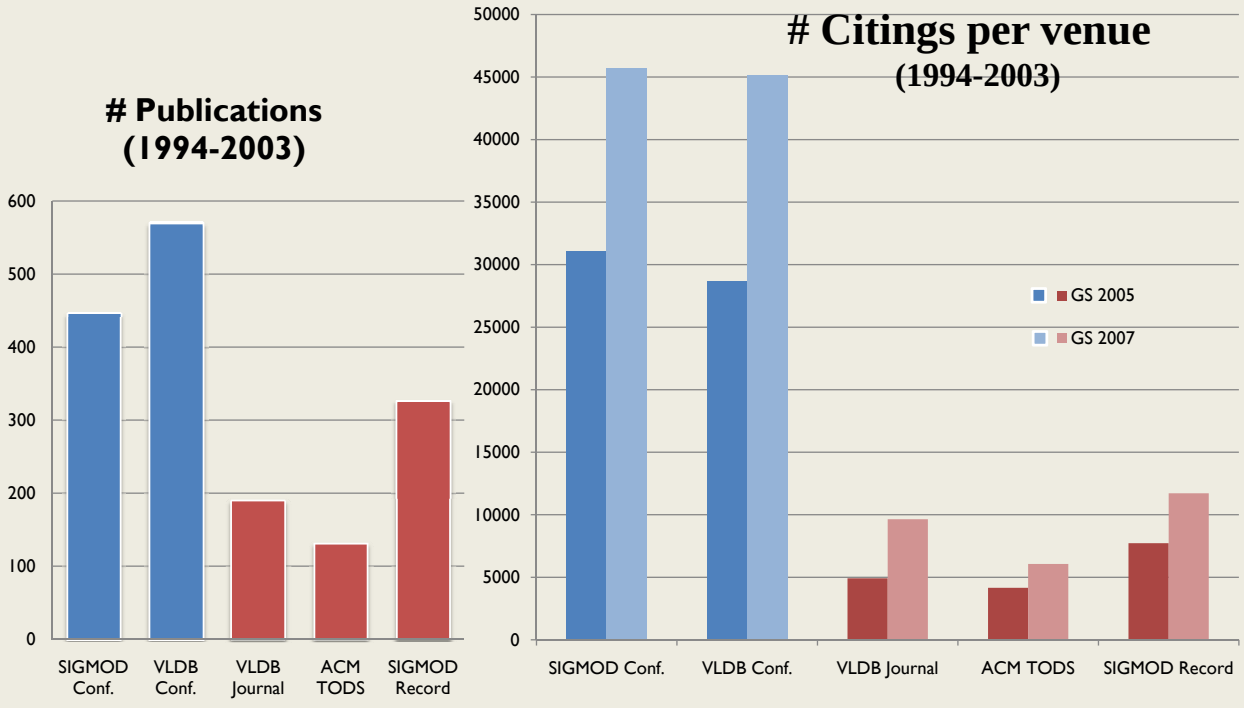
#citings per source
(to papers of considered venues and years)



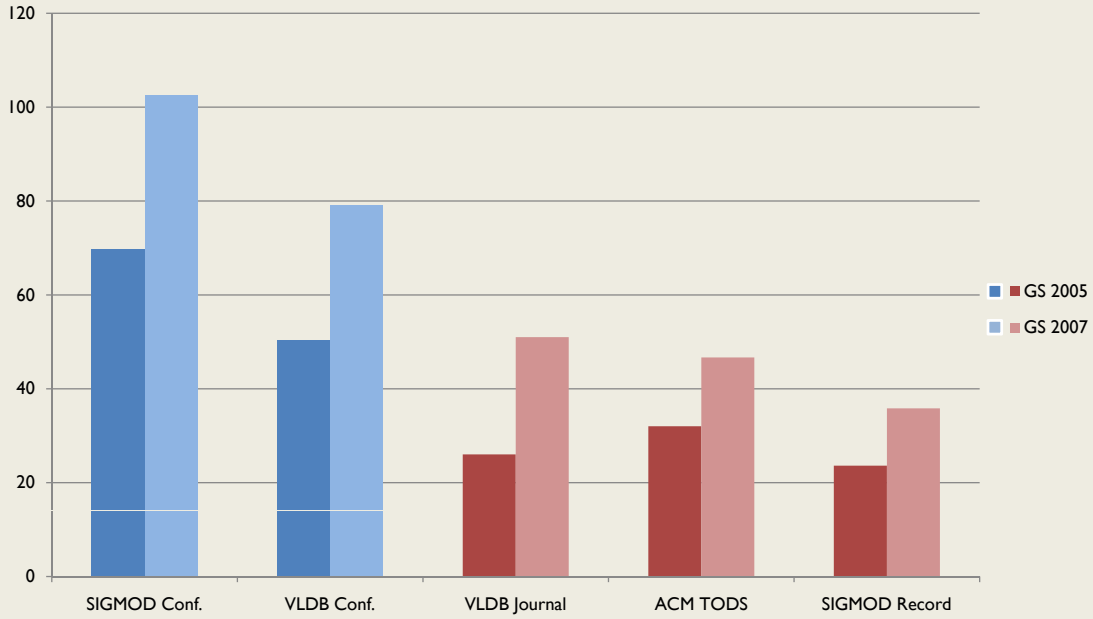
as of Dec. 2007

*Scopus does not cover VLDB conf
** ISI does not cover conferences; VLDBJ /SR since 1998/2000

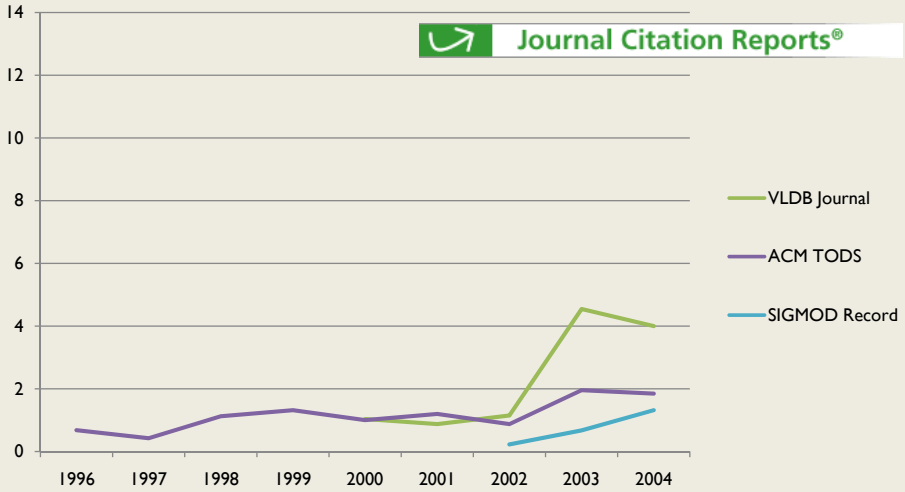
Conferences vs. Journals



Conf. vs. Journals: #citings per paper

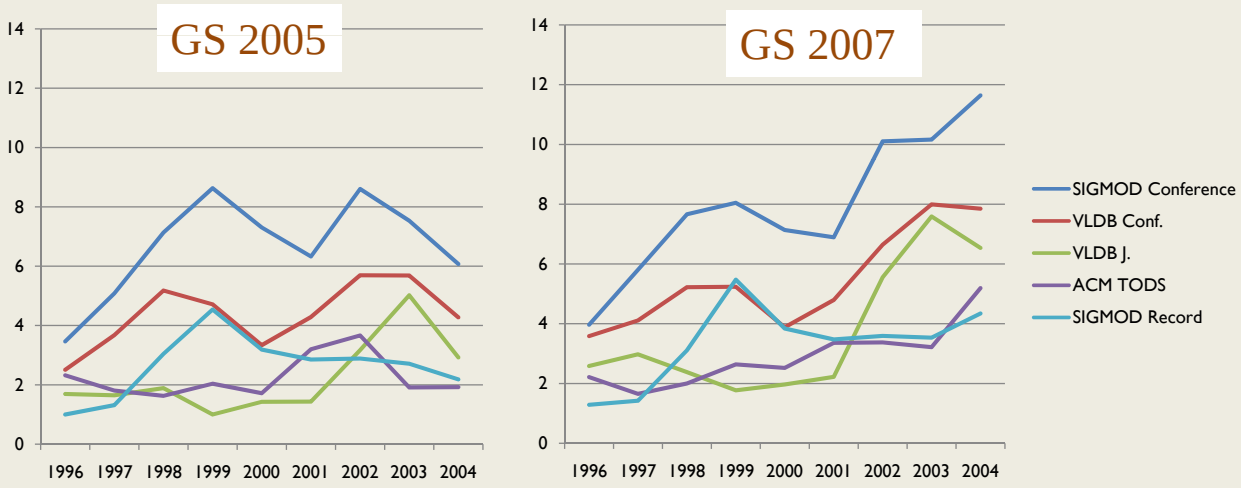


JCR impact factors for journals



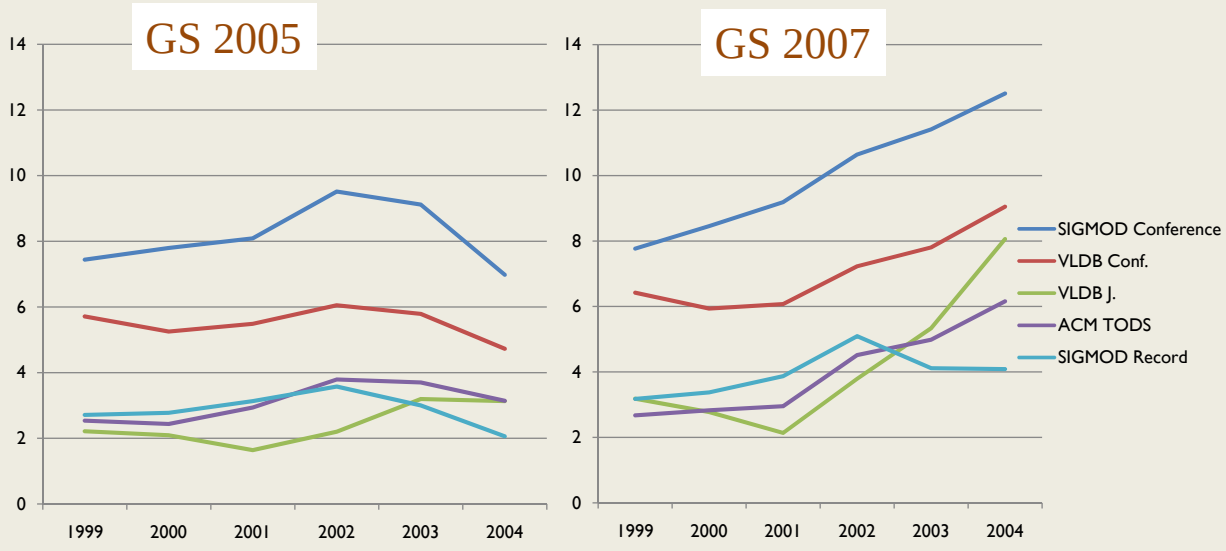
- Journal impact factor IF(X) = average #citings in year X for a journal article published in the 2 preceding years X-1 and X-2
- IF can also be determined for annual conference series
- Can be generalized to articels from k preceding years (e.g. k=5)

GS-based impact factors



- Consider only citing GS publications with year (ca. 77%)
- SIGMOD conf. > VLDB conf. > Journals
- 2007 data: higher impact factors than 2005 and than using JCR

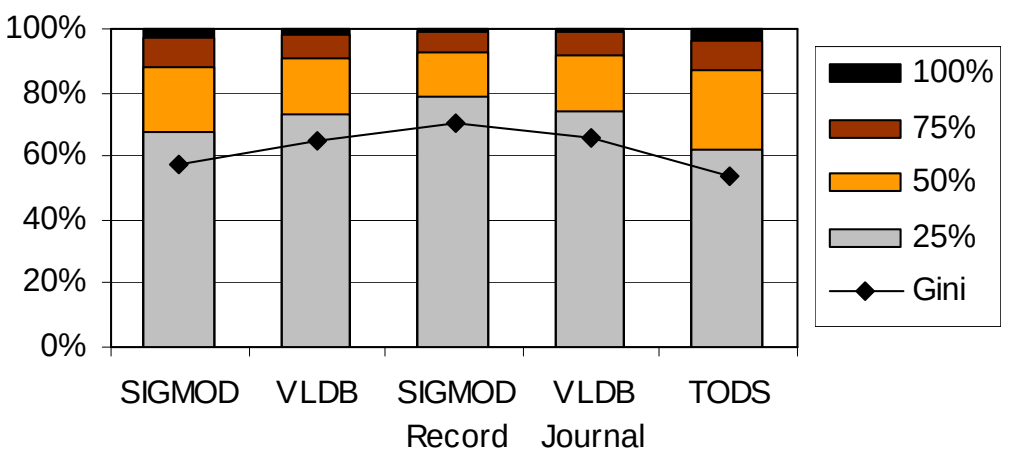
GS-based impact factors (5 years)



- Impact factors more stable for 5 years
- Conferences maintain higher impact than journals

Citation skew

- Citation distribution (splitted by quarters)
  - 25% top referenced publications → 60-80% citings
  - SR has highest skew, TODS is most balanced



Aggregated Citation Frequencies

	Country	# Cit.	in %	# Pub.
1.	USA	51783	72.7	599
2.	Germany	4445	6.2	74
3.	Canada	3342	4.7	38
4.	France	2255	3.2	31
5.	Italy	2079	2.9	25
6.	Israel	858	1.2	6
7.	Japan	753	1.1	8
8.	Switzerland	699	1.0	13
9.	Denmark	655	0.9	8
10.	Greece	623	0.9	14

Table 5: Citations by country

	Institution	# Cit.	# Pub.
1.	IBM	9593	73
2.	Stanford University	7064	63
3.	University of Wisconsin-Madison	5150	61
4.	Bell Labs & AT&T Labs	4573	59
5.	University of Maryland	3299	34
6.	Microsoft	2411	27
7.	University of California, Berkely	1925	25
8.	INRIA (France)	1887	22
9.	University of Washington	1506	16
10.	University of Munich (Germany)	1367	15

Table 6: Citations by institution

- based on institution of first author
- only papers with at least 20 citings (w/o self-citings) are considered

Agenda

- Motivation
- In-depth comparison for CS publications on databases
  - Data sources
  - Conference vs. journal impact factors
  - Citation skew, nation ranking, institution ranking
- Data integration of bibliographic web data
  - MOMA framework for record matching
  - Online citation service (OCS)
- Summary

Matching objects in web sources

```
@article{DBLP:journals/vldb/RahmB01,
author = {Erhard Rahm and Philip A. Bernstein},
title  = {A survey of approaches to automatic schema matching.}
journal= {VLDB J.}, year   = {2001}, ...
```

DBLP

A survey of approaches to automatic schema **matching** - group of 25 »
EJ Rahm, PAJ Bernstein - The VLDB Journal The International Journal on Very Large
... In the next section, we summarize some example applica- tions of **schema match**
5 provides a classification of different ways to perform Match **automatically**. ...
Cited by 585 - Web Search

Google Scholar

Information Fusion

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

ACM

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

Additional Information: abstract citings index terms collaborative colleagues peer to peer

Object matching framework MOMA

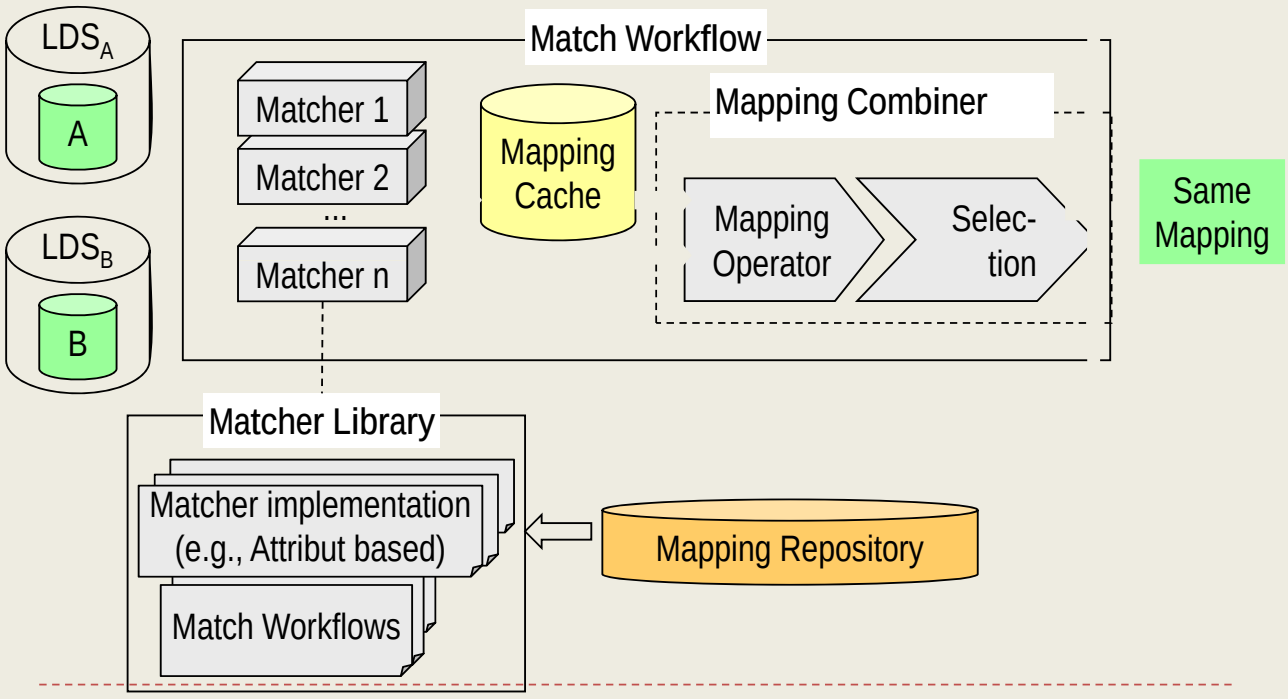
- MOMA = **M**apping based **O**bject **Ma**tching*
- Object consolidation framework
  - Matching objects from 2 sources
  - Generation of instance mappings (correspondences)
  - Special case: duplicate detection within 1 source (generation of self-mapping)
- Key features
  - Extensible matcher library
  - Mapping combination
  - Construction of match workflows
  - Storage of mappings for reuse in other match problems
- Implemented within iFuice data integration platform

$Source_A$	$Source_{A'}$	Sim
a_1	a'_1	1
a_2	a'_1	0.9
a_3	a'_3	0.8

same-mapping for authors

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

MOMA Architecture



On-demand citation analysis

- On-demand citation service (OCS)*
  - What are the most cited papers of conference X?
  - What is the average citation number of publications from 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 GS, Citeseer or Scopus
  - DBLP, Pubmed: high bibliographic data quality
  - GS: large coverage of citations counts
- Query problem: Given a set of DBLP publications → How to find the corresponding GS publications?
  - Query GS and match DBLP-GS

*Thor, Aumueller, Rahm: *Data Integration Support for Mashups*. Proc. IIWeb, 2007

Online Citation Service: Result overview

	Title	Authors	Venue	Year	Citation ▼
⊕	A survey of approaches to automatic schema matching.	Erhard Rahm, Philip A. Bernstein	VLDB J.	2001	1076
⊕	Generic Schema Matching with Cupid.	Jayant Madhavan, Philip A. Bernstein, Erhard Rahm	VLDB	2001	659
⊖	Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching. S Melnik, H Garcia-Molina, E Rahm: *Similarity* 149 *Flooding: A Versatile Graph Matching Algorithm* (2002) S Melnik, H Garcia-Molina, E Rahm: *Similarity* 239 *flooding: a versatile graph matching algorithm and itsapplication to schema matching* (2002) S Melnik, H Garcia-Molina, E Rahm: *Similarity* 1 *Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching* (2002) S Melnik, H Garcia-Molina, E Rahm: *Similarity* 1 *flooding: A versatile graph matching algorithm and its application to schema matching. In* (2002)	Sergey Melnik, Hector Garcia-Molina, Erhard Rahm	ICDE	2002	390
⊕	COMA - A System for Flexible Combination of Schema Matching Approaches.	Hong Hai Do, Erhard Rahm	VLDB	2002	306
⊕	Data Cleaning: Problems and Current Approaches.	Erhard Rahm, Hong Hai Do	IEEE Data Eng. Bull.	2000	234

Bibliographic data from DBLP

Sum of GS citations

Corresponding GS publications

OCS example: Top conference papers

OCS result for venue SIGMOD Conference 2005

- Found 118 GS publications for 107 DBLP publications.
- No GS publications found for 10 DBLP publications.
- Overall: 117 DBLP publications having 1993 citations.
- Average: 17,0 citations per publication.
- H-Index: 27
- Match configuration: 80% title similarity, max. 0 year(s) difference, 50% author similarity.

	Title	Authors	Venue	Year	Citation ▼
✥	Incognito: Efficient Full-Domain K-Anonymity.	Kristen LeFevre, David J. DeWitt, Raghu Ramakrishnan	SIGMOD Conference	2005	109
✥	Reference Reconciliation in Complex Information Spaces.	Xin Dong, Alon Y. Halevy, Jayant Madhavan	SIGMOD Conference	2005	94
✥	Middleware based Data Replication providing Snapshot Isolation.	Yi Lin, Bettina Kemme, Marta Patiño-Martínez, Ricardo Jiménez-Peris	SIGMOD Conference	2005	72
✥	Schema and ontology matching with COMA++.	David Aumueller, Hong Hai Do, Sabine Massmann, Erhard Rahm	SIGMOD Conference	2005	63
✥	Deriving Private Information from Randomized Data.	Zhengli Huang, Wenliang Du, Biao Chen	SIGMOD Conference	2005	56
✥	Robust and Fast Similarity Search for Moving Object Trajectories.	Lei Chen 0002, M. Tamer Özsu, Vincent Oria	SIGMOD Conference	2005	52
✥	Tributaries and Deltas: Efficient and Robust Aggregation in Sensor Network Streams.	Amit Manjhi, Suman Nath, Phillip B. Gibbons	SIGMOD Conference	2005	49

OCS example: Top journal papers

OCS result for venue Bioinformatics 2005

- Found 489 GS publications for 474 DBLP publications.
- No GS publications found for 300 DBLP publications.
- Overall: 774 DBLP publications having 9946 citations.
- Average: 12,9 citations per publication.
- H-Index: 43
- Match configuration: 80% title similarity, max. 0 year(s) difference, 50% author similarity.

	Title	Authors	Venue	Year	Citation ▼
✥	Haploview: analysis and visualization of LD and haplotype maps.	Jeffrey C. Barrett, B. Fry, Julian Maller, Mark Daly	Bioinformatics	2005	1037
✥	MatInspector and beyond: promoter analysis based on transcription factor binding sites.	K. Cartharius, Kornelie Frech, Korbinian Grote, B. Klocke, M. Haltmeier, Andreas Klingenhoff, Matthias Frisch, M. Bayerlein, Thomas Werner	Bioinformatics	2005	166
✥	Ontological analysis of gene expression data: current tools, limitations, and open problems.	Purvesh Khatri, Sorin Draghici	Bioinformatics	2005	146
✥	HyPhy: hypothesis testing using phylogenies.	Sergei L. Kosakovsky Pond, Simon D. W. Frost, Spencer V. Muse	Bioinformatics	2005	146
✥	Outcome signature genes in breast cancer: is there a unique set?.	Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, Eytan Domany	Bioinformatics	2005	145
✥	ProtTest: selection of best-fit models of protein evolution.	Federico Abascal, Rafael Zardoya, David Posada	Bioinformatics	2005	134
✥	RDP2: recombination detection and analysis from sequence alignments.	Darren Martin, C. Williamson, David Posada	Bioinformatics	2005	118
✥	PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis.	Jennifer L. Gardy, M. R. Laird, Fei Chen, S. Rey, C. J. Walsh, Martin Ester, Fiona S. L. Brinkman	Bioinformatics	2005	112

Agenda

- Motivation
- In-depth comparison for CS publications on databases
  - Data sources
  - Conference vs. journal impact factors
  - Citation skew, nation ranking, institution ranking
- Data integration of bibliographic web data
  - MOMA framework for record matching
  - Online citation service (OCS)
- Summary

Summary

- Large scientific impact of conference publications in computer science
  - Must be considered for a meaningful citation analysis
  - In some fields, e.g. database research, top conferences receive many more citings than top journals
- Impact factors should be extended to major conferences
- #citings are highly skewed within venues -> need for individual (per author/organization etc.) impact analysis
  - not just #publications and general venue impact
- Need für improved data integration on heterogeneous data sources (more automatic, high data quality)
- U Leipzig: new research prototypes for data integration, object matching and on-demand citation analysis