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

<table>
<thead>
<tr>
<th></th>
<th>#venues</th>
<th>#papers (all)</th>
<th>#cited (all)</th>
<th>#papers (top 100 venues)</th>
<th>#cited (top 100 venues)</th>
</tr>
</thead>
<tbody>
<tr>
<td>journals</td>
<td>471</td>
<td>321.000</td>
<td>1.655.000</td>
<td>190.000</td>
<td>1.434.000</td>
</tr>
<tr>
<td>Conference / workshop series</td>
<td>2.297</td>
<td>585.000</td>
<td>1.752.000</td>
<td>167.000</td>
<td>1.216.000</td>
</tr>
</tbody>
</table>

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

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

Further Citation Sources

- good coverage of CS venues
- manually curated, good quality
- no citation counts
- many citations
- very good coverage of computer science research
- data quality problems (duplicates, …) due to automatic information extraction

ACM Digital Library

---

* Rahm, E., A. Thor: Citation analysis of database publications, ACM Sigmod Record, Dec. 2005
# Citings per Source
(to papers of considered venues and years)

![Graph showing citings per source across years for different sources such as Google Scholar, MS Libra, Scopus, ACM DL, Citeseer, and Thompson. ISI.](image)

- As of Dec. 2007
- *Scopus does not cover VLDB conf
- **ISI does not cover conferences, VLDBJ/SR since 1998/2000

---

Conferences vs. Journals

![Bar graph comparing publications and citings per venue between conferences and journals from 1994 to 2003.](image)

- # Citings per Venue (1994-2003)
- SIGMOD Conf, VLDB Conf, VLDB Journal, ACM TODS, SIGMOD Record

---

*7th page (natural text)*
Conf. vs. Journals: #citings per paper

Conf. vs. Journals: #citings per paper

JCR impact factors for journals

- Journal impact factor $IF(X) = \text{average #citings in year } X$ 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 articles 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 (split by quarters)
  - 25% top referenced publications → 60-80% citings
  - SR has highest skew, TODS is most balanced

![Citation distribution chart](chart.png)

Aggregated Citation Frequencies

<table>
<thead>
<tr>
<th>Country</th>
<th># Cit.</th>
<th>in %</th>
<th># Pub.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. USA</td>
<td>51783</td>
<td>77.7</td>
<td>599</td>
</tr>
<tr>
<td>2. Germany</td>
<td>4445</td>
<td>6.2</td>
<td>74</td>
</tr>
<tr>
<td>3. Canada</td>
<td>3342</td>
<td>4.1</td>
<td>38</td>
</tr>
<tr>
<td>4. France</td>
<td>2255</td>
<td>3.2</td>
<td>31</td>
</tr>
<tr>
<td>5. Italy</td>
<td>2079</td>
<td>2.9</td>
<td>25</td>
</tr>
<tr>
<td>6. Israel</td>
<td>232</td>
<td>1.2</td>
<td>6</td>
</tr>
<tr>
<td>7. Japan</td>
<td>733</td>
<td>1.1</td>
<td>8</td>
</tr>
<tr>
<td>8. Switzerland</td>
<td>699</td>
<td>1.0</td>
<td>13</td>
</tr>
<tr>
<td>9. Denmark</td>
<td>635</td>
<td>0.9</td>
<td>8</td>
</tr>
<tr>
<td>10. Greece</td>
<td>623</td>
<td>0.9</td>
<td>14</td>
</tr>
</tbody>
</table>

*Table 5: Citations by country*

<table>
<thead>
<tr>
<th>Institution</th>
<th># Cit.</th>
<th># Pub.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. IBM</td>
<td>9593</td>
<td>73</td>
</tr>
<tr>
<td>2. Stanford University</td>
<td>7064</td>
<td>63</td>
</tr>
<tr>
<td>3. University of Wisconsin-Madison</td>
<td>5150</td>
<td>61</td>
</tr>
<tr>
<td>4. Bell Labs &amp; AT&amp;T Labs</td>
<td>4573</td>
<td>59</td>
</tr>
<tr>
<td>5. University of Maryland</td>
<td>3299</td>
<td>34</td>
</tr>
<tr>
<td>6. Microsoft</td>
<td>2411</td>
<td>27</td>
</tr>
<tr>
<td>7. University of California, Berkeley</td>
<td>1925</td>
<td>25</td>
</tr>
<tr>
<td>8. INRIA (France)</td>
<td>1387</td>
<td>22</td>
</tr>
<tr>
<td>9. University of Washington</td>
<td>1306</td>
<td>16</td>
</tr>
<tr>
<td>10. University of Munich (Germany)</td>
<td>1367</td>
<td>15</td>
</tr>
</tbody>
</table>

*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},
}

A survey of approaches to automatic schema matching - group of 25
EJ Rahm, PA Bernstein - The VLDB Journal The International Journal on Very Large...
In the next section, we summarize some example applications of schema matching. Experiments in classification of different ways to perform match automatically. ...
Cited by 385 - Web Search

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: 324 - 360
Year of Publication: 2001
ISSN: 1066-888X

Authors: Erhard Rahm
          Philip A. Bernstein

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

Additional information: abstract citations index terms collaborative colleagues peer to peer
Object matching framework MOMA

- MOMA = **Mapping based Object Matching**
- 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

<table>
<thead>
<tr>
<th>Source_A</th>
<th>Source_A'</th>
<th>Sim</th>
</tr>
</thead>
<tbody>
<tr>
<td>a_1</td>
<td>a'_1</td>
<td>1</td>
</tr>
<tr>
<td>a_2</td>
<td>a'_1</td>
<td>0.9</td>
</tr>
<tr>
<td>a_3</td>
<td>a'_3</td>
<td>0.8</td>
</tr>
</tbody>
</table>

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

### 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 1983 citations.
- Average: 17.0 citations per publication.
- H-Index: 27.
- Match configuration: 80% title similarity, max. 0 year(s) difference, 50% author similarity.

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Venue</th>
<th>Year</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Reconciliation in Complex Information Spaces</td>
<td>Xin Dong, Alon Y. Halevy, Jayant Madhavan</td>
<td>SIGMOD Conference</td>
<td>2005</td>
<td>94</td>
</tr>
<tr>
<td>Middleware-based Data Replication providing Snapshot Isolation</td>
<td>Yi Liu, Bettina Kemme, Maria Patino-Martinez, Ricardo Jimenez-Peiris</td>
<td>SIGMOD Conference</td>
<td>2005</td>
<td>72</td>
</tr>
<tr>
<td>Scheme and ontology matching with COMA++</td>
<td>David Aurueler, Hong Hai De, Sabine Massmann, Emilard Rahm</td>
<td>SIGMOD Conference</td>
<td>2005</td>
<td>63</td>
</tr>
<tr>
<td>Deriving Private Information from Randomized Data</td>
<td>Zhengli Huang, Wenliang Du, Shao Chen</td>
<td>SIGMOD Conference</td>
<td>2005</td>
<td>56</td>
</tr>
<tr>
<td>Robust and Fast Similarity Search for Moving Object Trajectories</td>
<td>Lei Chen 0002, M. Tarrer Ozsu, Vincent Oria</td>
<td>SIGMOD Conference</td>
<td>2005</td>
<td>52</td>
</tr>
<tr>
<td>Tributaries and Deltas: Efficient and Robust Aggregation in Sensor Network Streams</td>
<td>Amit Manjhi, Suman Nath, Phillip B. Gibbons</td>
<td>SIGMOD Conference</td>
<td>2005</td>
<td>49</td>
</tr>
</tbody>
</table>

### OCS example: Top journal papers

**OCS result for venue Bioinformatics 2005**

- Found 429 GS Publications for 474 DBLP publications.
- No GS Publications found for 300 DBLP publications.
- Overall: 774 DBLP publications having 9346 citations.
- Average: 12.9 citations per publication.
- H-Index: 43.
- Match configuration: 80% title similarity, max. 0 year(s) difference, 50% author similarity.

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Venue</th>
<th>Year</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haploview: analysis and visualization of LD and haplotype maps</td>
<td>Jeffrey C. Barrett, B. Fry, Julian Maller, Mark Daly</td>
<td>Bioinformatics</td>
<td>2005</td>
<td>1037</td>
</tr>
<tr>
<td>MultiInspector and beyond: promoter analysis based on transcription factor binding sites</td>
<td>K. Caranthos, Konstantina Troch, Korbinian Grube, B. Klock, M. Heitmeier, Andreas Klingenhorst, Matthias Fissich, M. Elsner, Thomas Werner</td>
<td>Bioinformatics</td>
<td>2005</td>
<td>168</td>
</tr>
<tr>
<td>Ontological analysis of gene expression data: current tools, limitations, and open problems</td>
<td>Purvesh Khatri, Sorin Draghici</td>
<td>Bioinformatics</td>
<td>2005</td>
<td>148</td>
</tr>
<tr>
<td>Outcome signature genes in breast cancer: is there a unique set?</td>
<td>Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, Eytan Domany</td>
<td>Bioinformatics</td>
<td>2005</td>
<td>145</td>
</tr>
<tr>
<td>ProTest: selection of best-fit models of protein evolution</td>
<td>Federico Abascal, Rafael Zardoya, David Pesada</td>
<td>Bioinformatics</td>
<td>2005</td>
<td>134</td>
</tr>
<tr>
<td>RDP2: recombination detection and analysis from sequence alignments</td>
<td>Darren Martin, C. Williamson, David Pesada</td>
<td>Bioinformatics</td>
<td>2005</td>
<td>119</td>
</tr>
<tr>
<td>PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis</td>
<td>Jennifer L. Gandy, M. R. Laird, Fe Chen, S. Rey, C. J. Walsh, Martin Ester, Fiona S. L. Brieleman</td>
<td>Bioinformatics</td>
<td>2005</td>
<td>112</td>
</tr>
</tbody>
</table>
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