Integrating and exploiting public metadata sources in a bibliographic information system

Ralf Schenkel
Dblp Overview

Recent activity: adding links to other authority providers, esp. ORCID and WikiData

~4 million publications, ~2 million authors, ~400,000 new publications per year
How is publication data added to dblp?
Dblp Data Ingestion Pipeline

Data Quality Control:
- Selection
- Correction
- Author disambiguation

Publishers

Web

Meta data in (some) structured form

Source Monitoring

Data Harvesting

Extracted meta data

HTML
Outline

• Meta Data Harvesting
• Author Disambiguation
• Existing Metadata Collections
• Citations
Harvesting is much more difficult now

Need to interact with Web site, parsing static HTML not enough
Harvesting is much more difficult now.

Successful harvesting needs to implement Javascript.

**OXPath-based Data Acquisition for dblp**
C Michelis, RR Fayzrakhmanov, M Ley... - ... (JCDL), 2017 ACM ..., 2017 - ieeexplore.ieee.org
We demonstrate how the contemporary problems of **data acquisition** for dblp can be tackled with **OXPath**. It enables web **data** extraction and wrapper maintenance for heterogeneous **data** sources on a simple declarative level. Its features render it a feasible instrument to

Zitieren

**MLA**

**APA**

**ISO 690**

Successful harvesting needs to implement Javascript.
Monitoring & harvesting: OXPath

Extension of XPath by University of Oxford (Georg Gottlob et al.)

- **Actions**: fill in forms, click buttons
- **Extraction**: specify what should be harvested
- **Transformation**: specify target XML format
- **Iteration**: loops, e.g., for paginated content

Example: Navigating Google Scholar

```
1 doc("https://scholar.google.com")
```
Example: Navigating Google Scholar

```xml
1. doc("https://scholar.google.com")
2. //*[@role="search"]//input[@type="text"]/"XPath"
```

Google Scholar

Stand on the shoulders of giants
Example: Navigating Google Scholar

```
1. doc("https://scholar.google.com")
2. //*[@role="search"]//input[@type="text"]/"XPath"
3. ./following-sibling::button/{click/}
```
Example: Navigating Google Scholar

**OXPath**

A language for scalable data extraction, automation, and crawling on the deep web

T Furche, G Gottlob, G Grasso, C Schallhart, A Sellers - The VLDB Journal, 2013 - Springer

Abstract The evolution of the web has outpaced itself: A growing wealth of information and increasingly sophisticated interfaces necessitate automated processing, yet existing automation and data extraction technologies have been overwhelmed by this very growth...

Cited by 46 Related articles More
Example: Navigating Google Scholar

**OXPath Expression**

```xml
1 doc("https://scholar.google.com")
2 //[@role="search"]//input[@type="text"]/{"OXPath"}
3 //following-sibling::button/{click/}
4 //[@id="gs_ylo_btn"]/click
```

**OXPath: A language for scalable data extraction, automation, and crawling on the deep web**

T Furche, G Gottlob, G Grasso, C Schallhart, A Sellers - The VLDB Journal, 2013 - Springer

Abstract The evolution of the web has outpaced itself: A growing wealth of information and increasingly sophisticated interfaces necessitate automated processing, yet existing automation and data extraction technologies have been overwhelmed by this very growth. ...

Cited by 46 Related articles More

Create alert
Example: Navigating Google Scholar

OXPath Expression

```xml
1 doc("https://scholar.google.com")
2 //*[@role="search"]//input[@type="text"]/{"OXPath"}
3 ../../following-sibling::button/{click/}
4 //*[@id="gs_ylo_btn"]/click}
5 //following::*[@id="gs_ylo_md"]//a[contains(.,
   "2016")]/{click/}
```
Example: Navigating Google Scholar

**OXPath Expression**

```
1 doc("https://scholar.google.com")
2 /*[@role="search"]//input[@type="text"]/"OXPath"
3 ./following-sibling::button/{click/}
4 /*[@id="gs_ylo_btn"]/"click"/}
5 //following::*[@id="gs_ylo_md"]//a[contains(.,
"2016")]/"click"/}
```
Example: Navigating Google Scholar

OXPath Expression

```xml
1 doc("https://scholar.google.com")
2  /*[@role="search"]//input[@type="text"]/{"OXPath"}
3  ../../../following-sibling::button/{click/}
4  /*[@id="gs_ylo_bnt"]/{click/}
5  //following::*[@id="gs_ylo_md"]/a[contains(.,
6  "2016")]//{click/}
7  //div[@class="gs_ri"]/h3/a:<title=string(.)>
```

XML Output

```xml
1 <?xml version="1.0" encoding="UTF-8"?>
2 <results>
3  <title>Tim Furche, Georg Gottlob, [...]</title>
4 </results>
```
Example: Navigating Google Scholar

**OXPath Expression**

```xml
1 doc("https://scholar.google.com")
2 /*[@role="search"]//input[@type="text"]/{"OXPath"}
3 ./../following-sibling::button/{click/}
4 /*[@id="gs_ylo_btn"]/{click} 
5 //following::*[@id="gs_ylo_md"]/a[contains(., "2016")]/{click/}
6 //*[contains(@class, "next")]/{click/})*
7 //div[@class="gs_ri"]/h3/a:<title=string(.)>
```

**XML Output**

```xml
1 <?xml version="1.0" encoding="UTF-8"?>
2 <results>
3 <title>Tim Furche, Georg Gottlob, [...]</title>
4 </results>
```
Example: Navigating Google Scholar

**OXPath Expression**

```xml
1 doc("https://scholar.google.com")
2 /*[@role="search"]//input[@type="text"]/{"OXPath"}
3 ./following-sibling::button/{click/}
4 /*[@id="gs_ylo_btn"]/a/{click/}
5 //following::*[@id="gs_ylo_md"]//a[contains(., "2016")]/a/{click/}
6 /*[contains(@class, "next")]/*/a/{click/}
7 //div[@class="gs_ri"]/h3/a: {title=string(.)}
```

**XML Output**

```xml
1 <?xml version="1.0" encoding="UTF-8"?>
2 <results>
3 <title>Tim Furche, Georg Gottlob, [...]</title>
4 </results>
```
Example: Navigating Google Scholar

```xml
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <title>Tim Furche, Georg Gottlob, [...]</title>
  <title>Special Issue: Big Data [...]</title>
</results>
```
Example: Navigating Google Scholar

**OXPath Expression**

```xml
1 doc("https://scholar.google.com")
2 //*[@role="search"]//input[@type="text"]/{"OXPath"}
3 //following-sibling::button/{click/}
4 //*[@id="gs_ylo_btn"]/{click/}
5 //following::*[id="gs_ylo_md"]/a[contains(., "2016")]/{click/}
6 //*[contains(@class, "next")]/{click/}*
7 //div[@class="gs_ri"]/h3/a:<title=string(.)>"
```

**XML Output**

```xml
1 <?xml version="1.0" encoding="UTF-8"?>
2 <results>
3  <title>Tim Furche, Georg Gottlob, [...]</title>
4  <title>Special Issue: Big Data [...]</title>
5  <!--[...]-->  
6 </results>
```
Advantages of OXPath

• **More powerful** than plain XPath: actions, extraction, transformation, iteration
• Possible to extract from **several pages** in one query
• Somewhat **robust to changes in layout**

Now in productive use at dblp
Outline

• Meta Data Harvesting
• Author Disambiguation
• Existing Metadata Collections
• Citations
Author Disambiguation: Homonyms

Multiple persons with the same name in the same profile

Hard problem for an algorithm (even for a human), may use
- paper titles/topics
- common coauthors
- publication years
- publication venues
- ...

DBLP computer science bibliography

Christian Sturm

> Home > Persons

2017
- [c19] Christian Sturm, Stefan Schönig, Claudio Di Cicco: Distributed Multi-Perspective Declare Discovery. PIM (Demos) 2017

2016

2015

2014
- [c16] Christian Sturm, Alice Oh, Sebastian Linxen, José L. Abdelnour-Nocera, Susan M. Dray, Katharina Reincke: How WEIRD is HCI?: Extending HCI Principles to other Countries and Cultures. CHI Extended Abstracts 2015: 2425-2428

2013

2013
Author Disambiguation: Homonyms

Affiliations would be useful, but usually not available
Author Disambiguation: Synonyms

The same person with different names in different profiles

Identify pairs of candidate profiles such that

- Small name difference
- Common coauthors
- Common venues
- Common topics
- ...

+ manual corrections
Author Disambiguation: Synonyms

The same person with different names in different profiles

Identify pairs of candidate profiles such that

- Small name difference
- Common coauthors
- Common venues
- Common topics
- ...

Last resort...
Observation:
Additional meta data can improve the quality of the detection of synonyms and homonyms.
Example: ORCID

- Provides **persistent digital identifier** for authors
- Includes additional **author-provided meta data** about publications, affiliations, ...
- API & dumps

Data often incomplete or not fully correct
ORCID for Homonym Detection

ORCIDs of authors with this name who claimed at least one publication in this profile

After import of 625,000 ORCIDs: 1,000 candidates for homonyms

Top candidate: 10 persons in one profile

BUT:
Profiles with common ORCID include papers from the same author (but maybe other papers as well due to homonyms)

After import of 625,000 ORCIDS: 4,500 candidates for synonyms

Top candidate: 6 profiles with same ORCID

X. Xu, X. W. Xu, X. William Xu, Xun Xu, Xun W. Xu, Xun William Xu

Source: Scopus to ORCID
Outline

- Meta Data Harvesting
- Author Disambiguation
- Existing Metadata Collections
- Citations
Useful information not (always) in dblp

- Author affiliations
- Keywords
- Topics
- Abstracts
- Full texts
- Incoming and outgoing citations
- Performance indicators
- ...
Sources for Bibliographic Metadata

- Dblp.org
- Semantic Scholar
- Aminer Open academic graph (includes Microsoft Academic Graph)
- Springer SciGraph
- CrossRef
- OpenCitations
- ...
## Overview: properties of sources

<table>
<thead>
<tr>
<th></th>
<th>Semantic Scholar</th>
<th>OAG - Aminer</th>
<th>OAG – Microsoft Academic</th>
<th>Springer SciGraph</th>
<th>CrossRef</th>
<th>Open Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>coverage</strong></td>
<td>CS</td>
<td>universal</td>
<td>universal</td>
<td>Springer</td>
<td>universal</td>
<td>universal</td>
</tr>
<tr>
<td><strong># publs</strong></td>
<td>7.2 million</td>
<td>154 million</td>
<td>166 million</td>
<td>~12 million</td>
<td>96 million</td>
<td>~300,000</td>
</tr>
<tr>
<td><strong>in dblp</strong></td>
<td>1.45 million</td>
<td>3.46 million</td>
<td>3.57 million</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td><strong>access</strong></td>
<td>dump</td>
<td>dump</td>
<td>dump</td>
<td>API, dump</td>
<td>API</td>
<td>API, dump</td>
</tr>
<tr>
<td><strong>size</strong></td>
<td>20 GB</td>
<td>39 GB</td>
<td>103 GB</td>
<td>~200 GB</td>
<td>-</td>
<td>3.5 GB</td>
</tr>
<tr>
<td><strong>Keywords</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Topics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Abstracts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>partial</td>
<td></td>
</tr>
<tr>
<td><strong>Full-texts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Citations</strong></td>
<td></td>
<td></td>
<td></td>
<td>planned</td>
<td>partial</td>
<td></td>
</tr>
<tr>
<td><strong>DOIs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Author aff.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>email</td>
<td>partial</td>
</tr>
<tr>
<td><strong>Funding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>partial</td>
</tr>
</tbody>
</table>
SpringerNature SciGraph

• Linked Open Data with rich ontology
• funders, research projects, conferences, affiliations and publications from SpringerNature and partners
• extension to citations, patents, clinical trials and usage numbers planned
• CC BY 4.0 license (NC for abstracts)

http://www.springernature.com/gp/researchers/scigraph
OpenCitations

- Initiative for Open Citations (I4OC): collaboration between scholarly publishers and researchers to promote the *unrestricted availability of scholarly citation data*
- As of January 2018, **50% of publications at CrossRef** with open references
- **OpenCitations:**
  publishes open citations from CrossRef as RDF-based collection, using SPAR ontology
CrossRef Example


"reference": [
...

http://api.crossref.org/works/10.1007/978-3-642-25073-6_38
Problems of these Collections

- Update Frequency
- Data Quality
- Completeness / Coverage / Sparsity
Strange names, not linked to a profile

No info on venue, year,…

@article{liang12keep,
  title={How to keep a knowledge base synchronized with its encyclopedia source},
  author={Liang12, Jiaqing and Zhang, Sheng and Xiao134, Yanghua}
}

How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source

Jiaqing Liang\textsuperscript{12}, Sheng Zhang\textsuperscript{1}, Yanghua Xiao\textsuperscript{134*}
\footnotesize
\textsuperscript{1}School of Computer Science, Shanghai Key Laboratory of Data Science
Fudan University, Shanghai, China
\textsuperscript{2}Shuyan Technology, Shanghai, China
\textsuperscript{3}Shanghai Internet Big Data Engineering Technology Research Center, China
\textsuperscript{4}Xiao Research, Shanghai, China

record conf/ijcai/LiangZX17

Requires data cleaning
Data Quality: What is a Publication?

Frequent problem: conference paper + followup journal paper

2012
Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, Gad Markovits:

2009
Anna Shtok, Oren Kurland, David Carmel:
Predicting Query Performance by Query-Drift Estimation. ICTIR 2009: 305-312
Data Quality: What is a Publication?

Frequent problem: conference paper + followup journal paper

*Predicting Query Performance by Query-Drift Estimation.*

**2012**
Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, Gad Markovits:

**2009**
Anna Shtok, Oren Kurland, David Carmel:
Predicting Query Performance by Query-Drift Estimation. ICTIR 2009: 305-312

*Semantic Scholar*

Predicting Query Performance by Query-Drift Estimation

Anna Shtok, Oren Kurland, David Carmel • ICTIR • 2009

Predicting query performance, that is, the effectiveness of a search performed in response to a query, is a highly important and challenging problem. We present a novel approach to this task... (More)

17 39
View on ACM Similar Papers

Cited By

Showing 1-10 of 70 extracted citations

Robust Standard Deviation Estimation for Query Performance Prediction

Haggai Roitman, Shai Erera, Bar Weiner • ICTIR • 2017

cites ICTIR

document score distribution models for query performance inference and prediction

Ronan Cummins • ACM Trans. Inf. Syst. • 2014

cites ICTIR

document score distribution models for query performance inference and prediction

Ronan Cummins • ACM Trans. Inf. Syst. • 2014

cites ICTIR

Query Performance Prediction for Aspect Weighting in Search Result Diversification

Ahmet Murat Ozdemiray, Ismail Sengör Altingövde • CIKM • 2014

cites TOIS

cites TOIS

cites TOIS
Towards Quantifying Coverage: Mapping papers to dblp

Preprocessing: Index all dblp entries in Lucene

Authors
Title
Venue
Year
DOI
...

DOI Index
Lucene

Title Index
Lucene

DOI

Authors

Title
Venue
Year
DOI
...

no DOI

Post-Filter by author overlap, venue similarity, temporal proximity, ...

Mapping quality in general very good, no systematic evaluation yet

key1, 14.1
key2, 12.7
key3, 11.5
...

dblp key or no match
Coverage of dblp and Overlap

- **Semantic Scholar**
  - 0.01 mio
  - 0.02 mio
- **Microsoft Academic**
  - 0.04 mio
  - 0.18 mio
- **AMiner**
  - 0.08 mio
  - 1.97 mio
  - 1.38 mio

- 0.26 mio from dblp missing

[not drawn to scale]
Overlap of dblp and CrossRef

DOI-based match in February 2018

• 4 million publications in dblp
• 3.2 million with DOI
• 3.1 million found in CrossRef
• 600,000 with citations (~15%)
  – 16 million citation instances
  – 4 million mapped based on DOI
  – ~1 million mapped based on reference string
Main Observation:

• All collections are too incomplete or too static to be useful for productive use.

• Initiative for Open Citations has effect, but still limited for computer science
Outline

- Meta Data Harvesting
- Author Disambiguation
- Existing Metadata Collections
- Citations
Bibliometrics: most frequently cited pubs

<table>
<thead>
<tr>
<th>Aminer*</th>
<th>SemanticScholar</th>
<th>Microsoft Academic*</th>
</tr>
</thead>
<tbody>
<tr>
<td>journals/ijcv/Lowe04</td>
<td>journals/ijcv/Lowe04</td>
<td>books/mg/CormenLS01</td>
</tr>
<tr>
<td>journals/tist/ChangL11</td>
<td>journals/tist/ChangL11</td>
<td>books/aw/Goldberg89</td>
</tr>
<tr>
<td>books/aw/Goldberg89</td>
<td>conf/nips/BleiN01</td>
<td>journals/ijcv/Lowe04</td>
</tr>
<tr>
<td>conf/cvpr/DalalT05</td>
<td>journals/carm/DeanG08</td>
<td>journals/corr/BoiyAJ15</td>
</tr>
<tr>
<td>journals/ml/Breiman01</td>
<td>conf/cvpr/DalalT05</td>
<td>journals/tmn/Chekeskyc97</td>
</tr>
<tr>
<td>books/mk/Quinlan93</td>
<td>conf/nips/KrizhevskySH12</td>
<td>journals/sigmobile/Shannon01</td>
</tr>
<tr>
<td>journals/ml/CortesV95</td>
<td>conf/cvpr/TurkP91</td>
<td>journals/tist/ChangL11</td>
</tr>
<tr>
<td>journals/tec/DebAPM02</td>
<td>conf/cvpr/ShiM97</td>
<td>books/daglib/0066829</td>
</tr>
<tr>
<td>books/daglib/0066829</td>
<td>conf/icml/LaffertyMP01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>journals/misq/Davis89</td>
<td>conf/sigcomm/StoicaMHKB01</td>
<td>journals/cacm/Hoare78</td>
</tr>
<tr>
<td>journals/acm/FischerB81</td>
<td>journals/tit/Donoho06</td>
<td>books/mk/Quinlan53</td>
</tr>
<tr>
<td>journals/tit/Donoho06</td>
<td>journals/ml/CortesV95</td>
<td>conf/cvpr/TurkP91</td>
</tr>
<tr>
<td>journals/jcvi/KassWT88</td>
<td>journals/cn/BrinP98</td>
<td>journals/ac/Kothari093</td>
</tr>
<tr>
<td>books/lib/RusselN03</td>
<td>books/ml/Breiman01</td>
<td>conf/13e/StuderAV03</td>
</tr>
<tr>
<td>journals/jcvi/KassWT88</td>
<td>journals/ml/Breiman01</td>
<td>journals/tcn/StoicaMLKKB03</td>
</tr>
<tr>
<td>journals/ml/Breiman96b</td>
<td>journals/ml/Breiman01</td>
<td>books/mg/CormenLS01</td>
</tr>
<tr>
<td>conf/sigmod/AgrawalS93</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>journals/cn/BrinP98</td>
<td>journals/ml/Breiman01</td>
<td>journals/cacm/Hoare78</td>
</tr>
<tr>
<td>conf/cvpr/ViolaJ01</td>
<td>journals/ml/Breiman01</td>
<td>books/mg/CormenLS01</td>
</tr>
<tr>
<td>conf/nips/KrizhevskySH12</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>journals/ml/Breiman96b</td>
<td>journals/ml/Breiman01</td>
<td>journals/cacm/Hoare78</td>
</tr>
<tr>
<td>journals/jacm/Kleinberg99</td>
<td>journals/ml/Breiman01</td>
<td>books/mg/CormenLS01</td>
</tr>
<tr>
<td>conf/middleware/RonstronD01</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>conf/cvpr/ViolaJ01</td>
<td>journals/ml/Breiman01</td>
<td>journals/cacm/Hoare78</td>
</tr>
<tr>
<td>journals/jt/Gupta00</td>
<td>journals/ml/Breiman01</td>
<td>books/mg/CormenLS01</td>
</tr>
<tr>
<td>journals/ml/Breiman96b</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>journals/sigmod/Alamouti98</td>
<td>journals/ml/Breiman01</td>
<td>books/mg/CormenLS01</td>
</tr>
<tr>
<td>journals/jn/AlkylidizSC02</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>journals/jt/LanemanTW04</td>
<td>journals/ml/Breiman01</td>
<td>books/mg/CormenLS01</td>
</tr>
<tr>
<td>journals/ett/Teletar99</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>journals/datavine/Burges98</td>
<td>journals/ml/Breiman01</td>
<td>books/mg/CormenLS01</td>
</tr>
<tr>
<td>journals/pam/BeihummerHK97</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>journals/ml/Quinlan86</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>conf/vldb/AgrawalS94</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>journals/jsac/Haykin05</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
<tr>
<td>journals/ijcvs/Ackerman865</td>
<td>journals/ml/Breiman01</td>
<td>books/swarm/P geliB07</td>
</tr>
</tbody>
</table>

Significantly different ranking derived from different collections – which one should one use?
**Bibliometrics: Most prominent authors**

<table>
<thead>
<tr>
<th>Autor</th>
<th>#paper</th>
<th>#cites</th>
<th>h-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jiawei Han 0001</td>
<td>828</td>
<td>33892</td>
<td>90</td>
</tr>
<tr>
<td>Andrew Zisserman</td>
<td>431</td>
<td>34226</td>
<td>86</td>
</tr>
<tr>
<td>Anil K. Jain</td>
<td>629</td>
<td>39336</td>
<td>86</td>
</tr>
<tr>
<td>Scott Shenker</td>
<td>301</td>
<td>33149</td>
<td>83</td>
</tr>
<tr>
<td>Philip S. Yu</td>
<td>1092</td>
<td>26027</td>
<td>79</td>
</tr>
<tr>
<td>Hector Garcia-Molina</td>
<td>455</td>
<td>20999</td>
<td>79</td>
</tr>
<tr>
<td>Christos Faloutsos</td>
<td>577</td>
<td>25758</td>
<td>76</td>
</tr>
<tr>
<td>Sebastian Thrun</td>
<td>305</td>
<td>21123</td>
<td>73</td>
</tr>
<tr>
<td>Jitendra Malik</td>
<td>248</td>
<td>29638</td>
<td>72</td>
</tr>
<tr>
<td>Don Towsley</td>
<td>626</td>
<td>18699</td>
<td>71</td>
</tr>
<tr>
<td>Ion Stoica</td>
<td>252</td>
<td>26341</td>
<td>71</td>
</tr>
<tr>
<td>Andrew Y. Ng</td>
<td>207</td>
<td>22617</td>
<td>71</td>
</tr>
<tr>
<td>Thomas S. Huang</td>
<td>946</td>
<td>25285</td>
<td>71</td>
</tr>
<tr>
<td>Luc J. Van Gool</td>
<td>686</td>
<td>29314</td>
<td>71</td>
</tr>
<tr>
<td>Michael I. Jordan</td>
<td>415</td>
<td>24734</td>
<td>70</td>
</tr>
<tr>
<td>David E. Culler</td>
<td>237</td>
<td>24351</td>
<td>69</td>
</tr>
<tr>
<td>Georgios B. Giannakis</td>
<td>745</td>
<td>18145</td>
<td>69</td>
</tr>
<tr>
<td>Cordelia Schmid</td>
<td>229</td>
<td>30525</td>
<td>68</td>
</tr>
<tr>
<td>Francisco Herrera</td>
<td>562</td>
<td>17041</td>
<td>68</td>
</tr>
<tr>
<td>HongJiang Zhang</td>
<td>396</td>
<td>18349</td>
<td>67</td>
</tr>
</tbody>
</table>

**Semantic Scholar**

<table>
<thead>
<tr>
<th>Autor</th>
<th>#paper</th>
<th>#cites</th>
<th>h-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scott Shenker</td>
<td>301</td>
<td>16287</td>
<td>65</td>
</tr>
<tr>
<td>Andrew Y. Ng</td>
<td>207</td>
<td>16971</td>
<td>57</td>
</tr>
<tr>
<td>Hector Garcia-Molina</td>
<td>455</td>
<td>11225</td>
<td>57</td>
</tr>
<tr>
<td>Ion Stoica</td>
<td>252</td>
<td>12819</td>
<td>55</td>
</tr>
<tr>
<td>Jiawei Han 0001</td>
<td>828</td>
<td>15198</td>
<td>55</td>
</tr>
<tr>
<td>Hari Balakrishnan</td>
<td>196</td>
<td>15397</td>
<td>55</td>
</tr>
<tr>
<td>Michael I. Jordan</td>
<td>415</td>
<td>18203</td>
<td>54</td>
</tr>
<tr>
<td>Yoshua Bengio</td>
<td>446</td>
<td>14935</td>
<td>52</td>
</tr>
<tr>
<td>Christos Faloutsos</td>
<td>577</td>
<td>10405</td>
<td>51</td>
</tr>
<tr>
<td>Deborah Estrin</td>
<td>243</td>
<td>11958</td>
<td>51</td>
</tr>
<tr>
<td>Sebastian Thrun</td>
<td>305</td>
<td>8001</td>
<td>50</td>
</tr>
<tr>
<td>Christopher D. Manning</td>
<td>254</td>
<td>12180</td>
<td>50</td>
</tr>
<tr>
<td>Andrew McCallum</td>
<td>207</td>
<td>10308</td>
<td>49</td>
</tr>
<tr>
<td>Daphne Koller</td>
<td>257</td>
<td>7274</td>
<td>49</td>
</tr>
<tr>
<td>Anil K. Jain</td>
<td>629</td>
<td>10733</td>
<td>49</td>
</tr>
<tr>
<td>Thomas A. Henzinger</td>
<td>449</td>
<td>9391</td>
<td>49</td>
</tr>
<tr>
<td>Alon Y. Halevy</td>
<td>253</td>
<td>8021</td>
<td>49</td>
</tr>
<tr>
<td>Jennifer Widom</td>
<td>216</td>
<td>8254</td>
<td>48</td>
</tr>
<tr>
<td>Jon M. Kleinberg</td>
<td>269</td>
<td>10906</td>
<td>47</td>
</tr>
<tr>
<td>Joseph M. Hellerstein</td>
<td>214</td>
<td>7684</td>
<td>47</td>
</tr>
</tbody>
</table>

Significantly different ranking derived from different collections – which one should one use?
Scientific Challenge:

Make bibliometric measures aware of incompleteness and possible errors

Provide confidence intervals for bibliometric measures
Possible uses of citations in dblp

• Estimate **importance of conferences** (to decide if and when a conference should be added)

  Dear DBLP team,
  I would like to ask you about the possibilities of indexation of the
  I hope to get a reply from you.

• Identify publication venues where **coverage in dblp is incomplete** (and missing part is important)

• Identify important **new publication venues**
DIY-Extraction from PDFs

- **ScienceParse** by Allen Institute for AI
- Reads (OCR‘ed) PDF as input
- Yields
  - Abstract
  - Authors with Emails
  - Full text with (some) structure
  - **Citations with (some) structure**

https://github.com/allenai/science-parse
Citations

References


"references" : [ {
  "title" : "ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints",
  "author" : [ "M. Acosta", "M. Vidal", "T. Lampo", "J. Castillo", "E. Ruckhaus" ],
  "venue" : "In ISWC’11,",
  "citeRegEx" : "1",
  "shortCiteRegEx" : "1",
  "year" : 2011
}, {
  "title" : "Describing Linked Datasets",
  "author" : [ "K. Alexander", "R. Cyganiak", "M. Hausenblas", "J. Zhao" ],
  "venue" : "In LDOW’09,",
  "citeRegEx" : "2",
  "shortCiteRegEx" : "2",
  "year" : 2009
}, ...}
Citation Contexts

ical operators. With limited access to statistics, however, most federated query engines rely on heuristics [1, 17] to reduce the huge space of possible plans or on dynamic programming (DP) [5, 7] to produce optimal plans. However, these plans may still exhibit

"referenceMentions" : [ {
   "referenceID" : 0,
   "context" : "Federated SPARQL query engines [1, 4, 7, 14, 17] answer SPARQL queries over a federation of SPARQL endpoints."
   "startOffset" : 31,
   "endOffset" : 48
 }, {
   "referenceID" : 0,
   "context" : "With limited access to statistics, however, most federated query engines rely on heuristics [1, 17] to reduce the huge space of possible plans or on dynamic programming (DP) [5, 7] to produce optimal plans."
   "startOffset" : 92,
   "endOffset" : 99
 }, ...
Evaluating Mapping Quality for Citations

96 papers from PVLDB Volume 10
• 3084 manually annotated citations
• 2700 with well-defined match in dblp

Results: (with best parameter setting, no systematic eval)
• Recall: ~80%
• Precision: ~97.5%
• Accuracy of match/nonmatch decisions: ~81%

A lot worse on old, OCR‘ed publications until ~2000 (finding citation & segmentation fails, OCR errors, ...)
## Experiment on CoRR Jan-Jun 2017

Most frequently extracted venues (after some normalization)

<table>
<thead>
<tr>
<th>venue</th>
<th>matched</th>
<th>not matched</th>
<th>overall</th>
<th>missing</th>
<th>found</th>
</tr>
</thead>
<tbody>
<tr>
<td>cvpr</td>
<td>5120</td>
<td>47</td>
<td>5167</td>
<td>0,91%</td>
<td>99,09%</td>
</tr>
<tr>
<td>advances in neural information processing systems</td>
<td>4205</td>
<td>66</td>
<td>4271</td>
<td>1,55%</td>
<td>98,45%</td>
</tr>
<tr>
<td>nips</td>
<td>2795</td>
<td>60</td>
<td>2855</td>
<td>2,10%</td>
<td>97,90%</td>
</tr>
<tr>
<td>ieee conference on computer vision and pattern recognition</td>
<td>2806</td>
<td>43</td>
<td>2849</td>
<td>1,51%</td>
<td>98,49%</td>
</tr>
<tr>
<td>corr</td>
<td>2327</td>
<td>45</td>
<td>2372</td>
<td>1,90%</td>
<td>98,10%</td>
</tr>
<tr>
<td>ieee transactions on information theory</td>
<td>2004</td>
<td>71</td>
<td>2075</td>
<td>3,42%</td>
<td>96,58%</td>
</tr>
<tr>
<td>ieee trans. inf. theory</td>
<td>2005</td>
<td>61</td>
<td>2066</td>
<td>2,95%</td>
<td>97,05%</td>
</tr>
<tr>
<td>iccv</td>
<td>1807</td>
<td>27</td>
<td>1834</td>
<td>1,47%</td>
<td>98,53%</td>
</tr>
<tr>
<td>eccv</td>
<td>1809</td>
<td>14</td>
<td>1823</td>
<td>0,77%</td>
<td>99,23%</td>
</tr>
<tr>
<td>journal of machine learning research</td>
<td>1519</td>
<td>281</td>
<td>1800</td>
<td>15,61%</td>
<td>84,39%</td>
</tr>
<tr>
<td>icml</td>
<td>1714</td>
<td>70</td>
<td>1784</td>
<td>3,92%</td>
<td>96,08%</td>
</tr>
<tr>
<td>phd thesis</td>
<td>577</td>
<td>1160</td>
<td>1737</td>
<td>66,78%</td>
<td>33,22%</td>
</tr>
<tr>
<td>ieee transactions on pattern analysis and machine intelligence</td>
<td>1553</td>
<td>69</td>
<td>1622</td>
<td>4,25%</td>
<td>95,75%</td>
</tr>
<tr>
<td></td>
<td>464</td>
<td>1146</td>
<td>1610</td>
<td>71,18%</td>
<td>28,82%</td>
</tr>
<tr>
<td>international conference on machine learning</td>
<td>1486</td>
<td>46</td>
<td>1532</td>
<td>3,00%</td>
<td>97,00%</td>
</tr>
<tr>
<td>ieee trans. wireless commun</td>
<td>1328</td>
<td>62</td>
<td>1390</td>
<td>4,46%</td>
<td>95,54%</td>
</tr>
<tr>
<td>technical report</td>
<td>235</td>
<td>1035</td>
<td>1270</td>
<td>81,50%</td>
<td>18,50%</td>
</tr>
<tr>
<td>ieee</td>
<td>868</td>
<td>350</td>
<td>1218</td>
<td>28,74%</td>
<td>71,26%</td>
</tr>
<tr>
<td>ieee trans. signal process</td>
<td>1049</td>
<td>45</td>
<td>1094</td>
<td>4,11%</td>
<td>95,89%</td>
</tr>
<tr>
<td>neural computation</td>
<td>1046</td>
<td>40</td>
<td>1086</td>
<td>3,68%</td>
<td>96,32%</td>
</tr>
<tr>
<td>ieee transactions on signal processing</td>
<td>949</td>
<td>73</td>
<td>1022</td>
<td>7,14%</td>
<td>92,86%</td>
</tr>
<tr>
<td>ieee transactions on automatic control</td>
<td>806</td>
<td>197</td>
<td>1003</td>
<td>19,64%</td>
<td>80,36%</td>
</tr>
</tbody>
</table>
### Venues with significant holes in dblp

<table>
<thead>
<tr>
<th>Venue</th>
<th>found</th>
<th>not found</th>
</tr>
</thead>
<tbody>
<tr>
<td>phd thesis</td>
<td>577</td>
<td>1160</td>
</tr>
<tr>
<td>technical report</td>
<td>464</td>
<td>1146</td>
</tr>
<tr>
<td>science</td>
<td>235</td>
<td>1035</td>
</tr>
<tr>
<td>master’s thesis</td>
<td>22</td>
<td>578</td>
</tr>
<tr>
<td>springer</td>
<td>195</td>
<td>320</td>
</tr>
<tr>
<td>journal of machine learning research</td>
<td>1519</td>
<td>281</td>
</tr>
<tr>
<td>ieee transactions on power systems</td>
<td>34</td>
<td>280</td>
</tr>
<tr>
<td>physical review letters</td>
<td>54</td>
<td>268</td>
</tr>
<tr>
<td>crc press</td>
<td>36</td>
<td>203</td>
</tr>
<tr>
<td>ieee transactions on automatic control</td>
<td>806</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>185</td>
</tr>
</tbody>
</table>
Experiment on CoRR Jan-Jun 2017

Venues that could not be matched to dblp

<table>
<thead>
<tr>
<th>venue</th>
<th>Not found</th>
</tr>
</thead>
<tbody>
<tr>
<td>the annals of mathematical statistics</td>
<td>168</td>
</tr>
<tr>
<td>psychological review</td>
<td>152</td>
</tr>
<tr>
<td>journal of the royal statistical society. series b</td>
<td>139</td>
</tr>
<tr>
<td>journal of personality and social psychology</td>
<td>96</td>
</tr>
<tr>
<td>journal of statistical software</td>
<td>85</td>
</tr>
<tr>
<td>american journal of sociology</td>
<td>69</td>
</tr>
<tr>
<td>behavior research methods</td>
<td>68</td>
</tr>
<tr>
<td>econometrica: journal of the econometric society</td>
<td>64</td>
</tr>
<tr>
<td>biglearn</td>
<td></td>
</tr>
<tr>
<td>Missing NIPS workshop (no longer available)</td>
<td></td>
</tr>
<tr>
<td>wiley online library</td>
<td>53</td>
</tr>
<tr>
<td>naval research logistics quarterly</td>
<td>49</td>
</tr>
<tr>
<td>cognitive psychology</td>
<td>46</td>
</tr>
<tr>
<td>the journal of physiology</td>
<td>45</td>
</tr>
<tr>
<td>annual review of sociology</td>
<td>45</td>
</tr>
<tr>
<td>journal of marketing research</td>
<td>44</td>
</tr>
<tr>
<td>monthly weather review</td>
<td>44</td>
</tr>
<tr>
<td>mathematische annalen</td>
<td>43</td>
</tr>
<tr>
<td>problemy peredachi informatsii</td>
<td>42</td>
</tr>
<tr>
<td>biometrics</td>
<td>40</td>
</tr>
</tbody>
</table>

Math
Sociology
Psychology
Other Sciences
Conclusion

• **Open meta data** is becoming more important and more available

• **Quality and scope** of available meta data is still unclear

• **Bibliometric measures** must take this uncertainty into account
Future Work for dblp

• Integrate with more data providers (currently ORCID and WikiData)

• Connect to bibliographic data providers from other domains

• Develop model for conference series and events

• Include references to published data (e.g., DataCite)
Future Work for Research

• Collect more extensive metadata for conferences
  – Organizers
  – Members of the program committee
  – Reviewers
  – Keynote speakers
  – ...

• Exploit this information for better estimation of the reputation of scientists (and of conferences)