Recent Advances in Bibliometric-Enhanced Information Retrieval

Philipp Mayr

ASIS&T Annual Meeting
Copenhagen, Denmark
Oct. 17, 2016
Who we are

Muthu Kumar Chandrasekaran
Kokil Jaidka
Min-Yen Kan

Ingo Frommholz
Guillaume Cabanac
Philipp Mayr
Dietmar Wolfram

- Our interests include Bibliometrics; Information Retrieval; Digital Libraries; Natural Language Processing; Text Mining, and Digital Libraries

Bibliometric-enhanced IR activities

• 2013: previous workshop at ISSI
• 2014: 1st workshop at ECIR
• 2015: special issue in Scientometrics, 2nd workshop at ECIR, CLBib workshop at ISSI
• 2016: 3rd workshop at ECIR, 4th workshop at JCDL, Panel at ASIST
• 2017: 5th workshop at ECIR, ??
Selection of BIR workshop topics

- IR and recommendation tool development and evaluation
- Bibliometric IR experiments and data sets
- Document Clustering for IR
- Citation Contexts and Analysis
IR and recommendation tool development

- Scholarly Article Recommendation Platform (Wesley-Smith et al.)
- Search Recommendations for Interactive IR (Mayr)


Bibliometric IR experiments and data sets

- Bradfordizing (Jack et al., Mayr)
- iSearch corpus (Dabrowska & Larsen, Carevic & Schaer) & PLOS corpus (Bertin & Atanassova), CL SciSumm corpus

<table>
<thead>
<tr>
<th>Collection</th>
<th>#documents</th>
<th>#topics</th>
<th>references</th>
<th>citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cranfield II</td>
<td>1,400</td>
<td>225</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>CACM</td>
<td>3,204</td>
<td>63</td>
<td>(yes)</td>
<td>yes</td>
</tr>
<tr>
<td>Fox (1983)</td>
<td>1,239</td>
<td>100</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>TREC Genomics Track</td>
<td>4.5 million (2004-2005)</td>
<td>50 (2004)</td>
<td>(Some can be extracted through PubMed?)</td>
<td>no</td>
</tr>
<tr>
<td>ACL (Kitchie 2008)</td>
<td>162,259 (2006-2007)</td>
<td>28 (2006)</td>
<td>(Some can be extracted through PubMed?)</td>
<td>no</td>
</tr>
<tr>
<td>ACL</td>
<td>9084 (full text + citation contexts)</td>
<td>82</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>iSearch (Lycke et al., 2010)</td>
<td>434,817 (PDF/LaTeX from arXiv.org)</td>
<td>65</td>
<td>12+ million</td>
<td>3.7 million</td>
</tr>
<tr>
<td>CORE .ac.uk</td>
<td>2.5 million (full text PDFs)</td>
<td>no</td>
<td>(in PDFs?)</td>
<td>no</td>
</tr>
<tr>
<td>CiteSeerX</td>
<td>6 million (full text PDFs)</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>PubMed OA</td>
<td>1.1 million (full text PDFs)</td>
<td>no</td>
<td>(in PDFs)</td>
<td>(through PubMed)</td>
</tr>
</tbody>
</table>

Fig. 1. Density of Verbs in Citation Context

Document Clustering for IR

- Polyrepresentative Clustering (Abbasi & Frommholz)
- CitNetExplorer (van Eck & Waltman)

Fig. 1. Polyrepresentation and Clustering. The left hand side shows the relevance w.r.t. combinations of different representations. The right hand side shows the induced rankings. Small circles denote relevant documents.


Citation Contexts and Analysis

- Citation contexts (Bertin & Atanassova)
- Bag of works retrieval (White, 2016)

Table 3. Top 12 titles co-cited with Bates (1989)

<table>
<thead>
<tr>
<th>TF*IDF</th>
<th>Sole or First Author, Date, and Title of Co-cited Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.61</td>
<td>ELLIS D, 1989, A behavioural approach to information retrieval design</td>
</tr>
<tr>
<td>11.22</td>
<td>BATES MJ, 1990, Where should the person stop and the information search interface start?</td>
</tr>
<tr>
<td>10.9</td>
<td>KUHLTHAU CC, 1991, Inside the search process: Information seeking from the user’s perspective</td>
</tr>
<tr>
<td>10.88</td>
<td>BELKIN NJ, 1995, Cases, scripts and information seeking strategies: Design of interactive information retrieval systems</td>
</tr>
<tr>
<td>10.84</td>
<td>MARCHIONINI G, 1995, Information Seeking in Electronic Environments</td>
</tr>
<tr>
<td>10.75</td>
<td>BELKIN NJ, 1993, BRAQUE: Design of an interface to support user interaction in information retrieval</td>
</tr>
</tbody>
</table>

Fig. 1: MIR in the 7 PLOS Journals


Overview

- .. and BIR editorials
Outlook

Special issue in IJDL to appear mid 2017

*International Journal on Digital Libraries*

Call for Papers

Special Issue on
Bibliometric-Enhanced Information Retrieval and
Natural Language Processing for Digital Libraries
Thank you

Contact:

Dr Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences, Germany
Email: philipp.mayr@gesis.org
Twitter: @philipp_mayr