Editorial for the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) at JCDL 2016

Philipp Mayr[1], Ingo Frommholz[2], Guillaume Cabanac[3], and Dietmar Wolfram[4]

[1] GESIS - Leibniz-Institute for the Social Sciences, Cologne, Germany,
philipp.mayr@gesis.org
[2] Institute for Research in Applicable Computing, University of Bedfordshire, Luton, UK,
ingo.frommholz@beds.ac.uk
[3] University of Toulouse, Computer Science Department, IRIT UMR 5505, France
guillaume.cabanac@univ-tlse3.fr
[4] School of Information Studies, University of Wisconsin-Milwaukee, USA
dwolfram@uwm.edu

1 Introduction

After the success of two parent workshops series – the 1st NLPIR4DL workshop in 2009, and the series of three Bibliometric-enhanced Information Retrieval (BIR) workshops in 2014, 2015 and 2016 – BIRNDL[5] at JCDL 2016 [1] will investigate how natural language processing, information retrieval, scientometric and recommendation techniques can advance the state-of-the-art in scholarly document understanding, analysis and retrieval at scale. Researchers are in need of assistive technologies to track developments in an area, identify the approaches used to solve a research problem over time and summarize research trends. Digital libraries require semantic search, question-answering as well as automated recommendation and reviewing systems to manage and retrieve answers from scholarly databases. Full document text analysis can help to design semantic search, translation and summarization systems; citation and social network analyses can help digital libraries to visualize scientific trends, bibliometrics and relationships and influences of works and authors. These approaches can be supplemented with the metadata supplied by digital libraries, such as usage data.

This workshop will be relevant to scholars in several fields of computer science, information science and computational linguistics; it will also be of importance for all stakeholders in the publication pipeline: implementers, publishers and policymakers – with this workshop we hope to bring a number of these contributors together. Today's publishers continue to seek new ways to be relevant to their consumers, in disseminating the right published works to their audience.

[5] http://wing.comp.nus.edu.sg/birndl-jcdl2016/

Formal citation metrics are increasingly a factor in decision-making by universities and funding bodies worldwide, making the need for research in such topics more pressing.

The BIRNDL event was split into two parts: the regular research paper track and the CL-SciSumm Shared Task system track.

2 Overview of the papers

The workshop featured one keynote talk, three paper sessions and one poster and demo interactive session. The BIRNDL organizers have accepted 5 long and 4 short papers for presentation in the research paper track. The CL-SciSumm organizers have accepted 9 system papers in the CL-SciSumm track. All papers in both tracks are included in the proceedings. The following briefly outlines the keynote and three paper sessions. The system papers in the CL-SciSumm track are outline in an overview paper [2].

2.1 Keynote

Dietmar Wolfram provided the keynote address on "Bibliometrics, Information Retrieval and Natural Language Processing: Natural Synergies to Support Digital Library Research"[3]. Until recently, methods developed for IR and bibliometrics that can be mutually beneficial have not been widely explored. This is changing as evidenced by recent themed meetings that have brought together researchers with interests that bridge both areas. Similarly, applications of language-based methods have provided new tools for research in bibliometrics and IR. The presenter discussed examples of the synergies that exist at the intersections of these three areas, not only for IR system design and evaluation, but also to provide insights into the structure of disciplines and their research communities.

2.2 Session 1

In their article "Multiple In-text Reference Phenomenon", Bertin and Atanassova studied the distribution of multiple in-text references (MIR), which are based on sentences with more than one reference [4]. A corpus of 80,000 PLOS papers was used for the analysis and references were counted based on the publications' IMRaD structure. The results revealed, for instance, that 41% of sentences with citations contain MIRs, with more than half of them in the introduction. Potential applications of this study comprised works on clustering, co-citation networks and summarization.

Citations to retracted paper were the focus of the contribution "Post Retraction Citations in Context" by Halevi and Bar-Ilan [5]. Citations to retracted articles might put the credibility of scientific work in jeopardy, hence it is a field worth studying. The authors discuss 5 case studies of retracted papers and the

negative, positive and neutral citations they received after retraction. The authors expressed their concern about the fact that retracted articles still attract citations, and provide some recommendation for publishers.

In his paper "Incorporating Satellite Documents into Co-citation Networks for Scientific Paper Searches", Masaki Eto examined the use of enlarged co-citation networks to improve IR search performance for documents from the Open Access Subset of PubMed Central [6]. Satellite documents to expand the network of linkages beyond direct co-citations were identified based on search terms appearing in documents co-cited with a seed document. Results of the study revealed that the proposed method provided better search performance than a baseline approach that did not incorporate the enlarged network.

To master the huge amount of scientific literature produced nowadays and make sense of the rich pool of knowledge they provide, Ronzano et *al.* introduced the Scientific Knowledge Miner project [7]. Based on a previous text mining project, SKM aims at extending the existing Dr. Inventor Scientific Text Mining Framework, and offers services like summarization and citation recommendation.

2.3 Session 2

Ha Jin Kim, Juyoung An, Yoo Kyung Jeong and Min Song presented the results of their research on "Exploring the Leading Authors and Journals in Major Topics by Citation Sentences and Topic Modeling" [8]. The authors employed an Author-Journal-Topic (AJT) model to identify leading journals and authors in the area of Oncology along with major topics that are shared among researchers. A key finding was that influential authors and journals identified using topic modeling did not necessarily correspond to those identified using citation-based measures. The authors concluded that the AJT model may be used to identify latent meaning in citation sentences.

Aravind Sesagiri Raamkumar, Schubert Foo, and Natalie Pang tackled a compelling question every scientist wonders while writing: "What papers should I cite from my reading list? User evaluation of a manuscript preparatory assistive task" [9]. They introduced techniques for shortlisting papers from a personal bibliography and discussed their effectiveness based on user evaluations. A panel of 116 users — balanced between students and staff members — rated the recommendations according to a variety of criteria, such as relevance, usefulness, importance, and certainty. Their positive feedback stresses the usefulness and relevance of this paper recommendation contribution.

2.4 Session 3

Jevin West and Jason Portenoy focusedon a largely ignored facet of scholarly papers – the equations [10], in their paper, "Delineating Fields Using Mathematical Jargon". They extracted mathematical symbols from Latex source files in the arXiv repository, performed an analysis of the distribution of these symbols across different fields and calculated the "jargon distance" between fields. The

main research goal of their paper was to find ways to utilize equations and formal notation in scholarly recommendation.

Joseph Mariani, Gil Francopoulo, and Patrick Paroubek discussed "A study of reuse and plagiarism in speech and natural language processing papers" [11]. They designed an algorithm based on n-gram comparisons to detect (self-)reuse and (self-)plagiarism. It was tested on the NLP4NLP dataset comprising about 65k NLP papers published during the past five decades. Results stress frequent self-plagiarism while uncommon plagiarism in the scientific literature of NLP.

Philipp Mayr presented a case study "How do practitioners, PhD students and postdocs in the social sciences assess topic-specific recommendations?" where different types of researchers in the social sciences assessed the relevance of search term, author name and journal name recommendations according to their research topics [12]. His results showed that simple bibliometric-enhanced recommendation services can be useful where they are integrated in an interactive retrieval task.

2.5 CL-SciSumm Shared Task

As part of this workshop, our colleagues at the National University of Singapore organized the CL-SciSumm Shared Task 2016[6] – a shared task on scientific paper summarization in the Computational Linguistics domain. This proceedings includes an outline of their Shared Task, as well as detailed system reports from the ten participating systems who completed the Task [2].

3 Outlook

This workshop is the first step to foster a reflection on the interdisciplinarity and the benefits that the disciplines Bibliometrics, IR and NLP can drive from it in a digital libraries context. In the future we plan follow-up workshops at IR, NLP and Digital Libraries venues. Furthermore we are working with the International Journal on Digital Libraries to offer a special issue on topics discussed at BIRNDL, for extended versions of BIRNDL workshop papers, shared task descriptions, as well as a general call for submissions.[7]

4 Acknowledgments

We are indebted to the referees who contributed to the review process: Colin Batchelor, Joeran Beel, Patrice Bellot, Marc Bertin, Guillaume Cabanac, Cornelia Caragea, Zeljko Carevic, Muthu Kumar Chandrasekaran, Jason S. Chang, Ingo Frommholz, Lee Giles, Bela Gipp, Daniel Hienert, Rahul Jha, Min-Yen Kan, Noriko Kando, Roman Kern, Claus-Peter Klas, Cyril Labbé, Birger Larsen, Elizabeth Liddy, Stasa Milojevic, Prasenjit Mitra, Marie-Francine Moens, Peter Mutschke, Doug Oard, Cécile Paris, Philipp Schaer, Andrea Scharnhorst, Henry Small, Simone Teufel, Mike Thelwall, Alex Wade, and Dietmar Wolfram.

[6] `http://wing.comp.nus.edu.sg/cl-scisumm2016/`

[7] See information at `http://wing.comp.nus.edu.sg/birndl-jcdl2016`.

References

1. Cabanac, G., Chandrasekaran, M.K., Frommholz, I., Jaidka, K., Kan, M.Y., Mayr, P., Wolfram, D.: Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016). In: JCDL '16: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, ACM New York, NY, USA (2016) 299–300
2. Jaidka, K., Chandrasekaran, M.K., Rustagi, S., Kan, M.Y.: Overview of the CL-SciSumm 2016 Shared Task. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016)
3. Wolfram, D.: Bibliometrics, Information Retrieval and Natural Language Processing: Natural Synergies to Support Digital Library Research. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 6–13
4. Bertin, M., Atanassova, I.: Multiple In-text Reference Aggregation Phenomenon. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 14–22
5. Halevi, G., Bar-Ilan, J.: Post Retraction Citations in Context. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 23–29
6. Eto, M.: Incorporating Satellite Documents into Co-citation Networks for Scientific Paper Searches. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 30–35
7. Ronzano, F., Freire, A., Saez-Trumper, D., Saggion, H.: Making Sense of Massive Amounts of Scientific Publications: the Scientific Knowledge Miner Project. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 36–41
8. Kim, H.J., An, J., Jeong, Y.K., Song, M.: Exploring the leading authors and journals in major topics by citation sentences and topic modeling. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 42–50
9. Raamkumar, A.S., Foo, S., Pang, N.: What papers should I cite from my reading list? User evaluation of a manuscript preparatory assistive task. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 51–62
10. West, J., Portenoy, J.: Delineating Fields Using Mathematical Jargon. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 63–71
11. Mariani, J., Francopoulo, G., Paroubek, P.: A study of reuse and plagiarism in speech and natural language processing papers. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 72–83
12. Mayr, P.: How do practitioners, PhD students and postdocs in the social sciences assess topic-specific recommendations? In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). (2016) 84–92