WORKSHOP REPORT

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

Guillaume Cabanac
University of Toulouse, IRIT UMR 5505, France
guillaume.cabanac@univ-tlse3.fr

Muthu Kumar Chandrasekaran
NUS School of Computing, Singapore
muthu.chandra@comp.nus.edu.sg

Ingo Frommholz
University of Bedfordshire, IRAC, Luton, UK
ingo.frommholz@beds.ac.uk

Kokil Jaidka
University of Pennsylvania, USA
jaidka@sas.upenn.edu

Min-Yen Kan
NUS School of Computing, Singapore
kanmy@comp.nus.edu.sg

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences, Germany
philipp.mayr@gesis.org

Dietmar Wolfram
University of Wisconsin-Milwaukee, USA
dwolfram@uwn.edu

Abstract

The large scale of scholarly publications poses a challenge for scholars in information seeking and sensemaking. Bibliometric, information retrieval (IR), text mining, and NLP techniques could help in these activities, but are not yet widely used in digital libraries. The BIRNDL workshop was held at the Joint Conference on Digital Libraries (JCDL 2016) in Newark, NJ. It intended to stimulate IR researchers and digital library professionals to elaborate on new approaches in natural language processing, information retrieval, scientometric, and recommendation techniques that can advance the state-of-the-art in scholarly

document understanding, analysis, and retrieval at scale. The workshop incorporated three paper sessions and the 2nd edition of the CL-SciSumm Shared Task.

1 Introduction

Current digital libraries collect and allow access to digital papers and their metadata—inclusive of citations—but mostly do not analyze the items they index. The scale of scholarly publications poses a challenge for scholars in their search for relevant literature.

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 [13]—the first joint BIRNDL workshop[1] at JCDL'16 [1] focused on scholarly publications and data [3]. The workshop investigated how natural language processing, information retrieval, as well as 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, relationships, and influences of works and authors. These approaches can be supplemented with article metadata and usage statistics collected by digital libraries.

This workshop was targeted at scholars in several fields of information science, information retrieval, and computational linguistics working in the context of digital libraries. In addition, it was intended to be of importance for all stakeholders in the publication pipeline—implementers, publishers, and policymakers—who continue to seek new ways to be relevant to their patrons, in disseminating the most relevant published works to their audience. Finally, the approaches developed herein suggested ways in which universities and funding bodies could better assess research process, quality and impact.

The workshop was split into two parts: Full and short paper presentations (Section 2) and the 2nd edition of the CL-SciSumm Shared Task (Section 3).

2 Workshop Topics

Our goal was to encourage insights from bibliometrics, scientometrics, and informetrics to applications in digital libraries. We invited submissions presenting full-text analyses, multimedia and multilingual analyses and alignment, citation-based NLP, information retrieval, information seeking, and digital libraries (DL). We further enumerated the following themes of interest:

- Summarization of scientific articles; automatic creation of reviews and automatic qualitative assessment of submissions; question-answering for scholarly DLs.
- Recommendation for scholarly papers, reviewers, citations, and publication venues.

[1] https://web.archive.org/web/2016/http://wing.comp.nus.edu.sg/birndl-jcdl2016/

- Navigation, searching, and browsing in scholarly DLs; niche search in scholarly DLs; new information access methods for scientific papers.
- Network analysis and citation analysis in scholarly DLs; citation function/motivation analysis; novel bibliometric metrics; topical modeling analysis; information retrieval for scholarly text, e.g., citation-based IR.
- Knowledge discovery and analysis of the ancestry of ideas.
- Translation, multilingual and multimedia analysis and alignment of scholarly works; analyses of writing style in scholarly publications.
- Metadata and controlled vocabularies for resource description and discovery; automatic metadata discovery, such as language identification.
- Disambiguation issues in scholarly DLs using NLP or IR techniques; data cleaning and data quality.

The workshop started with an inspirational keynote by Dietmar Wolfram (University of Wisconsin–Milwaukee) followed by paper presentations: 5 long and 4 short papers. A special session with poster presentations of the 9 participating groups in the CL-SciSumm Shared Task. A fishbowl-style panel concluded the workshop.

2.1 Keynote

Dietmar Wolfram delivered a keynote address on "Bibliometrics, Information Retrieval and Natural Language Processing: Natural Synergies to Support Digital Library Research"[18]. 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 bridging both areas. Similarly, applications of language-based methods have provided new tools for research in bibliometrics and IR. Wolfram 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 titled "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 [2]. 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 paper titled "Post Retraction Citations in Context" by Halevi and Bar-Ilan [7]. Citations to retracted articles might put the credibility of scientific work in jeopardy, hence it is a field worth studying. The authors discussed five 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,” 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 [15]. 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

Kim et *al.* presented the results of their paper titled “Exploring the Leading Authors and Journals in Major Topics by Citation Sentences and Topic Modeling” [9]. 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.

Raamkumar et *al.* addressed a question faced by every scholar: “What papers should I cite from my reading list? User evaluation of a manuscript preparatory assistive task” [14]. The authors 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

West and Portenoy focused on a largely ignored facet of scholarly papers—the equations—in their paper, “Delineating Fields Using Mathematical Jargon” [17]. 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.

Mariani et *al.* presented their paper titled “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.

Mayr presented a case study—“How do practitioners, PhD students and postdocs in the social sciences assess topic-specific recommendations?” [12]. In this work, 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. His results showed that simple bibliometric-enhanced recommendation services can be useful when integrated in an interactive retrieval task.

3 The 2nd CL-SciSumm Shared Task

As a part of the workshop, we conducted the 2nd Computational Linguistics Scientific Summarization Shared Task, sponsored by Microsoft Research Asia. This is the first medium-scale shared task on scientific document summarization in the computational linguistics (CL) domain. Fifteen teams from six countries registered for the Shared Task, of which ten teams ultimately submitted and presented their results. The task was run on an annotated corpus of 30 target papers—currently the largest of such available corpora. The corpus is made available for free download and use at `https://github.com/WING-NUS/scisumm-corpus`.

In this task, participants were provided a training corpus of 30 topics, each comprising a Reference Paper (RP) and 10 or more Citing Papers (CPs) that all contain citations to the RP. In each CP, the text spans (i.e., citances) that pertain to a particular citation to the RP have been identified. Participants were required to solve three sub-tasks in automatic research paper summarization on a text corpus. The development corpus was an extended version of the dataset used at the CL Pilot Task at the Text Analysis Conference 2014 (TAC 2014).

Ten participating systems proposed a variety of heuristical, lexical, and supervised approaches. The final leader board of results (see Figure 1) shows that hybrid supervised approaches outperformed all other systems in Task 1 [10]. In general, those systems which implemented weights based on term- and document-frequency tended to perform better than those which did not [4, 16]. The results from Task 2 suggest that automatic summarization systems for scientific literature, may work across disciplines—this is because the best-performing system in the summary generation task had originally been developed for

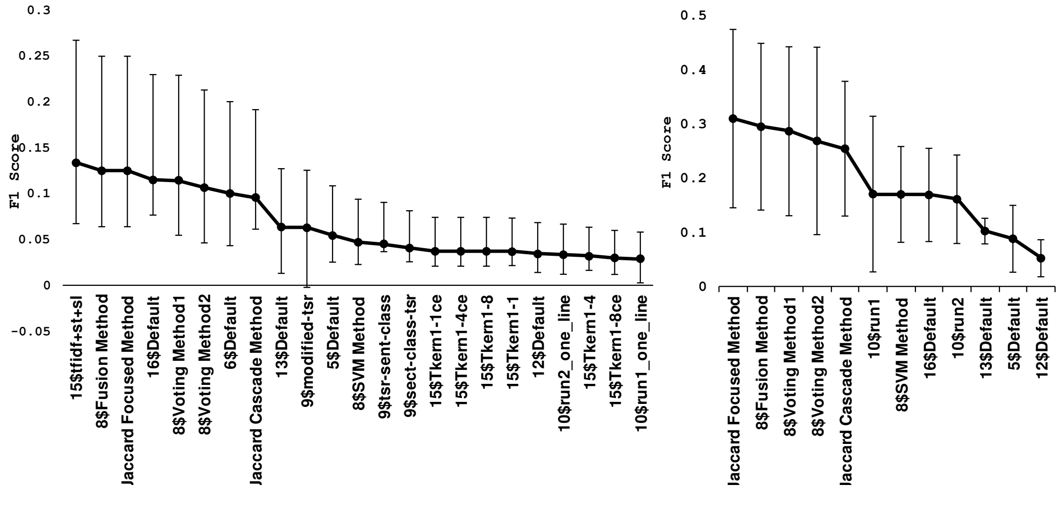

Figure 1: CL-SciSumm 2016 shared task leader board for Task1a (left) and Task1b (right). Detailed results can be found in the overview paper [8] or the slidedeck (`http://bit.ly/cl-scisumm16-slides`)

generating biomedical human summaries [5]. Based on the interest of the community and the participants' feedback, we believe that the CL-SciSumm Shared Task and the related corpus has broad applicability to related problems in computational linguistics and natural language processing, especially in the sub-disciplines of text summarization, discourse structure in scholarly discourse, paraphrase, textual entailment, and text simplification.

Figure 2: BIRNDL 2016 audience (`http://twitter.com/knmnyn/status/746004228523565057`)

4 Outlook

This workshop was the first step to foster the reflection on the interdisciplinarity and the benefits that the disciplines Bibliometrics, IR, and NLP can drive from it in a digital libraries context. Based on feedback from the audience, in the future we plan to host follow-up BIRNDL workshops co-located with key IR, NLP, and Digital Libraries conferences, as well as to partner with other related workshops being conducted at these venues, in order to consolidate the community. We are also planning the next edition of the CL-SciSumm shared task with a bigger corpus, as part of a proposed 3rd Natural Language Processing and Information Retrieval for Digital Libraries (NLPIR4DL) Workshop.

Furthermore, as follow-up of this workshop, the *International Journal on Digital Libraries*[2] will publish a special issue in mid 2017 on "Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries."

5 Acknowledgments

We thank Microsoft Research Asia for its generous support regarding the development, annotation, and dissemination of the dataset as well as the organization of the CL-SciSumm Shared Task.

[2] `http://bit.ly/ijdl-birndl`

References

[1] Nabil R. Adam, Lillian Cassel, and Yelena Yesha, editors. *JCDL'16: Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, 2016. ACM.

[2] Marc Bertin and Iana Atanassova. 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)*, pages 14–22, 2016.

[3] Guillaume Cabanac, Muthu Kumar Chandrasekaran, Ingo Frommholz, Kokil Jaidka, Min-Yen Kan, Philipp Mayr, and Dietmar Wolfram, editors. *BIRNDL'16: Proceedings of the 1st Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries co-located with the 16th ACM/IEEE Joint Conference on Digital Libraries (JCDL'16)*, volume 1610, RWTH Aachen University, 2016. CEUR Workshop Proceedings.

[4] Ziqiang Cao, Wenjie Li, and Dapeng Wu. Polyu at cl-scisumm 2016. In *Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016)*, pages 132–138, Newark, NJ, USA, June 2016.

[5] John Conroy and Sashka Davis. Vector space and language models for scientific document summarization. In *NAACL-HLT*, pages 186–191, Newark, NJ, USA, 2015. Association of Computational Linguistics.

[6] Masaki Eto. 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)*, pages 30–35, 2016.

[7] Gali Halevi and Judit Bar-Ilan. Post Retraction Citations in Context. In *Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016)*, pages 23–29, 2016.

[8] Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 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)*, pages 93–102, 2016.

[9] Ha Jin Kim, Juyoung An, Yoo Kyung Jeong, and Min Song. 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)*, pages 42–50, 2016.

[10] Lei Li, Liyuan Mao, Yazhao Zhang, Junqi Chi, Taiwen Huang, Xiaoyue Cong, and Heng Peng. CIST System for CL-SciSumm 2016 Shared Task. In *Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016)*, pages 156–167, 2016.

[11] Joseph Mariani, Gil Francopoulo, and Patrick Paroubek. 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)*, pages 72–83, 2016.

[12] Philipp Mayr. 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)*, pages 84–92, 2016.

[13] Philipp Mayr, Ingo Frommholz, and Guillaume Cabanac. Report on the 3rd International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016). *SIGIR Forum*, 50(1):28–34, 2016.

[14] Aravind Sesagiri Raamkumar, Schubert Foo, and Natalie Pang. 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)*, pages 51–62, 2016.

[15] Francesco Ronzano, Ana Freire, Diego Saez-Trumper, and Horacio Saggion. 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)*, pages 36–41, 2016.

[16] Horacio Saggion, Ahmed AbuRa'Ed, and Francesco Ronzano. Trainable Citation-enhanced Summarization of Scientific Articles. In *Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016)*, pages 175–186, 2016.

[17] Jevin West and Jason Portenoy. Delineating Fields Using Mathematical Jargon. In *Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016)*, pages 63–71, 2016.

[18] Dietmar Wolfram. 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)*, pages 6–13, 2016.