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

Guillaume Cabanac[1], Muthu Kumar Chandrasekaran[2], Ingo Frommholz[3],
Kokil Jaidka[4], Min-Yen Kan[2], Philipp Mayr[5], Dietmar Wolfram[6]
[1]University of Toulouse, France; [2]NUS School of Computing, Singapore;
[3]University of Bedfordshire, UK; [4]Adobe Systems Inc., India;
[5]GESIS - Leibniz Institute for the Social Sciences, Germany; [6]Univ. of Wisconsin-Milwaukee, USA
[1]guillaume.cabanac@univ-tlse3.fr; [2]{muthu.chandra,kanmy}@comp.nus.edu.sg;
[3]ingo.frommholz@beds.ac.uk; [4]jaidka@adobe.com; [5]philipp.mayr@gesis.org; [6]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. This workshop is intended to stimulate IR researchers and digital library professionals to elaborate on new approaches in natural language processing, information retrieval, scientometric and recommendation techniques which can advance the state-of-the-art in scholarly document understanding, analysis and retrieval at scale.

CCS Concepts

•**Information systems → Information retrieval;** *Link and co-citation analysis;* •**Applied computing → Digital libraries and archives;**

Keywords

Bibliometrics; Information Retrieval; Digital Libraries; Natural Language Processing; Text Mining

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 – BIRNDL[1] will focus on scholarly publications and data. The workshop 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 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. 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 event is split into two parts: the paper presentations and the CL-SciSumm Shared Task.

2. WORKSHOP TOPICS AND FORMAT

Our goal is to encourage insights from bibliometrics, scientometrics and informetrics to applications in digital libraries. We invite stimulating submissions on topics including – but not limited to – full-text analysis, multimedia and multilingual analysis and alignment as well as citation-based NLP, information retrieval, information seeking and digital libraries (DL). Specific examples of fields of interests include (but are not limited to):

- Summarization of scientific articles; automatic creation of reviews and automatic qualitative assessment of submissions; question-answering for scholarly DLs

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

JCDL '16 June 19-23, 2016, Newark, NJ, USA
© 2016 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4229-2/16/06.
DOI: http://dx.doi.org/10.1145/2910896.2926734

- Recommendation for scholarly papers, reviewers, citations and publication venues
- 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 will start with an inspirational keynote by Dietmar Wolfram (University of Wisconsin–Milwaukee) followed by paper presentations. There will be a special session with presentations of the participating groups in the CL-SciSumm Shared Task as well as a planned fishbowl-style panel.

3. THE CL-SCISUMM SHARED TASK

To highlight and bring together the scholars working on above topics, BIRNDL will also host a shared task on scientific document summarization on open access literature in the computational linguistics (CL) domain. The output summaries will be of two types: faceted summaries of the traditional self-summary (the abstract) and the community summary (the collection of citation sentences or *citances*) [4]. This task follows up on the successful CL Pilot Task conducted as a part of the BiomedSumm Track at the Text Analysis Conference 2014 (TAC 2014), and re-uses the SciSumm14 manually-annotated dataset [2], to enhance impact and visibility. In this shared task, we will extend the SciSumm14 dataset of ten, by releasing pairs of training and test datasets: each pair comprising the annotated citing sentences for a research paper, and summaries of the research paper. The resulting CLSciSumm16 corpus is expected to be of interest to a broad community including those working 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. Microsoft Research Asia is generously supporting the development, annotation and dissemination of the dataset as well as the organization of this shared task.

4. PREVIOUS RELATED WORKSHOPS

Our workshop is a continuation of several previous ones on similar topics. We present a summary of some relevant recent events, which underpin our claim of the workshop topic being spot-on and relevant.

- The 1st Workshop on text and citation analysis for scholarly digital libraries (NLPIR4DL) was held in conjunction with ACL-IJCNLP 2009, Singapore.
- Scholarly Big Data: AI Perspectives, Challenges, and Ideas at AAAI 2016 - This workshop is related to our topics but appears earlier in 2016. It indicates a high degree of interest for our topic, and will be synergistic due to its complementary date.
- 3rd Workshop on Argumentation Mining at ACL 2016 - This related workshop is synergistic and complementary. We overlap to a small extent in being interested in argumentation (their workshop) in scientific documents (our workshop).
- 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR2016) at ECIR 2016 [3]. The scope of the BIR workshops (2014, 2015 and 2016) were on information retrieval, information seeking, science modelling, network analysis, and digital libraries, applying insights from bibliometrics, scientometrics, and informetrics.
- 1st Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics at ISSI 2015 [1] brought together researchers to study the ways Bibliometrics can benefit from large-scale text analytics and sense mining of scientific papers, thus exploring the interdisciplinarity of Bibliometrics and NLP.

5. OUTLOOK

This workshop is 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. 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. Dates for first submission of camera-ready papers will likely be around September 2016, with a target of producing an issue by mid 2017.

6. REFERENCES

[1] Iana Atanassova, Marc Bertin, and Philipp Mayr. Proceedings of the First Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics. Istanbul, Turkey, 2015.

[2] Kokil Jaidka, Muthu Kumar Chandrasekaran, Beatriz Fisas Elizalde, Rahul Jha, Christopher Jones, Min-Yen Kan, Ankur Khanna, Diego Molla-Aliod, Dragomir R Radev, Francesco Ronzano, et al. The computational linguistics summarization pilot task. In *Proceedings of Text Ananlysis Conference*, Gaithersburg, USA, 2014.

[3] Philipp Mayr, Ingo Frommholz, and Guillaume Cabanac. Proceedings of the Third Workshop on Bibliometric-enhanced Information Retrieval. Padova, Italy, 2016.

[4] Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*, pages 81–88, 2004.