

Introduction to the special issue on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)

Philipp Mayr¹  · Ingo Frommholz²  · Guillaume Cabanac³  ·
Muthu Kumar Chandrasekaran⁴  · Kokil Jaidka⁵  · Min-Yen Kan⁴  ·
Dietmar Wolfram⁶

© Springer-Verlag GmbH Germany 2017

Abstract The large scale of scholarly publications poses a challenge for scholars in information seeking and sensemaking. Bibliometric, information retrieval (IR), text mining, and natural language processing techniques can assist to address this challenge, but have yet to be widely used in digital libraries (DL). This special issue on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL) was compiled after the first joint BIRNDL workshop that was held at the joint conference on digital libraries (JCDL 2016) in Newark, New Jersey,

USA. It brought together IR and DL researchers and 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. This special issue includes 14 papers: four extended papers originating from the first BIRNDL workshop 2016 and the BIR workshop at ECIR 2016, four extended system reports of the CL-SciSumm Shared Task 2016 and six original research papers submitted via the open call for papers.

✉ Philipp Mayr
philipp.mayr-schlegel@gesis.org
Ingo Frommholz
ifrommholz@acm.org
Guillaume Cabanac
guillaume.cabanac@univ-tlse3.fr
Muthu Kumar Chandrasekaran
muthu.chandra@comp.nus.edu.sg
Kokil Jaidka
jaidka@sas.upenn.edu
Min-Yen Kan
kanmy@comp.nus.edu.sg
Dietmar Wolfram
dwolfram@uwm.edu

Keywords Computational linguistics · Scientometrics · Scientific document summarization · Shared task · Information seeking · Academic search

1 Introduction

After the success of two parent workshops series—the 1st NLP4DL workshop in 2009, and the series of three Bibliometric-enhanced Information Retrieval (BIR) workshops in 2014, 2015 and 2016—BIRNDL¹ papers presented at JCDL 2016 [3,5] investigated 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.

Digital libraries present unique challenges for representation and retrieval that make them an ideal environment to investigate applications of bibliometric and natural language processing methods to support discovery and retrieval. DL users need more effective search and discovery tools that go beyond simple keyword access to increasingly large, hetero-

¹ GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
² Institute for Research in Applicable Computing, University of Bedfordshire, Luton, UK
³ Computer Science Department, IRIT UMR 5505 CNRS, University of Toulouse, Toulouse, France
⁴ NUS School of Computing, Singapore, Singapore
⁵ University of Pennsylvania, Philadelphia, PA, USA
⁶ University of Wisconsin-Milwaukee, Milwaukee, WI, USA

¹ <http://wing.comp.nus.edu.sg/birndl-jcdl2016/>.

geneous digital collections that may consist of more formal, structured documents and unstructured text and other media. Language-based models to support IR are equally applicable in digital library environments and can support text-based bibliometric analysis [19]. Similarly, hyperlink access to structured and unstructured multimedia documents provides additional avenues for discovery that parallel citation concepts studied in bibliometrics and to which citation and social network analysis methods may be applied.

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 special issue will be relevant to scholars in several fields, including computer science, information science and computational linguistics; it will also be of importance for all stakeholders in the publication pipeline: implementers, publishers and policymakers. 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.

2 Special issue papers

The BIRNDL event at JCDL 2016 was split into two parts: the regular research paper track and the CL-SciSumm Shared Task system track.²

This special issue resulting from the BIRNDL workshop includes 14 papers: four extended papers presented at the first BIRNDL workshop and the BIR workshop at ECIR 2016 [2, 8, 14, 18], three extended system reports of the CL-SciSumm Shared Task 2016 [1, 13, 16] and one overview paper [11] and six original research papers submitted via the open call for papers [6, 7, 9, 10, 12, 17].

2.1 BIRNDL and BIR workshop papers

- Authors Mariani, Francopoulo and Paroubek investigate “Reuse and plagiarism in Speech and Natural Lan-

guage Processing publications” [14] to detect extrinsic instances of self-reuse, self-plagiarism, reuse and plagiarism in NLP and speech processing articles. By comparing word sequences in publications from the NLP4NLP corpus, consisting of more than 65,000 documents, the authors found that self-reuse (i.e., the reuse of text from another document on which there is a common author and that is cited) is relatively common, but reuse of content from others papers that have been cited and plagiarism, where there is no attribution, were quite rare and remained within ethical limits.

- In their article “The context of multiple in-text references and their signification” [2], Marc Bertin and Iana Atanassova studied the distribution of multiple in-text references (MIR), which are based on sentences with more than one reference. A corpus of 80,000 PLOS papers was used for the analysis. References were counted based on the publications IMRaD structure and their linguistic contexts, based on POS-tagging. 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 include the clustering and summarization of research papers based on co-citation networks.
- In “Bag of works retrieval: TF*IDF weighting of works co-cited with a seed” [18], Howard D. White introduces a concept of a novel TF*IDF weighting adaptation. He combines co-citation linkages of documents and TF*IDF weighting of terms. “Bag of Works” can be operationalized with the user entering a string identifying a work (“a seed document”) to retrieve the strings identifying other works that are co-cited with the seed. The author describe how the two counts are plugged into the standard formula for TF*IDF weighting such that all the co-cited items can be ranked for relevance to the seed, given that the entire retrieval is relevant to it by evidence from multiple co-citing authors. The “Bag of Works” approach is not implemented yet but White illustrates some properties of the “Bag of Works”-specific ranking by works co-cited with three well-known seeds (well-cited papers). In the end, possible users and uses are discussed.
- In “The references of references: a method to enrich humanities library catalogs with citation data” [8], Giovanni Colavizza, Matteo Romanello, and Frederic Kaplan enhanced a digital library by collecting references from domain-specific reference monographs in the Humanities. They present a workflow that identifies the core works from a domain based on a set of reference monographs. Their experiment on a corpus dedicated to the history of Venice stresses the necessity of including such overlooked references to improve search effectiveness in such corpora.

² The BIRNDL 2016 proceedings including research papers and system papers are available at <http://ceur-ws.org/Vol-1610/> [4].

2.2 CL-SciSumm shared task system track

This section describes the systems which participated in the CL-SciSumm Shared Task. Two of the papers in the open call, described in the following section, have also used the CL-SciSumm dataset released as a part of this task for their experiments.

- The first article in this track, Jaidka et al.’s “Insights from CL-SciSumm 2016: the Faceted Scientific Document Summarization Shared Task” [11] by the shared task coordinators, provides an overview of the participating systems and the official results of the task. Aside from the summary of system results, the key novel content of this paper is the participant feedback, which highlight the data quality aspects and amendments to future editions of the task. A meta-analysis revealed that the performance of kernel approaches had the largest variances and merit further experiments and adjustment of parameters. An analysis of the test set also suggested that the summarization task is of variable difficulty. An ideal system should be able to generalize well on different instances of the scientific summarization task.
- Li et al.’s “Computational linguistics literature and citations oriented citation linkage, classification and summarization” [13] provides the approach of one of the top-ranked systems in the shared task. The authors used knowledge-rich lexica and rules combined by the use of a support vector machine with voting and hierarchical topic modeling for summary generation. This implementation also expanded the corpus vocabulary by adding WordNet hypernyms and hyponyms for every word in the training corpus. This paper also built one set of topic models each, for every set of reference papers and citing papers, and found the best results when the topic count was set to 30.
- Moraes et al.’s “Identifying reference spans: topic modeling and word embeddings help IR” [16], describes another top-ranked system and its approach in the CL-SciSumm Shared Task. Using a WordNet expansion, they address the fallibility of tf-idf methods which cannot map citances to reference sentences when there are no common words among the two. Furthermore, they demonstrate how topic models can better represent the limited vocabulary of citances and lead to better performances in the citation-mapping task.
- Al Saied et al.’s “Automatic summarization of scientific publications using a feature selection approach” [1] explores the flexibility and relevance of an optimized, language-agnostic feature-based approach, evaluated both on the DUC ACQUAINT corpus and the SciSumm corpus. Its strongest advantage is that it is not dependent on the availability of a training set. Their

experiments show that this approach performed at par with the best-performing systems in terms of recall and placed fifth in terms of precision.

2.3 Open call for papers

- In their paper “Task-oriented search for evidence-based medicine” [12], the authors investigate information retrieval and natural language processing methods for accessing literature in digital libraries specifically for clinical decision support, which is crucial for evidence-based medicine. Main clinical tasks comprise searching for diagnoses, for tests and for treatments, respectively. An empirical evaluation using the TREC Clinical Decision Support Challenges showed that retrieval effectiveness can be improved by considering above clinical tasks. Task-oriented filtering also turns out to be cost-saving, where cost is defined as the number of articles users have to view before reaching a certain gain.
- In “Investigating exploratory search activities based on the stratagem level in digital libraries” [6], Carevic et al. present a user study on exploratory search for a given search task in the social science digital library sowiport. The users, 32 participants in total (16 postdocs and 16 students), were asked to perform a 10 minutes exploratory search task “finding and collecting similar documents” starting from a predefined relevant seed document. The participants were observed via an eyetracker and their gaze data recorded. The authors use a novel tree graph representation to visualize the users’ search patterns and introduce a way to combine multiple search session trees. The results show that search activities on the stratagem level are frequently utilized by both user groups. The most heavily used search activities were keyword search, followed by browsing through references and citations, and author searching. The authors found a tendency of the postdoctoral researchers to examine the metadata records more intensively with regard to dwell time and the number of fixations. The authors were able to identify common patterns like economic (explorative) and exhaustive (navigational) behavior in the search session trees of the users.
- The role of linked open data (LOD) for bibliographic recommendation is investigated in the article “Retrieval by recommendation: using LOD technologies to improve digital library search” [17]. A Web-based experiment is conducted to test the feasibility of LOD-based recommendations in the digital library context. While the LOD-based recommendation did not outperform a text-based one, the authors emphasize some advantages, such as the fact that LOD recommendations can be calculated without the need of further preprocessing. The authors

also propose two new content-based recommendation approaches: flexible similarity detection (suitable for browsing and exploratory search) and constraint-based recommendations (to filter recommendation results). Their findings show that the linked open data approach to recommendation come closes to rivaling common, text-based recommendation systems based on open-source information retrieval libraries (in their case, Apache Solr). Such LOD-based approaches hold the promise of bettering text-based systems in the future when linked and open resources are improved.

- The paper titled “Extracting discourse elements and annotating scientific documents using the SciAnnotDoc model: a use case in gender documents” [10] presents an ontology of discourse types in an annotated corpus of gender studies. The motivation behind this study is much aligned with the rest of this proceedings as it seeks to explore the complex tasks involved in the writing of a research paper or its literature review, such as finding all the definitions of a term, finding all the results that support a claim and so on. The paper is based on real cases from interviews with social scientists, and implemented as an annotation model of metadata, textual content, discourse elements and relational elements. The authors report encouraging findings from their user study and suggest that an interface based on their ontology retrieves more useful results than a general keyword search for a task-based evaluation.
- The paper “Scientific document summarization via citation contextualization and scientific discourse” [7] addresses the challenge in verifying the accuracy of claims made in citations, which in themselves are often too short to include the evidence and context from the referring document. They propose three methods—query reformulation, word embeddings and supervised learning—to contextualize citations to their source in the referring document. Their experiments are conducted on the TAC 2014 scientific summarization dataset and the CL-SciSumm dataset released as a part of the aforementioned Shared Task.
- The paper titled “Section mixture models for scientific document summarization” [9] describes a mixture model approach followed on the CL-SciSumm dataset for scientific summarization. A bigram mixture model was trained on the main sections of the scientific document and its citing sentences to estimate the weight of the terms, and sentences were selected for the final summary based on a combinatorial optimization approach. This paper exploited the well-defined structure of scientific articles and pre-trained models on the TAC 2014 corpus for the CL-SciSumm Shared Task. Surprisingly, the models transferred over to the new domain with little effort,

and produced the highest scores when evaluated against human summaries.

3 Future work

In August 2017, we organized the 2nd BIRNDL event³ at SIGIR 2017, where we had over 40 participants for a half-day workshop. Motivated by the enthused response, we are encouraged to continue the tradition of organizing an annual workshop, together with a Shared Task and the release of new annotated corpora. We look forward to engaging with the bibliometrics, IR and NLP community on the emerging challenges in digital libraries.

Acknowledgements We wish to thank all those who have contributed to the special issue: all those who contributed papers, the many reviewers who generously gave their time, the various people involved in publishing the issue and the participants of BIRNDL workshop in 2016. We hope the articles in the issue will provide a starting point for future explorations in the field.

References

1. Al Saied, H., Dugué, N., Lamirel, J.C.: Automatic summarization of scientific publications using a feature selection approach. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0214-x>
2. Bertin, M., Atanassova, I.: The context of multiple in-text references and their signification. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0225-7>
3. 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, pp. 299–300*. <https://doi.org/10.1145/2910896.2926734> (2016)
4. Cabanac, G., Chandrasekaran, M.K., Frommholz, I., Jaidka, K., Kan, M.Y., Mayr, P., Wolfram, D. (eds.): *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) Colocated with the Joint Conference on Digital Libraries 2016 (JCDL 2016), CEUR Workshop Proceedings, vol. 1610*. CEUR-WS.org. <http://ceur-ws.org/Vol-1610/> (2016)
5. Cabanac, G., Chandrasekaran, M.K., Frommholz, I., Jaidka, K., Kan, M.Y., Mayr, P., Wolfram, D.: Report on the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016). *SIGIR Forum* **50**(2), 36–43. <http://sigir.org/wp-content/uploads/2017/01/p036.pdf> (2016)
6. Carevic, Z., Lusky, M., van Hoek, W., Mayr, P.: Investigating exploratory search activities based on the stratagem level in digital libraries. *Int. J. Digit. Libr.* <https://doi.org/10.1007/s00799-017-0226-6>. <https://arxiv.org/abs/1706.06410> (2017)
7. Cohan, A., Goharian, N.: Scientific document summarization via citation contextualization and scientific discourse. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0216-8>

³ <http://wing.comp.nus.edu.sg/birndl-sigir2017/> and workshop proceedings at <http://ceur-ws.org/Vol-1888/> [15].

8. Colavizza, G., Romanello, M., Kaplan, F.: The references of references: a method to enrich humanities library catalogs with citation data. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0210-1>
9. Conroy, J.M., Davis, S.T.: Section mixture models for scientific document summarization. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0218-6>
10. de Ribaupierre, H., Falquet, G.: Extracting discourse elements and annotating scientific documents using the SciAnnotDoc model: a use case in gender documents. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0227-5>
11. Jaidka, K., Chandrasekaran, M.K., Rustagi, S., Kan, M.Y.: Insights from CL-SciSumm 2016: the faceted scientific document summarization Shared Task. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0221-y>
12. Koopman, B., Russell, J., Zuccon, G.: Task-oriented search for evidence-based medicine. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0209-7>
13. Li, L., Mao, L., Zhang, Y., Chi, J., Huang, T., Cong, X., Peng, H.: Computational linguistics literature and citations oriented citation linkage, classification and summarization. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0219-5>
14. Mariani, J., Francopoulo, G., Paroubek, P.: Reuse and plagiarism in Speech and Natural Language Processing publications. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0211-0>
15. Mayr, P., Chandrasekaran, M.K., Jaidka, K. (eds.): Proceedings of the 2nd Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017) Co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), Tokyo, Japan, August 11, 2017, CEUR Workshop Proceedings, vol. 1888. CEUR-WS.org. <http://ceur-ws.org/Vol-1888> (2017)
16. Moraes, L., Baki, S., Verma, R., Lee, D.: Identifying reference spans: topic modeling and word embeddings help IR. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0220-z>
17. Wenige, L., Ruhland, J.: Retrieval by recommendation: using LOD technologies to improve digital library search. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0224-8>
18. White, H.D.: Bag of works retrieval: TF*IDF weighting of works co-cited with a seed. *Int. J. Digit. Libr.* (2017). <https://doi.org/10.1007/s00799-017-0217-7>
19. Wolfram, D.: Bibliometrics, information retrieval and natural language processing: natural synergies to support digital library research. In: Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) Co-located with the Joint Conference on Digital Libraries 2016 (JCDL 2016), Newark, NJ, USA, June 23, 2016, pp. 6–13. <http://ceur-ws.org/Vol-1610/paper1.pdf> (2016)