Report on the 5th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)

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

Ingo Frommholz
Institute for Research in Applicable Computing
University of Bedfordshire, Luton, UK
ifrommholz@acm.org

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

Abstract

This workshop report presents the output of the fifth Bibliometric-enhanced Information Retrieval (BIR) workshop, which has been co-located with the 39th European Conference on Information Retrieval (ECIR 2017) in Aberdeen, UK. We motivate our workshop and outline the papers (one keynote, six regular papers and five poster papers) presented at BIR 2017. Finally, we conclude with an outlook and future directions of this workshop activity.

1 Introduction

Following the successful workshops at ECIR 2014[1], 2015[2], 2016[3] [8] and JCDL 2016[4], respectively, this workshop was the fifth in a series of events that brought together experts of communities which often have been perceived as different ones: bibliometrics / scientometrics / informetrics on the one hand and information retrieval on the other. Our motivation as organizers of the workshop started from the observation that main discourses in both fields

[1] `http://ceur-ws.org/Vol-1143/`

[2] `http://ceur-ws.org/Vol-1344/`

[3] `http://ceur-ws.org/Vol-1567/`

[4] `http://ceur-ws.org/Vol-1610/`

are different, that communities, though both embedded in information science, are only partly overlapping [19] and from the belief that a knowledge transfer would be profitable for both sides [10].

This fifth full-day Bibliometric-enhanced Information Retrieval (BIR) workshop[5] at ECIR 2017 aimed to foster a common ground for the incorporation of bibliometric-enhanced services into scholarly search engine interfaces. In particular we addressed specific communities, as well as studies on large, cross-domain collections like Web of Science, Scopus or Mendeley. This fifth BIR workshop addressed explicitly both scholarly and industrial researchers.

2 Workshop Theme and Topics

Researchers from different domains, such as information retrieval, information seeking, science modelling, bibliometrics, scientometrics, network analysis, and digital libraries were invited to contribute to the workshop and to move toward a deeper understanding of related research challenge. To support the previously described goals the workshop topics included (but were not limited to) the following:

- IR for digital libraries and academic search engines
- IR for scientific domains, e.g. social sciences, life sciences, etc.
- Bibliometrics, citation analysis, and network analysis for IR
- Query expansion and relevance feedback approaches
- Science Modelling (both formal and empirical)
- Task based user modelling, interaction, and personalisation
- (Long-term) Evaluation methods and test collection design
- Collaborative information handling and information sharing
- Classification, categorisation, and clustering approaches
- Information extraction (including topic detection, entity and relation extraction)
- Recommendations based on explicit and implicit user feedback
- Recommendation for scholarly papers, reviewers, citations and publication venues
- (Social) Book Search
- Information extraction (including topic detection, entity and relation extraction)

3 Paper Presentations

This year 16 papers were submitted to the workshop, 11 of which were finally accepted for presentation and inclusion in the proceedings: 6 regular papers and 5 posters. The workshop featured one keynote talk, three full paper sessions and one poster session. The workshop proceedings have been published with CEUR Workshop Proceedings, see Vol-1823[6] [9]. The following section briefly describes the keynote and paper sessions.

[5] http://www.gesis.org/en/services/events/events-archive/conferences/ecir-workshops/ecir-workshop-2017/

[6] http://ceur-ws.org/Vol-1823/

3.1 Keynote

The invited paper "Real-World Recommender Systems for Academia: The Pain and Gain in Building, Operating, and Researching them" [2] by Joeran Beel (Trinity College Dublin, Ireland) gives an insightful overview of the practical experiences in building scholarly document recommender systems for Digital Libraries. The authors Beel and Dinesh report about their research with three different recommender systems which have been implemented and operated in the last six years. They present empirical results of various studies, discuss challenges like running A/B testing with real-world scholarly recommender systems and perform research against competitive benchmarks.

3.2 Session 1: Full papers

In the paper "Manuscript Matcher: A Content and Bibliometrics-based Scholarly Journal Recommendation System" [13], Jason Rollins, Meredith McCusker, Joel Carlson and Jon Stroll present a scholarly journal recommendation system called Manuscript Matcher which is developed and run by Clarivate (formerly Thomson Reuters). The use case of the tool is uploading manuscript title, abstract and references to Manuscript Matcher and getting back bibliometric-informed recommendations of journals ("best fit" publications). The authors present user feedback of the recommendation system and future directions.

In their paper "Use of Locality Sensitive Hashing (LSH) Algorithm to Match Web of Science and SCOPUS" [1], Mehmet Ali Abdulhayoglu and Bart Thijs report on an attempt to match the records of the two flagship bibliographic databases Web of Science and SCOPUS. They considered various metadata (e.g., publication title, venue name, bylines) whilst disregarding identifiers such as DOIs, as these are not always available or assigned. Their efficient approach based on LSH found a 70% intersection between these in about an hour. This research contributes to the understanding of the coverage of leading bibliographic databases.

3.3 Session 2: Full papers

The paper "Academic Search in Response to Major Scientific Events" [7] by Li and de Rijke describes search behaviour of academic and web users in occurrence of major scientific events (the Nobel Prize announcements of Chemistry, Physics and Medicine in 2014). The authors compare the query patterns in the query log of the academic search engine ScienceDirect with the data provided by Google Trends. Google Trend is used as a proxy to observe users on the web. They found unique trends for the academic searchers which are different from users of a web search engine.

The paper "Exploring Choice Overload in Related-Article Recommendations in Digital Libraries" [3] by Beierle, Aizawa and Beel studies choice overload in scholarly document recommendation in the social sciences search engine sowiport. The authors used click-through rates of different amounts of recommendations as a measure of recommendation effectiveness. Their preliminary results show lower click-through rates for higher numbers of recommendations. According to the experiments, users in the social sciences seem to feel quickly overloaded by increasing choice.

3.4 Session 3: Full papers

The article “Computing Interdisciplinarity of Scholarly Objects using an Author-Citation-Text Model” [15] by Seo, Jung, Kim and Myaeng discusses the computation of the degree of interdisciplinarity of a scholarly object (e.g., an article). To this end, three different sources are used: the author network, the citation network and the actual text. Furthermore, an alternative to measure interdisciplinarity is discussed. Experiments show that the combination of the three aspects author, citations and text of articles can accurately predict the discipline distributions.

In their paper “Detecting Automatically Generated Sentences with Grammatical Structure Similarity” [11], Nguyen Minh Tien and Cyril Labbé tackle the issue of spotting machine generated texts at the sentence level. They introduce a grammatical structure similarity and benchmark it to detect passages stemming from known generators: 80% positive detection rate and less than 1% false detection rate. Editorial workflows could integrate this effective approach to detect questionable manuscripts that editorial staff should check before sending to peer review.

3.5 Poster session

Langer and Beel discuss the use of Lucene in the Docear research paper recommender in their article “Apache Lucene as Content-Based-Filtering Recommender System: 3 Lessons Learned” [6]. They compare Lucene’s relevance score to the click-through rate of a document, finding that Lucene’s scores indeed can be used to determine relevance. The authors also observed that returning ten recommendations out of the top 50 results might be sensible. Furthermore, Lucene is suitable to approximate the recommendation effectiveness.

In their paper “Extending Scientific Literature Search by Including the Author’s Writing Style” [12], Andi Rexha, Mark Kröll, Hermann Ziak, and Roman Kern consider authors’ writing style as a potential feature for paper retrieval and recommendation. They report the results of a pilot study questioning the extent to which individuals identify similarities in authorship. This is a challenging task, even for humans.

In his paper “Drakkar: a graph based All-Nearest Neighbour search algorithm for bibliographic coupling” [17], Bart Thijs discusses the creation of bibliographic coupling graphs based on citations. The proposed algorithm utilizes a bipartite graph constituted by the citing publications and the cited references as well as directed citations.

Siebert, Dinesh and Feyer discuss how scientific recommender systems can be improved by incorporating scientometric measures. In their paper “Extending a Research-Paper Recommendation System with Scientometric Measures” [16] the authors evaluate different reranking approaches in the context of the Mr. DLib research paper recommender system. Readership data is used as an approximation for citation.

In their paper “Semantic embedding for information retrieval”, Wang and Koopman [18] combine bibliometric measures with word embeddings. Word embedding results of well-known systems such as Word2Vec/Doc2Vec and GloVe are compared to the Ariadne approach, showing that Ariadne exhibits a competitive performance in a document embedding for information retrieval task.

4 Discussion Session

The talks sparked stimulating discussions and stressed various initiatives and news at the crossroads between bibliometrics, information retrieval and associated disciplines. Recent announcements from the industry suggest strengthening ties between the two communities.

New initiatives stemming from academia also stress an increasing interest in the topics addressed at the BIR workshops. For instance, the Initiative for Open Citations[7] [14] "promote[s] the unrestricted availability of scholarly citation data" and, as March 2017 "the fraction of publications with open references has grown from 1% to more than 40% out of the nearly 35 million articles with references deposited with Crossref." The availability of citation data at large scale is an opportunity for our community.

An other initiative was discussed at the workshop. The Topic Extraction Challenge[8] tackles the issue of comparing how various approach delineate scientific fields [5]. They release a data set and formulate the challenge as: "We challenge you to comparatively discuss advantages and disadvantages of approaches to topic identification and thus to contribute to a cumulative body of knowledge on the suitability of data models and algorithms for the identification of topics" [4].

5 Conclusion and Future Directions

BIR 2017 has been a successful continuation of past workshops and a further step towards the integration of bibliometrics and IR. With the continuing workshop series, a special issue on "Combining Bibliometrics and Information Retrieval" in the leading *Scientometrics* journal [10] and a further special issue on "Bibliometric-enhanced Information Retrieval" to be published in Scientometrics we have built up a sequence of explorations, visions, results documented in scholarly discourse, and created a sustainable bridge between bibliometrics and IR.

While past events focused in particular on Information Retrieval aspects, we now broaden the scope of our workshop series by offering a Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries[9] (BIRNDL 2017) at SIGIR 2017. BIRNDL will be co-organized with the natural language processing group at National University of Singapore, which includes a shared task (the CL-SciSumm Shared Task[10]). This shared task tackles automatic paper summarization in the Computational Linguistics (CL) domain.

We are working with the *International Journal on Digital Libraries (IJDL)* to offer a special issue on topics discussed at BIR and BIRNDL, for extended versions of workshop papers, shared task descriptions, as well as a general call for submissions.

6 Acknowledgements

We express our gratitude to all PC members for doing an excellent job in reviewing papers for the workshop. This workshop was also supported by the COST Action TD1210 KnowEscape.

[7] https://i4oc.org/

[8] http://www.topic-challenge.info/

[9] `http://wing.comp.nus.edu.sg/birndl-jcdl2017/`

[10] `http://wing.comp.nus.edu.sg/cl-scisumm2017/`

References

[1] M. A. Abdulhayoglu and B. Thijs. Use of Locality Sensitive Hashing (LSH) algorithm to match Web of Science and SCOPUS. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 30–40. CEUR-WS.org, 2017.

[2] J. Beel and S. Dinesh. Real-World Recommender Systems for Academia: The Pain and Gain in Developing, Operating, and Researching them. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 6–17. CEUR-WS.org, 2017.

[3] F. Beierle, A. Aizawa, and J. Beel. Exploring Choice Overload in Related-Article Recommendations in Digital Libraries. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 51–61. CEUR-WS.org, 2017.

[4] K. Boyack, W. Glänzel, J. Gläser, F. Havemann, A. Scharnhorst, B. Thijs, N. J. van Eck, T. Velden, and L. Waltmann. Topic identification challenge. *Scientometrics*, 111(2):1223–1224, 2017.

[5] J. Gläser, W. Glänzel, and A. Scharnhorst. Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, 111(2):981–998, 2017.

[6] S. Langer and J. Beel. Apache Lucene as Content-Based-Filtering Recommender System: 3 Lessons Learned. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 85–92. CEUR-WS.org, 2017.

[7] X. Li and M. de Rijke. Academic Search in Response to Major Scientific Events. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 41–50. CEUR-WS.org, 2017.

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

[9] P. Mayr, I. Frommholz, and G. Cabanac, editors. *Proceedings of the Fifth Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 39th European Conference on Information Retrieval (ECIR 2017), Aberdeen, UK, April 9th, 2017*, volume 1823 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.

[10] P. Mayr and A. Scharnhorst. Scientometrics and Information Retrieval: weak-links revitalized. *Scientometrics*, 102(3):2193–2199, 2015.

[11] M. T. Nguyen and C. Labbé. Detecting Automatically Generated Sentences with Grammatical Structure Similarity. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 73–84. CEUR-WS.org, 2017.

[12] A. Rexha, M. Kröll, H. Ziak, and R. Kern. Extending Scientific Literature Search by Including the Author's Writing Style. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 93–100. CEUR-WS.org, 2017.

[13] J. Rollins, M. McCusker, J. Carlson, and J. Stroll. Manuscript Matcher: A Content and Bibliometrics-based Scholarly Journal Recommendation System. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 18–29. CEUR-WS.org, 2017.

[14] Q. Schiermeier. Initiative aims to break science's citation paywall. *Nature*, apr 2017.

[15] M.-G. Seo, S. Jung, K.-m. Kim, and S.-H. Myaeng. Computing Interdisciplinarity of Scholarly Objects using an Author-Citation-Text Model. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 62–72. CEUR-WS.org, 2017.

[16] S. Siebert, S. Dinesh, and S. Feyer. Extending a Research Paper Recommendation System with Bibliometric Measures. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 112–121. CEUR-WS.org, 2017.

[17] B. Thijs. Drakkar: a graph based All-Nearest Neighbour Search Algorithm for Bibliographic Coupling. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 101–111. CEUR-WS.org, 2017.

[18] S. Wang and R. Koopman. Semantic Embedding for Information Retrieval. In *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, pages 122–132. CEUR-WS.org, 2017.

[19] S. Yang, R. Han, D. Wolfram, and Y. Zhao. Visualizing the intellectual structure of information science (2006-2015): Introducing author keyword coupling analysis. *Journal of Informetrics*, 10(1):132–150, 2016.