Report on the 3rd International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)

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
ingo.frommholz@beds.ac.uk

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

Abstract

To foster collaboration and knowledge transfer between the fields of bibliometrics / scientometrics / informetrics on the one hand and information retrieval on the other hand, we successfully ran a workshop series on bibliometric-enhanced information retrieval (BIR). This workshop report presents the BIR 2016 workshop, which has been co-located with ECIR for the third time this year. We motivate our workshop and outline the papers (one keynote and seven regular papers) presented at ECIR 2016 in Padua, Italy. Finally we reflect on past BIR workshops and conclude with an outlook of future directions.

1 Introduction

Following successful workshops at ECIR 2014[1] and 2015,[2] respectively, the BIR 2016 workshop was the third 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 hand. Our motivation as organizers of

[1] http://gesis.org/en/events/events-archive/conferences/ecirworkshop2014
[2] http://gesis.org/en/events/events-archive/conferences/ecirworkshop2015

the workshop started from the observation that main discourses in both fields are different and communities are only partly overlapping, as well as from the belief that a knowledge transfer is profitable for both sides [15].

The first BIR workshop in 2014 set the research agenda by introducing each group to the other, illustrating state-of-the-art methods, reporting on current research problems, and brainstorming about common interests. The second workshop in 2015 further elaborated these themes. The third full-day BIR workshop[3] at ECIR 2016 aimed at establishing 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 Mendeley and ResearchGate. The third 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 scientific information portals
- IR for scientific domains, e.g. social sciences, life sciences, etc.
- Information seeking behaviour
- 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

3 Paper Presentations

This year 15 papers were submitted to the workshop, 7 of which were finally accepted for presentation and inclusion in the proceedings.[4] The workshop featured one keynote talk and three paper sessions. The first session discussed text and reference mining approaches while the second session focused on bibliometric and IR tools. The final position paper session gave an outlook on further research. The following briefly describes the keynote and sessions.

[3] `http://gesis.org/en/events/events-archive/conferences/ecirworkshop2016`
[4] Available at `http://ceur-ws.org/Vol-1567/`

3.1 Keynote

The keynote "Bibliometrics in online book discussions: Lessons for complex search tasks" [13] was given by Marijn Koolen from the University of Amsterdam. Koolen explored the potential relationships between book search information needs and bibliometric analysis. The CLEF Social Book Search Lab was introduced, which utilizes data from Amazon, LibraryThing (LT), the Library of Congress, and the British Library. LT discussions indicate some complex search tasks. Users catalogue, tag, and relate books to each other. The hypothesis is that reviews, catalogues, and discussion threads could be interpreted as (implicit) co-citation and citation structures. Analyzing comments and reviews, several information need patterns were identified. Koolen also discussed how the data at hand can be utilized for information retrieval.

3.2 Text and Reference Mining

In their paper "Weak links and strong meaning: The complex phenomenon of negational citations" [5], Marc Bertin and Iana Atanassova designed a method to extract negational citations from full-text publications. They revealed the frequency distribution of such citations appearing throughout the regular IMRaD structure of about 80,000 PLOS papers. Qualifying the polarity of citations has many practical applications. This valuable knowledge might inform the scientific community about papers attracting negative feedback that should be reconsidered and potentially retracted.

Andi Rexha, Stefan Klampfl, Mark Kröll, and Roman Kern aimed to chunk papers according to stylometric features in their paper "Towards a more fine grained analysis of sceintific authorship: Predicting the number of authors using stylometric features" [17]. The resulting segments were then attributed to the corresponding author(s) listed in the byline of the paper (i.e., the individuals who co-signed the paper). This contribution is likely to enhance paper/passage retrieval by author name.

In their contribution "The references of references: Enriching library catalogs via domain-specific reference mining" [8], Giovanni Colavizza, Matteo Romanello, and Frédéric Kaplan enhanced a digital library by collecting references from domain-specific reference monographs in the Humanities. Their experiment on a corpus dedicated to the history of Venice stressed the necessity of including such overlooked references to improve search effectiveness in such scholarly corpora.

3.3 Tools for Bibliometric IR

In the paper "*Bibliometrics*: a publication analysis tool" [16] by Rosa Padrós-Cuxart, Clara Riera-Quintero, and Francesc March-Mir, the authors presented a bibliometric data management and consultation tool that can be utilized to study and analyze an institution's scientific activity. The tool is able to generate bibliometric reports on scientific outputs at different analysis levels like author, journal, and institution. It includes data from various sources like WOS/Scopus and provides different indicators like productivity, visibility, impact, and collaboration intensity among scientists.

Detecting fake academic paper is an emerging issue in the scientifc world, which is addressed in the paper "Engineering a tool to detect automatically generated papers" [20]. Here, Nguyen Minh Tien and Cyril Labbé focused on detecting fake academic papers that

were automatically created. The authors worked on detection approaches based on distance/similarity measurement and introduced a tool able to detect automatically generated papers: the SciDetect system. The authors evaluated it against pattern matching and Kullback-Leibler Divergence on three different text corpora.

3.4 IR Position Papers

In his article "Bag of works retrieval: TF*IDF weighting of co-cited works", Howard D. White proposed an analogy to the well-known bag of words model called *bag of works* [23]. This model can in particular be used for finding similar documents to a given seed one. In the proposed bag of works model, the *tf* and *idf* measures are re-defined based on (co-)citation counts. The properties of the retrieved documents were discussed and an example provided.

In their article "On the need for and provision for an 'IDEAL' scientific information retrieval test collection" [14], Birger Larsen and Christina Lioma argued there is a need for test collections tailored to bibliometric IR. They discussed several challenges coming along with creating such a collection (e.g., regarding size, domain-specific dissemination and retrieval, realistic queries and relevance judgements, pooling strategies as well as format). Furthermore, procedures to create an ideal test collection were examined.

4 Discussion Sessions

The talks sparked stimulating discussions and stressed various initiatives and news at the crossroads between bibliometrics and information retrieval.

Recent announcements from the industry suggest strengthening ties between the two communities. For instance, Microsoft released the *Microsoft Academic Graph* [18] in June 2015.[5] This is a frequently updated large bibliographic dataset of more than 120 million papers. This volume is to be compared to the yearly 1.3 million papers published worldwide according to the most recent estimates [19]. Search engines dedicated to scholarly materials were also launched recently, such as the *Microsoft Academic* platform[6] (redux from *Microsoft Academic Search*) and the Allen Institute's *Semantic Scholar*[7] [12].

New initiatives stemming from academia also stress an increasing interest in the topics addressed at the BIR workshops. For instance, the TREC OpenSearch Academic Edition[8] will run in 2016 to assess *ad hoc* literature search engines. Another initiative will be held at JCDL'16 under name CL-SciSumm.[9] SciSumm aims to evaluate scientific document summarization.

Eventually, the combined availability of bibliographic data and the development of dedicated search engines and performance indicators might raise several issues. A wide range of pitfalls have been identified by the bibliometric community and a state-of-the-art account is given in the *Leiden Manifesto* [10]. Such guidelines should be kept in mind while designing information systems that guarantee an ethical use of bibliographic and bibliometric data.

[5] http://research.microsoft.com/en-us/projects/mag
[6] http://academic.research.microsoft.com
[7] http://allenai.org/semantic-scholar
[8] http://web.archive.org/web/20160406200108/http://trec-open-search.org
[9] http://wing.comp.nus.edu.sg/cl-scisumm2016

5 Previous BIR Workshops

Previous Bibliometric-enhanced IR workshops at ECIR in Amsterdam 2014 and in Vienna 2015 have generated a wide range of papers.[10] The main directions of these workshop papers have been:

- IR and recommendation tool development and evaluation [21, 22, 6]
- Bibliometric IR experiments and data sets [11, 7, 9]
- Document Clustering for IR [1, 2]
- Citation Contexts and Analysis [3, 4, 24].

6 Conclusion and Future Directions

BIR 2016 has been a successful continuation of past workshops and a further step towards the integration of bibliometrics and IR. With the continuing workshop series and a special issue on "Combining Bibliometrics and Information Retrieval" in the leading *Scientometrics* journal [15] 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[11] (BIRNDL 2016) at JCDL 2016. BIRNDL will be co-organized with the natural language processing group of Min-Yen Kan, National University of Singapore, which includes a shared task (the CL-SciSumm Shared Task[12]). 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. Dates for first submission of papers will likely be around September 2016, with a target of producing an issue by mid 2017.

7 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 an ongoing COST Action TD1210 KnowEscape.

References

[1] M. K. Abbasi and I. Frommholz. Exploiting Information Needs and Bibliographics for Polyrepresentative Document Clustering. In *Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval (BIR 2014)*, pages 21–28, 2014.

[10] Proceedings are available at `http://ceur-ws.org/Vol-1344/` and `http://ceur-ws.org/Vol-1143/`

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

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

[2] M. K. Abbasi and I. Frommholz. Polyrepresentative Clustering: A Study of Simulated User Strategies and Representations. In *Proc. of the 2nd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2015)*, pages 47–54, 2015.

[3] M. Bertin and I. Atanassova. A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. In *Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval (BIR 2014)*, pages 5–12, 2014.

[4] M. Bertin and I. Atanassova. Factorial Correspondence Analysis Applied to Citation Contexts. In *Proc. of the 2nd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2015)*, pages 22–29, 2015.

[5] M. Bertin and I. Atanassova. Weak links and strong meaning: The complex phenomenon of negational citations. In *Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*, pages 14–25, 2016.

[6] Z. Carevic and P. Mayr. Extending search facilities via bibliometric-enhanced stratagems. In *Proc. of the 2nd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2015)*, pages 40–46, 2015.

[7] Z. Carevic and P. Schaer. On the Connection Between Citation-based and Topical Relevance Ranking: Results of a Pretest using iSearch. In *Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval (BIR 2014)*, pages 37–44, 2014.

[8] G. Colavizza, M. Romanello, and F. Kaplan. The references of references: Enriching library catalogs via domain-specific reference mining. In *Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*, pages 32–43, 2016.

[9] A. Dabrowska and B. Larsen. Exploiting Citation Contexts for Physics Retrieval. In *Proc. of the 2nd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2015)*, pages 14–21, 2015.

[10] D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols. Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548):429–431, 2015.

[11] K. Jack, P. López-García, M. Hristakeva, and R. Kern. {{citation needed}}: Filling in Wikipedia's Citation Shaped Holes. In *Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval (BIR 2014)*, pages 45–52, 2014.

[12] N. Jones. Artificial-intelligence institute launches free science search engine. *Nature News*, 2015.

[13] M. Koolen. Bibliometrics in online book discussions: Lessons for complex search tasks. In *Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*, pages 5–13, 2016.

[14] B. Larsen and C. Lioma. On the need for and provision for an 'IDEAL' scholarly information retrieval test collection. In *Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*, pages 73–81, 2016.

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

[16] R. Padrós-Cuxart and C. Riera-Quintero. *Bibliometrics*: a publication analysis tool. In *Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*, pages 44–53, 2016.

[17] A. Rexha, S. Klampfl, M. Kröll, and R. Kern. Towards a more fine grained analysis of scientific authorship: Predicting the number of authors using stylometric features. In *Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*, pages 26–31, 2016.

[18] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang. An overview of microsoft academic service (MAS) and applications. In *WWW'15 Companion: Proceedings of the 24th International Conference on World Wide Web*, pages 243–246, New York, NY, 2015. ACM.

[19] L. Soete, S. Schneegans, D. Eröcal, B. Angathevar, and R. Rasiah. A world in search of an effective growth strategy. In S. Schneegans, editor, *UNESCO Science Report: Towards 2030*, UNESCO Reference Works, chapter 1, pages 20–55. Paris, 2015.

[20] N. M. Tien and C. Labbé. Engineering a tool to detect automatically generated papers. In *Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*, pages 54–62, 2016.

[21] N. J. van Eck and L. Waltman. Systematic Retrieval of Scientific Literature based on Citation Relations: Introducing the CitNetExplorer Tool. In *Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval (BIR 2014)*, pages 13–20, 2014.

[22] I. Wesley-Smith, R. Dandrea, and J. West. An Experimental Platform for Scholarly Article Recommendation. In *Proc. of the 2nd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2015)*, pages 30–39, 2015.

[23] H. D. White. Bag of works retrieval: TF*IDF weighting of co-cited works. In *Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*, pages 63–72, 2016.

[24] H. Zhao and X. Hu. Language Model Document Priors based on Citation and Co-citation Analysis. In *Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval (BIR 2014)*, pages 29–36, 2014.