Bibliometric-enhanced Information Retrieval: 5th International BIR Workshop

Philipp Mayr[1], Ingo Frommholz[2], and Guillaume Cabanac[3]

[1] GESIS - Leibniz-Institute for the Social Sciences, Cologne, Germany,
`philipp.mayr@gesis.org`
[2] Institute for Research in Applicable Computing, University of Bedfordshire,
Luton, UK,
`ifrommholz@acm.org`
[3] University of Toulouse, Computer Science Department, IRIT UMR 5505, France
`guillaume.cabanac@univ-tlse3.fr`

Abstract. Bibliometric-enhanced Information Retrieval (BIR) workshops serve as the annual gathering of IR researchers who address various information-related tasks on scientific corpora and bibliometrics. The workshop features original approaches to search, browse, and discover value-added knowledge from scientific documents and related information networks (e.g., terms, authors, institutions, references). We welcome contributions elaborating on dedicated IR systems, as well as studies revealing original characteristics on how scientific knowledge is created, communicated, and used. In this paper we introduce the BIR workshop series and discuss some selected papers presented at previous BIR workshops.

Keywords: Bibliometrics, Scientometrics, Informetrics, Information Retrieval, Digital Libraries

1 Introduction

Following the successful workshops at ECIR 2014[4], 2015[5], 2016[6] and JCDL 2016[7], respectively, this workshop is 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 hand. Our motivation as organizers of the workshop started from the observation that main discourses in both fields are different, that communities are only partly overlapping and from the belief that a knowledge transfer would be profitable for both sides [1,2]. The need for researchers to keep up-to-date with their respective field given the highly increasing number of

---

[4] `http://ceur-ws.org/Vol-1143/`
[5] `http://ceur-ws.org/Vol-1344/`
[6] `http://ceur-ws.org/Vol-1567/`
[7] `http://ceur-ws.org/Vol-1610/`

publications available has led to the establishment of scientific repositories that allow us to use additional evidence coming for instance from citation graphs to satisfy users' information needs.

The first BIR workshops in 2014 and 2015 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 third workshop in 2016 [3] further elaborated on these themes. For the fourth workshop, co-located with the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) 2016, we broadened the workshop scope and interlinked the BIR workshop with the natural language processing (NLP) and computational linguistics field [4]. This 5th full-day BIR workshop at ECIR 2017 aims to foster a common ground for the incorporation of bibliometric-enhanced services (incl. text mining functionality) into scholarly search engine interfaces. In particular we address specific communities, as well as studies on large, cross-domain collections like Mendeley and ResearchGate. This fifth BIR workshop again addresses explicitly both scholarly and industrial researchers.

2 Goals, Objectives and Outcomes

Our workshop aims to engage the IR community with possible links to bibliometrics. Bibliometric techniques are not yet widely used to enhance retrieval processes in digital libraries, yet they offer value-added effects for users [5]. Hence, our objective is to bring together information retrieval, information seeking, science modelling, network analysis, and digital libraries to apply insights from bibliometrics, scientometrics, informetrics and text mining to concrete, practical problems of information retrieval and browsing. We discuss some examples from previous workshops in Section 5. More specifically we ask questions like:

- How can we generalize paper tracking on social media?
  a.k.a. altmetrics on steroïds: beyond DOI spotting.
- How can we detect fake reviews [6] to sustain the peer review process?
- How can we improve homonym detection (e.g., Li Li) in bibliographic records [7]?
- To what degree can we automate fact-checking [8,9] in academic papers?
- How can we support researchers in finding relevant scientific literature, e.g., by integrating ideas from information retrieval, information seeking and searching and bibliometrics [10,11]?
- How can we build scholarly information systems that explicitly use bibliometric measures at the user interface (e.g. contextual bibliometric-enhanced features [12])?
- How can models of science be interrelated with scholarly, task-oriented searching?
- How can we combine classical IR (with emphasis on recall and weak associations) with more rigid bibliometric recommendations [13,14]?
- How can we create suitable testbeds (like iSearch corpus) [15]?

3 Format and Structure of the Workshop

The workshop will start with an inspirational keynote "Real-World Recommender Systems for Academia: The Pain and Gain in Developing, Operating, and Researching them" by Joeran Beel (Trinity College Dublin, the School of Computer Science and Statistics) to kick-start thinking and discussion on the workshop topic. This will be followed by paper presentations in a format that we found to be successful at previous BIR workshops: each paper is presented as a 10 minute lightning talk and discussed for 20 minutes in groups among the workshop participants followed by 1-minute pitches from each group on the main issues discussed and lessons learned. The workshop will conclude with a round-robin discussion of how to progress in enhancing IR with bibliometric methods.

4 Audience

The audiences of IR and bibliometrics overlap [1,2]. Traditional IR serves individual information needs, and is – consequently – embedded in libraries, archives and collections alike. Scientometrics, and with it bibliometric techniques, has a matured serving science policy. We therefore will hold a full-day workshop that brings together IR researchers with those interested in bibliometric-enhanced approaches. Our interests include information retrieval, information seeking, science modelling, network analysis, and digital libraries. The workshop is closely related to the past BIR workshops at ECIR 2014, 2015, 2016 and strives to feature contributions from core bibliometricians and core IR specialists who already operate at the interface between scientometrics and IR. While the past workshops laid the foundations for further work and also made the benefit of bringing information retrieval and bibliometrics together more explicit, there are still many challenges ahead. One of them is to provide infrastructures and testbeds for the evaluation of retrieval approaches that utilise bibliometrics and scientometrics. To this end, a focus of the proposed workshop and the discussion will be on real experimentations (including demos) and industrial participation. This line was started in a related workshop at JCDL (BIRNDL 2016), but with a focus on digital libraries and computational linguistics and not on information retrieval and information seeking and searching.

5 Selected papers and past Keynotes

Past BIR workshops had invited talks of several experts working in the field of bibliometrics and information retrieval. Last year, Marijn Koolen gave a keynote on "Bibliometrics in online book discussions: Lessons for complex search tasks" [16]. Koolen explored the potential relationships between book search information needs and bibliometric analysis and introduced the Social Book Search Lab, triggering a discussion on the relationship between book search and

bibliometric-enhanced IR. In 2015, the keynote "In Praise of Interdisciplinary Research through Scientometrics" [17] was given by Guillaume Cabanac. Cabanac accentuated the potential of interdisciplinary research at the interface of information retrieval and bibliometrics. He came up with many research questions that lie at the crossroad of scientometrics and other fields, namely information retrieval, digital libraries, psychology and sociology.

Recent examples of BIR workshop publications have shown the potential of informing the information retrieval process with bibliometrics. These examples comprise topics like IR and recommendation tool development, bibliometric IR evaluation and data sets, and the application and analysis of citation contexts for instance for cluster-based search.

As an example of recommendation tool development utilising bibliometrics, Wesley-Smith et al. [18] describe an experimental platform constructed in collaboration with the repository Social Science Research Network (SSRN) in order to test the effectiveness of different approaches for scholarly article recommendations. Jack et al. [19] present a case study on how to increase the number of citations to support claims in Wikipedia. They analyse the distribution of more than 9 million citations in Wikipedia and found that more than 400,000 times an explicit marker for a needed citation is present. To overcome this situation they propose different techniques based on Bradfordizing and popularity number of readers in Mendeley to implement a citation recommending system. The authors conclude that a normal keyword-based search engine like Google Scholar is not sufficient to be used to provide citation recommendation for Wikipedia articles and that altmetrics like readership information can improve retrieval and recommendation performance.

Utilising a collection based on PLOS articles, Bertin and Atanassova [20] try to further unravel the riddle of meaning of citations. The authors analyse the word use in standard parts of articles, such as Introduction, Methods, Results and Discussion, and reveal interesting distributions of the use of verbs for those sections. The authors propose to use this work in future citation classifiers, which in the long-term might also be implemented in citation-based information retrieval.

As an application of citation analysis, Abbasi and Frommholz [21] investigate the benefit of combining polyrepresentation with document clustering, where representations are informed by citation analysis. The evaluation of the proposed model on the basis of the iSearch collection shows some potential of the approach to improve retrieval quality. A further application example reported by Nees Jan van Eck and Ludo Waltman [22] considers the problem of scientific literature search. The authors suggest that citation relations between publications can be a helpful instrument in the systematic retrieval process of scientific literature. They introduce a new software tool called CitNetExplorer that can be used for citation-based scientific literature retrieval. To demonstrate the use of CitNetExplorer, they employ the tool to identify publications dealing with the topic of "community detection in networks". They argue that their approach can be especially helpful in situations in which one needs a comprehensive overview

of the literature on a certain research topic, for instance in the preparation of a review article.

Howard D. White proposes an alternative 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 bag of works model, tf and idf measures are re-defined based on (co-)citation counts. The properties of the retrieved documents are discussed and an example is provided.

6 Output

In 2015 we published a first special issue on "Combining Bibliometrics and Information Retrieval" in Scientometrics [1]. A special issue on "Bibliometrics, Information Retrieval and Natural Language Processing in Digital Libraries" is currently under preparation for the International Journal on Digital Libraries. For this year's ECIR workshop we continue the tradition of producing follow-up special issues. Authors of accepted papers at this year's BIR workshop will again be invited to submit extended versions to a special issue on "Bibliometric-enhanced IR" to be published in Scientometrics.

References

1. Mayr, P., Scharnhorst, A.: Scientometrics and Information Retrieval: weak-links revitalized. Scientometrics **102**(3) (2015) 2193–2199
2. Wolfram, D.: The symbiotic relationship between information retrieval and informetrics. Scientometrics **102**(3) (2015) 2201–2214
3. Mayr, P., Frommholz, I., Cabanac, G.: Report on the 3rd International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016). SIGIR Forum **50**(1) (2016) 28–34
4. 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) (2016) 36–43
5. Mutschke, P., Mayr, P., Schaer, P., Sure, Y.: Science models as value-added services for scholarly information systems. Scientometrics **89**(1) (2011) 349–364
6. Bartoli, A., De Lorenzo, A., Medvet, E., Tarlao, F.: Your paper has been accepted, rejected, or whatever: Automatic generation of scientific paper reviews. In: CD-ARES'16: Proceedings of the IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference. Number 9817 in LNCS (2016) 19–28
7. Momeni, F., Mayr, P.: Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names. In: 20th International Conference on Theory and Practice of Digital Libraries (TPDL 2016). (2016) 386–391
8. Baker, M.: Smart software spots statistical errors in psychology papers. Nature (2015)
9. Ziemann, M., Eren, Y., El-Osta, A.: Gene name errors are widespread in the scientific literature. Genome Biology **17**(1) (2016) 1–3
10. Abbasi, M.K., Frommholz, I.: Cluster-based Polyrepresentation as Science Modelling Approach for Information Retrieval. Scientometrics **102**(3) (2015) 2301–2322

11. Mutschke, P., Mayr, P.: Science Models for Search. A Study on Combining Scholarly Information Retrieval and Scientometrics. Scientometrics **102**(3) (2015) 2323–2345

12. Carevic, Z., Mayr, P.: Survey on High-level Search Activities Based on the Stratagem Level in Digital Libraries. In: 20th International Conference on Theory and Practice of Digital Libraries (TPDL 2016). (2016) 54–66

13. Zitt, M.: Meso-level retrieval: Ir-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation. Scientometrics **102**(3) (2015) 2223–2245

14. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. International Journal on Digital Libraries **17**(4) (2016) 305–338

15. Larsen, B., Lioma, C.: On the need for and provision for an 'ideal' scholarly information retrieval test collection. In: Proc. of the Third Workshop on Bibliometric-enhanced Information Retrieval. (2016) 73–81

16. Koolen, M.: Bibliometrics in online book discussions: Lessons for complex search tasks. In: Proceedings of the Third Workshop on Bibliometric-enhanced Information Retrieval co-located with the 38th European Conference on Information Retrieval (ECIR 2016), Padova, Italy, March 20, 2016. (2016) 5–13

17. Cabanac, G.: In praise of interdisciplinary research through scientometrics. In: Proc. of the Second Workshop on Bibliometric-enhanced Information Retrieval. (2015) 5–13

18. Wesley-Smith, I., Dandrea, R.J., West, J.D.: An experimental platform for scholarly article recommendation. In: Proc. of the Second Workshop on Bibliometric-enhanced Information Retrieval. (2015) 30–39

19. Jack, K., López-García, P., Hristakeva, M., Kern, R.: {{citation needed}}: Filling in wikipedia's citation shaped holes. In: Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval. (2014) 45–52

20. Bertin, M., Atanassova, I.: A study of lexical distribution in citation contexts through the imrad standard. In: Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval. (2014) 5–12

21. Abbasi, M.K., Frommholz, I.: Exploiting information needs and bibliographics for polyrepresentative document clustering. In: Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval. (2014) 21–28

22. van Eck, N.J., Waltman, L.: Systematic retrieval of scientific literature based on citation relations: Introducing the citnetexplorer tool. In: Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval. (2014) 13–20

23. White, H.D.: Bag of works retrieval: Tf*idf weighting of co-cited works. In: Proc. of the Third Workshop on Bibliometric-enhanced Information Retrieval. (2016) 63–72