Editorial for the 3rd Bibliometric-Enhanced Information Retrieval Workshop at ECIR 2016

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,
ingo.frommholz@beds.ac.uk
[3] University of Toulouse, Computer Science Department, IRIT UMR 5505, France
guillaume.cabanac@univ-tlse3.fr

1 Introduction

Following the successful workshops at ECIR 2014[4] and 2015[5], respectively, this 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. 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]. 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. This third full-day BIR workshop[6] at ECIR 2016 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 Mendeley and ResearchGate. This third BIR workshop addressed explicitly both scholarly and industrial researchers.

2 Overview of the papers

This year 15 papers were submitted to the workshop, 7 of which were finally accepted for presentation and inclusion in the proceedings. 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.

[4] http://gesis.org/en/events/events-archive/conferences/ecirworkshop2014
[5] http://gesis.org/en/events/events-archive/conferences/ecirworkshop2015
[6] http://gesis.org/en/events/events-archive/conferences/ecirworkshop2016

2.1 Keynote

The keynote "Bibliometrics in online book discussions: Lessons for complex search tasks" [2] was given by Marijn Koolen from the University of Amsterdam. Koolen explores the potential relationships between book search information needs and bibliometric analysis. The Social Book Search Lab is 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 can be identified. Koolen also discusses how the data at hand can be utilized for information retrieval.

2.2 Text and Reference Mining

In their paper "Weak links and strong meaning: The complex phenomenon of negational citations" [3], 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.

In their paper "Towards a more fine grained analysis of sceintific authorship: Predicting the number of authors using stylometric features" [4] Andi Rexha, Stefan Klampfl, Mark Kröll, and Roman Kern aimed to chunk papers according to stylometric features. 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 paper "The references of references: Enriching library catalogs via domain-specific reference mining" [5], 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 stresses the necessity of including such overlooked references to improve search effectiveness in such corpora.

2.3 Tools for Bibliometric IR

In the paper "*Bibliometrics*: a publication analysis tool" [6] by Rosa Padrós-Cuxart, Clara Riera-Quintero, and Francesc March-Mir, the authors present 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. The tool includes data from various sources

like WOS/Scopus and provides different indicators like productivity, visibility, impact, and collaboration.

In the paper "Engineering a tool to detect automatically generated papers" [7] by Nguyen Minh Tien and Cyril Labbé, the authors are focussing on detecting fake academic papers that are automatically created. The authors work on detection approaches based on distance/similarity measurement and introduce a tool which is able to detect automatically generated papers, the SciDetect system. The authors evaluate the SciDetect system against pattern matching and Kullback-Leibler Divergence on three different text corpora.

2.4 IR Position Papers

In his article "Bag of works retrieval: TF*IDF weighting of co-cited works", Howard D. White proposes an alternative to the well-known bag of words model called *bag of works* [8]. 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 are discussed and an example is provided.

In their article "On the need for and provision for an 'IDEAL' scientific information retrieval test collection" [9], Birger Larsen and Christina Lioma argue there is a need for test collections tailored to bibliometric IR. They discuss 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 are examined.

3 Outlook

With this continuing workshop series we have built up a sequence of explorations, visions, results documented in scholarly discourse, and created a sustainable bridge between bibliometrics and IR.

As a next iteration we will organize a Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016)[7] at the JCDL conference 2016. The BIRNDL workshop will be co-organized together with the natural language processing group of Min-Yen Kan, National University of Singapore, which includes a shared task (the CL-SciSumm Shared Task[8]). The shared task tackles automatic paper summarization in the Computational Linguistics (CL) domain.

[7] http://wing.comp.nus.edu.sg/birndl-jcdl2016/
[8] http://wing.comp.nus.edu.sg/cl-scisumm2016/

References

1. Mayr, P., Scharnhorst, A.: Scientometrics and Information Retrieval: weak-links revitalized. Scientometrics **102**(3) (2015) 2193–2199
2. Koolen, M.: Bibliometrics in online book discussions: Lessons for complex search tasks. In: Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR2016). 5–13
3. Bertin, M., Atanassova, I.: Weak links and strong meaning: The complex phenomenon of negational citations. In: Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR2016). 14–25
4. Rexha, A., Klampfl, S., Kröll, M., Kern, R.: 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 (BIR2016). 26–31
5. Colavizza, G., Romanello, M., Kaplan, F.: The references of references: Enriching library catalogs via domain-specific reference mining. In: Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR2016). 32–43
6. Padrós-Cuxart, R., Riera-Quintero, C.: *Bibliometrics*: a publication analysis tool. In: Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR2016). 44–53
7. Tien, N.M., Labbé, C.: Engineering a tool to detect automatically generated papers. In: Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR2016). 54–62
8. White, H.D.: Bag of works retrieval: TF*IDF weighting of co-cited works. In: Proc. of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR2016). 63–72
9. Larsen, B., Lioma, C.: 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 (BIR2016). 73–81