Report on the 11th Bibliometric-enhanced Information Retrieval Workshop (BIR 2021)

Ingo Frommholz
University of Wolverhampton, UK
ifrommholz@acm.org

Guillaume Cabanac
University of Toulouse, France
guillaume.cabanac@univ-tlse3.fr

Philipp Mayr
GESIS, Cologne, Germany
philipp.mayr@gesis.org

Suzan Verberne
Leiden University, the Netherlands
s.verberne@liacs.leidenuniv.nl

Abstract

The 11th Bibliometric-enhanced Information Retrieval Workshop (BIR 2021) was held online on April 1st, 2021, at ECIR 2021 as a virtual event. The interdisciplinary BIR workshop series aims to bring together researchers from different communities, especially Scientometrics/Bibliometrics and Information Retrieval. We report on the 11th BIR, its invited talks and accepted papers. Lessons learned from BIR 2021 are discussed and potential future research questions identified that position Bibliometric-enhanced IR as an exciting special yet important branch of IR research.

1 Introduction

The BIR (Bibliometric-enhanced Information Retrieval) workshop series has a long-established tradition of bringing together researchers from different communities, especially Scientometrics/Bibliometrics and Information Retrieval. BIR was launched at ECIR in 2014 and was held at ECIR each year since then.[1] BIR 2021 was organised by the authors of this report. Due to the ongoing pandemic, the workshop took place online only in conjunction with ECIR 2021. The workshop attracted around 57 participants at peak times but a larger number throughout due to participants dropping in and out. Workshop proceedings were published at CEUR-WS [Cabanac et al., 2021a].

2 Accepted Papers and Keynote Talks

BIR consisted of invited talks as well as presentations of peer-reviewed, accepted papers. This year five papers were accepted as full papers and four papers as short papers. Most of the talks

[1] All pointers to past and future workshops as well as to proceedings are hosted at https://sites.google.com/view/bir-ws/

were recorded; the recordings are available in the BIR 2021 YouTube playlist.[2] In addition, the workshop featured three keynote talks:

- **Ludo Waltman (CWTS, the Netherlands)** addressed openness, transparency, and inclusivity in science, and in particular the question: what does it mean for information retrieval? This was nicely related to the panel about open access during ECIR the day before. Ludo discussed how research is moving in a direction of increased openness, transparency, and inclusivity and the new possibilities this offers for scholarly literature search. Calls for increased transparency and inclusivity raise complex questions about the responsibilities of those who manage search systems for scholarly literature and about the benefits as well as the risks of new AI-based approaches to scholarly literature search. While acknowledging that there are no easy answers, Ludo shared his thoughts on the various issues that the BIR community may need to reflect on, including open metadata, references, and abstracts.

- **Lucy Lu Wang (Allen Institute for AI, USA)** discussed text mining insights from the COVID-19 pandemic. She described the emergence of novel information retrieval and NLP tasks with the potential to change the way information from the scientific literature is communicated to healthcare providers and public health researchers. Lucy discussed some of the ways the computing community came together to tackle this challenge, with the release of open data resources like CORD-19 and the introduction of various shared tasks for evaluation. She also presented her work on scientific fact-checking, a novel NLP task that looks to address issues around scientific misinformation, and its practical uses in managing conflicting information arising from COVID-19 pandemic publishing.

- **Jimmy Lin (University of Waterloo, Canada)** presented approaches to domain adaptation for scientific texts and discussed the limits of scale. He argued that a fundamental assumption behind bibliometric-enhanced information retrieval is that ranking models need to be adapted to handle scientific text, which is very different from the typical corpora (Wikipedia, books, web crawls, etc.) used to pretrain large-scale transformers. One common approach is to take a large "general-domain" model and then apply domain adaptation techniques to "customize" it for a specific (scientific) domain. However, it appears that the far less satisfying approach of "just throwing more data at the problem" with increasingly larger pretrained transformers seems to be more effective. In fact, over the last year, Jimmy's group has "won" multiple community-wide shared evaluations focused on texts related to the novel coronavirus SARS-CoV-2 using exactly this approach: document ranking (TREC-COVID, TREC Health Misinformation), question answering (EPIC-QA), and fact verification (SciFcat). Jimmy shared their efforts to grapple with the issues of why "smarter" is not better than "larger", and opened up the discussion to try to understand why.

The following research papers were presented (a more topic-focused discussion is provided in the next section):

- *Shintaro Yamamoto, Anne Lauscher, Simone Paolo Ponzetto, Goran Glavaš and Shigeo Morishima*: Self-Supervised Learning for Visual Summary Identification in Scientific Publications [Yamamoto et al., 2021];

[2] https://www.youtube.com/playlist?list=PLK4WlSr348zmuZHOzC15W2DYR3gdF6n9n

- *Pablo Accuosto, Mariana Neves and Horacio Saggion*: Argumentation mining in scientific literature: From computational linguistics to biomedicine [Accuosto et al., 2021];
- *Frederique Bordignon, Liana Ermakova and Marianne Noel*: Preprint abstracts in times of crisis: a comparative study with the pre-pandemic period [Bordignon et al., 2021];
- *Hiran H. Lathabai, Abhirup Nandy and Vivek Kumar Singh*: Expertise based institutional recommendation in different thematic areas [Lathabai et al., 2021];
- *Ahmed Abura'Ed and Horacio Saggion*: A select and rewrite approach to the generation of related work reports [Abura'Ed and Saggion, 2021];
- *Ken Voskuil and Suzan Verberne*: Improving reference mining in patents with BERT [Voskuil and Verberne, 2021];
- *Manajit Chakraborty, David Zimmermann and Fabio Crestani*: PatentQuest: A User-Oriented Tool for Integrated Patent Search [Chakraborty et al., 2021];
- *Jacqueline Sachse*: Bibliometric Indicators and Relevance Criteria – An Online Experiment [Sachse, 2021]; and
- *Daria Alexander and Arjen P. de Vries*: "This research is funded by...": Named Entity Recognition of financial information in research papers [Alexander and de Vries, 2021].

3 Discussion

We summarize and reflect on the main takeaway messages from BIR 2021 before looking further to discuss emerging research directions in the context of Bibliometric-enhanced IR.

3.1 Summary and Reflection

The post-anniversary 11th BIR workshop 2021[3] was a great success and showed again the relevance of the interdisciplinary BIR workshop series to the communities involved. The accepted papers and keynote talks demonstrated that BIR addresses important topics that should be tackled interdisciplinary by the Bibliometrics/Scientometrics and IR communities, as well as incorporating contributions coming from the NLP and Machine Learning communities. Traditionally, the Scientometrics and IR communities were intertwined [White and McCain, 1998] and one of BIR's aims is to connect these communities again. This enables us to address pressing problems, for instance, how to handle rapid publication cycles and make a large number of scientific preprints accessible in times of crisis like the current pandemic, or how the communities can contribute to openness, transparency, and inclusivity in science, as it was discussed in Waltman's keynote talk.

Authors and keynote speakers of the BIR workshop addressed several topics and application areas. The keynotes by Wang and Lin demonstrate the focus on the current pandemic, a theme that was also picked up by other authors. Accuosto et al. [2021] reported on work in the more

[3]See the anniversary workshop summary with an overview of the BIR workshop series in Cabanac et al. [2020].

general field of biomedicine utilising argument mining as a means to access relevant information more rapidly; Bordignon et al. [2021] look at preprint abstracts as a direct response to the current crisis to speed up knowledge dissemination. Summarization was another big theme at BIR 2021, for instance as a means to generate related work sections [Abura'Ed and Saggion, 2021] or to identify visual summaries for better access to the content of a scientific publication [Yamamoto et al., 2021]. Bibliometrics are traditionally used to rank researchers and institutions, which was the focus of the work by Lathabai et al. [2021] who covered the use case that researchers need institutes to complement their expertise, for instance for project proposals. The work by Alexander and de Vries [2021] is an example of the importance of natural language processing and information extraction for both the Bibliometrics and IR communities. A core and well-established part of IR research is evaluation, which leads to the question of whether scientific literature requires special care in this respect. To this end, Sachse [2021] investigates the relevance criteria and the role bibliometric indicators play to establish relevance. Finally, some authors looked at a special area of both Bibliometrics and IR, which is patent search. Chakraborty et al. [2021] suggest a system that deals as a single point of access for patent information from different sources. Voskuil and Verberne [2021] look at the extraction of scientific references from patents, which allows us to answer questions about the types of scientific references that eventually lead to innovation.

BIR 2021 demonstrated again that a collaboration between IR and Bibliometrics can lead to a fruitful exchange to tackle important and timely problems. An interdisciplinary effort is made, for instance, to ensure researchers have access to high-quality publications while ensuring rapid turnaround times. The workshop itself attracted a large number of participants despite not being free for authors, which we interpret as a strong signal for the high interest in this topic.

3.2 Future Research Directions

Which kind of topics should we address in the future? What are current and new avenues of research? We identified a number of potential research directions from this year's and previous BIR workshops:

- Interactive search and recommendation in scholarly big data collections. On the one hand, it seems the Bibliometrics community has not yet been exposed to the latest developments in interactive IR and benchmark evaluations. On the other hand, the latest IR approaches, theories and concepts were not applied to scholarly big data collections or took insights from the Bibliometrics community into account. This includes user models and the application of established theories such as Information Foraging Theory [Liu et al., 2010; Maxwell and Azzopardi, 2018] and polyrepresentation [Abbasi and Frommholz, 2015]. Also, searching scientific data is often a known-item search ("I remember I read this in one paper, which one was it?"). IR research has developed many promising approaches that can be applied in this context. In general, researchers, students and policy-makers (e.g., politicians who need to base their decisions on evidence) usually have specific information needs that should be addressed by specialised search and recommendation solutions. This also includes the latest developments in neural IR and semantic embeddings such as BERT, with the challenge that scientific text might be very different from the typical corpora, as indicated in Jimmy Lin's keynote. Search in scholarly big data collections is not restricted to documents, but can also include the search for suitable *data sets* and code, for instance, to re-run analyses to confirm

and reproduce results, and gain further insights [Koesten et al., 2018]. This adheres to the growing and welcome tendency to publish data sets alongside literature.

- Related to the previous point, the IR community has seen a larger interest in combining Human-Computer Interaction with IR, resulting in the ACM CHIIR conference series. Indeed, appropriate user interfaces and visualizations play a very important role in supporting users' information seeking and searching, which, for example, not only triggered research in search user interfaces [Hearst, 2010] but also inspired the creation of formal models for interactive retrieval interfaces [Zhang and Zhai, 2015]. However, to our knowledge, none of those models have been extensively applied to scholarly information seeking and searching.

- Another important research area is summarization and along with that, text simplification.[4] The goal is to make the material easier accessible and also more digestible for non-specialists. More refined approaches utilise argument mining (e.g., [Accuosto et al., 2021]) to give an insight into the argumentative structure of scientific discussions. Fact-checking, handling misinformation and managing conflicting information arising scientific questions are further very important aspects to address, as outlined by Lucy Lu Wang's keynote.

- Entity extraction and the creation of knowledge graphs of scholarly publications can potentially benefit from bibliographic metadata [Turki et al., 2021]. This might also foster the exploration of new knowledge through text mining to discover entities on the fly during exploration, potentially based on entities highlighted by the user, similar to relevance feedback in IR. Systems supporting such techniques might be able to aid decisions for instance about vaccinations, by listing benefits and potential risks, to enable users to make an informed decision.[5] It might also help to emphasize differences between pre-prints and final published (and peer-reviewed) versions, to learn how scientific results have been improved.

- Evaluation has a long tradition in IR through benchmarking and evaluation initiatives, but focused efforts are still missing in the context of Bibliometric-enhanced IR. However, works in this workshop give us an indication about the special relevance criteria users of scholarly databases might have [Sachse, 2021], including potential quality criteria such as trusted researchers, prestigious groups, the application of a proper methodology, etc. More studies are needed in this respect. Furthermore, suitable test collections should be created. There have been early efforts in this respect [Lykke et al., 2010; Gläser et al., 2017], from the computational linguistics community in the context of the CL-SciSumm Shared Task [Chandrasekaran et al., 2019] as well as TREC-COVID [Wang and Lo, 2021; Voorhees et al., 2020] but these efforts either do not have information retrieval in mind or are focused on a specific topic such as the COVID-19 pandemic.

- Another use case is that of finding innovation and links between work. For instance, researchers linking two papers that were not linked before by citing them in a paper, possibly coming from different disciplines. This interdisciplinary linking allows for the association of ideas across disciplines. It also allows for the identification of science hotspots of science.

[4] See the SimpleText workshop at CLEF 2021 https://www.irit.fr/simpleText/

[5] This, of course, does also include issues around conflicting information and misinformation, as mentioned before.

With the adoption of open science, linking is needed between the various outputs of scientific studies: protocols, code, data, preprints, (open) peer reviews, publications, press articles, and reactions on social media which are all scattered on various platforms. The timely linking of COVID19 preprints to the subsequently published articles proved critical to revise the knowledge curated in the living systematic reviews used by health professionals during the pandemic [Cabanac et al., 2021b].

- An important aspect of scholarly work are manual *annotations*, for instance in the form of notes, margin comments or links [Agosti et al., 2004]. Annotation-based retrieval models have been proposed a while ago to satisfy specific information needs [Frommholz and Fuhr, 2006] but have not yet been explored as a means to address scholars' information needs or to foster scientific publishing based on annotations and open peer review. Furthermore, advances in tablet technology gave rise to interactive and multimodal tools such as, for instance, LiquidText[6], which provide a novel way to work with publications by means of annotations, going beyond paper. Again, the potential to utilise this data created and residing in personal libraries has not yet been explored.

- A more general question is about what can Scientometrics and Bibliometrics further contribute to IR? The Bibliometrics community has a long tradition when it comes to topics such as visualizations, exploratory data analysis, citation networks, and ranking institutions and authors. As much as the Bibliometrics community can learn from IR when it comes to, e.g., evaluation and advanced, interactive ranking methods, the IR community could benefit from the outcomes of Scientometric/Bibliometrics research to create better information services for scholars, students and decision-makers.

We can see from this discussion that Bibliometric-enhanced IR forms an exciting branch of Information Retrieval research. Its main differences compared to other IR directions are with respect to: *document structure* (scientific papers, patents, data/code repositories follow a certain pattern that can be exploited), the *specific information needs* and *relevance criteria* of scholars, students and policy-makers (who are often not scientists themselves) that require more expert search modes [Verberne et al., 2019], the *heterogeneous nature of the data at hand* (textual, multimedia, data sets, sensor data, user-generated annotations, bibliographic metadata, citation networks), the different forms of *visualizations* to present results, the need to ensure the *quality* of the results and the *human-computer interfaces* that are required to make large pools of data and documents accessible to users. Despite its differences, the discussed problems are typical IR research questions and as such interesting for the broader IR community.

The above discussion is based on BIR 2021, past BIR and BIRNDL workshops as well as the aftermath discussion of the BIR organisers and not meant to be complete.

4 Conclusion

In this report, we presented the 11th Bibliometric-enhanced Information Retrieval workshop that took place along ECIR 2021. We informed about accepted papers and keynote talks and discussed future directions in Bibliometric-enhanced IR. We interpret the high interest in the topic

[6] https://www.liquidtext.net/

from both communities and the identified research questions as a strong sign and motivation for further research. Indeed, Bibliometric-enhanced IR, which is different in its nature from other IR directions, is capable of tackling some pressing problems that have arisen from the need to make scholarly material more effectively accessible for different groups of interested users, such as scholars, students and decision-makers. There is still a lot both communities can and should learn from each other, which motivates future BIR workshops and closer collaboration between the communities.

References

Muhammad Kamran Abbasi and Ingo Frommholz. Cluster-based Polyrepresentation as Science Modelling Approach for Information Retrieval. *Scientometrics*, 102(3):2301–2322, 2015. doi: 10.1007/s11192-014-1478-1.

Ahmed Abura'Ed and Horacio Saggion. A select and rewrite approach to the generation of related work reports. In Cabanac et al. [2021a], pages 53–68. URL http://ceur-ws.org/Vol-2847/#paper-06.

Pablo Accuosto, Mariana Neves, and Horacio Saggion. Argumentation mining in scientific literature: From computational linguistics to biomedicine. In Cabanac et al. [2021a], pages 20–36. URL http://ceur-ws.org/Vol-2847/#paper-03.

Maristella Agosti, Nicola Ferro, Ingo Frommholz, and Ulrich Thiel. Annotations in Digital Libraries and Collaboratories – Facets, Models and Usage. In Rachel Heery and Liz Lyon, editors, *Research and Advanced Technology for Digital Libraries. Proc. European Conference on Digital Libraries (ECDL 2004)*, Lecture Notes in Computer Science, pages 244–255, Heidelberg et al., 2004. Springer.

Daria Alexander and Arjen P. de Vries. "This research is funded by...": Named Entity Recognition of Financial Information in Research Papers. In Cabanac et al. [2021a], pages 102–110. URL http://ceur-ws.org/Vol-2847/#paper-10.

Frederique Bordignon, Liana Ermakova, and Marianne Noel. Preprint Abstracts in Times of Crisis: a Comparative Study with the Pre-pandemic Period. In Cabanac et al. [2021a], pages 37–44. URL http://ceur-ws.org/Vol-2847/#paper-04.

Guillaume Cabanac, Ingo Frommholz, and Philipp Mayr. Bibliometric-enhanced information retrieval 10th anniversary workshop edition. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 641–647, Cham, 2020. Springer International Publishing. URL https://doi.org/10.1007/978-3-030-45442-5_85.

Guillaume Cabanac, Ingo Frommholz, Philipp Mayr, and Suzan Verberne, editors. *Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval (BIR)*, number 2847 in CEUR Workshop Proceedings, Aachen, 2021a. URL http://ceur-ws.org/Vol-2847/.

Guillaume Cabanac, Theodora Oikonomidi, and Isabelle Boutron. Day-to-day discovery of preprint–publication links. *Scientometrics*, 126(6):5285–5304, 2021b. doi: 10.1007/s11192-021-03900-7.

Manajit Chakraborty, David Zimmermann, and Fabio Crestani. Patentquest: A user-oriented tool for integrated patent search. In Cabanac et al. [2021a], pages 89–101. URL http://ceur-ws.org/Vol-2847/#paper-09.

Muthu Kumar Chandrasekaran, Michihiro Yasunaga, Dragomir Radev, Dayne Freitag, and Min-Yen Kan. Overview and results: Cl-scisumm shared task 2019, 2019.

Ingo Frommholz and Norbert Fuhr. Probabilistic, Object-oriented Logics for Annotation-based Retrieval in Digital Libraries. In Michael Nelson, Cathy Marshall, and Gary Marchionini, editors, *Proc. of the 6th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2006)*, pages 55–64, New York, 2006. ACM. ISBN 1595933549. doi: 10.1145/1141753.1141764.

Jochen Gläser, Wolfgang Glänzel, and Andrea Scharnhorst. Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, 111(2):981–998, 2017. doi: 10.1007/s11192-017-2296-z.

Marti A. Hearst. *Search User Interfaces*. Cambridge University Press, 2010. ISBN 9780521113793. URL https://searchuserinterfaces.com/book/.

Laura Koesten, Philipp Mayr, Paul Groth, Elena Simperl, and Maarten de Rijke. Report on the DATA:SEARCH'18 workshop - Searching Data on the Web. *SIGIR Forum*, 52(2):117–124, 2018. URL http://sigir.org/wp-content/uploads/2019/01/p117.pdf. Editorial.

Hiran H. Lathabai, Abhirup Nandy, and Vivek Kumar Singh. Expertise based institutional recommendation in different thematic areas. In Cabanac et al. [2021a], pages 45–52. URL http://ceur-ws.org/Vol-2847/#paper-05.

Haiming Liu, Paul Mulholland, Dawei Song, Victoria Uren, and Stefan Rüger. Applying information foraging theory to understand user interaction with content-based image retrieval. In *Proceeding of the third symposium on Information Interaction in Context - IIiX '10*, pages 135–144, New York, New York, USA, 2010. ACM Press. ISBN 9781450302470. doi: 10.1145/1840784.1840805. URL http://portal.acm.org/citation.cfm?doid=1840784.1840805.

Marianne Lykke, Birger Larsen, Haakon Lund, and Peter Ingwersen. Developing a Test Collection for the Evaluation of Integrated Search. In *Proceedings ECIR 2010*, pages 627–630, 2010.

David Maxwell and Leif Azzopardi. Information scent, searching and stopping: Modelling SERP level stopping behaviour. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10772 LNCS(i):210–222, 2018. ISSN 16113349. doi: 10.1007/978-3-319-76941-7_16.

Jacqueline Sachse. Bibliometric indicators and relevance criteria – an online experiment. In Cabanac et al. [2021a], pages 69–77. URL http://ceur-ws.org/Vol-2847/#paper-07.

Houcemeddine Turki, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Grischa Fraumann, Christian Hauschke, and Lambert Heller. Enhancing knowledge graph extraction and validation from scholarly publications using bibliographic metadata. *Frontiers in Research Metrics and Analytics*, 6:36, 2021. ISSN 2504-0537. doi: 10.3389/frma.2021.694307. URL https://www.frontiersin.org/article/10.3389/frma.2021.694307.

S. Verberne, J. He, U. Kruschwitz, G. Wiggers, B. Larsen, T. Russell-Rose, and A. P. de Vries. First international workshop on professional search. *ACM SIGIR Forum*, 52(2):153–162, 2019.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. *SIGIR Forum*, 54(1):1–12, 2020. URL http://arxiv.org/abs/2005.04474.

Ken Voskuil and Suzan Verberne. Improving Reference Mining in Patents with BERT. In Cabanac et al. [2021a], pages 78–88. URL http://ceur-ws.org/Vol-2847/#paper-08.

Lucy Lu Wang and Kyle Lo. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics*, 22(2):781–799, 2021. ISSN 14774054. doi: 10.1093/bib/bbaa296.

Howard D. White and Katherine W. McCain. Visualizing a discipline: An author co-citation analysis of Information Science, 1972–1995. *Journal of the American Society for Information Science*, 49(4):327–355, 1998. doi: b57vc7.

Shintaro Yamamoto, Anne Lauscher, Simone Paolo Ponzetto, Goran Glavaš, and Shigeo Morishima. Self-supervised learning for visual summary identification in scientific publications. In Cabanac et al. [2021a], pages 5–19. URL http://ceur-ws.org/Vol-2847/#paper-02.

Yinan Zhang and Chengxiang Zhai. Information Retrieval as Card Playing : A Formal Model for Optimizing Interactive Retrieval Interface. In *Proceedings SIGIR 2015*, pages 685–694. ACM, 2015. ISBN 9781450336215.