Scientometrics manuscript No.
(will be inserted by the editor)

Scholarly literature mining with Information Retrieval and Natural Language Processing: Preface

Guillaume Cabanac · Ingo Frommholz · Philipp Mayr

Received: October 9, 2020

1 Introduction

This special issue features the work of authors originally coming from different communities: bibliometrics/scientometrics (SCIM), information retrieval (IR) and, as an emerging player gaining more relevance for both aforementioned fields, natural language processing (NLP). The work presented in their papers combine ideas from all these fields, having in common that they all are using the scholarly data well known in scientometrics and solving problems typical to scientometric research. They model and mine citations, as well as metadata of bibliographic records (authorships, titles, abstracts sometimes), which is common practice in SCIM. They also mine and process fulltexts (including in-text references and equations) which is common practice in IR and requires established NLP text mining techniques. IR collections are utilised to ensure reproducible evaluations; creating and sharing test collections in evaluation initiatives such as CLEF eHealth[1] is common IR tradition that is also prominent in NLP, e.g., by the CL-SciSumm shared task.[2]

From an IR perspective, surprisingly, scholarly information retrieval and recommendation, though gaining momentum, have not always been the focus of research in the past. Besides operating on a rich set of data for researchers in all three disciplines to play with, scholarly search poses challenges in particular for

G. Cabanac
University of Toulouse, Computer Science Department, IRIT UMR 5505 CNRS,
118 route de Narbonne, F-31062 Toulouse cedex 9, France
E-mail: guillaume.cabanac@univ-tlse3.fr, ORCID: 0000-0003-3060-6241

Ingo Frommholz
University of Bedfordshire
Luton LU1 3JU, UK
E-mail: ifrommholz@acm.org, ORCID: 0000-0002-5622-5132

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany
E-mail: philipp.mayr@gesis.org, ORCID: 0000-0002-6656-1658

[1] https://clefehealth.imag.fr
[2] https://github.com/WING-NUS/scisumm-corpus

IR due to the complex information needs that require different approaches than known from, e.g., Web search, where information needs are simpler in many cases. As an example, the current COVID-19 crisis shows that hybrid SCIM/IR/NLP approaches are increasingly required to ensure researchers get access to important relevant and high-quality information, often only available on preprint servers, in a short period of time [2,5,7,10]. These kinds of complex information needs pose challenges which have been recognised by the Information Retrieval community that quickly launched the TREC-COVID initiative run by NIST [11], demonstrating the timeliness of our endeavour and this special issue. Working on scholarly material thus has incentives for researchers in Information Retrieval but we believe the challenges can only be tackled effectively by all three communities as a whole. The NLP community has initiated a similar activity with a dedicated workshop series NLP COVID-19 Workshop[3] which is running at major NLP conferences (ACL & EMNLP) in 2020.

With the surge of "scholarly big data" [6], Bibliometrics and Information Retrieval in combination with NLP methods have seen a recent renaissance that resulted in a series of special issues:

- "Combining Bibliometrics and Information Retrieval" [9] in *Scientometrics* (2015).
- "Bibliometric-enhanced Information Retrieval" [3] in *Scientometrics* (2018).
- "Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries" [8] in *International Journal on Digital Libraries* (2018).
- "Mining Scientific Papers: NLP-enhanced Bibliometrics" [1] in *Frontiers in Research Metrics and Analytics* (2019).

2 Special issue papers

This special issue on "Scholarly literature mining with Information Retrieval and Natural Language Processing" presents works intersecting Bibliometrics and Information Retrieval, utilising Natural Language Processing (NLP). The special issue was announced *via* an open call for papers[4]. In response to the CFP, we received 24 submissions which were reviewed by two to three reviewers (for overlapping papers, e.g., IR and NLP, we selected reviewers from both domains). Eventually, the guest editors accepted 14 papers. Nine papers have been rejected and one paper was withdrawn by the authors during the reviewing rounds.

In the following we provide an overview of the 14 papers organised into 3 clusters. We introduce the paper ordering of the special issue in Tables 1–2. To generate a lightweight overview of the variety of the papers we identified the research *Tasks* and *Area of Application*, the used *Corpus*, *Objects*, and *Methods* of each contribution.

The papers in this special issue appear in the following sequence. We decided to start with a set of more classical papers featuring scientometric methods like network analysis and bibliographic data from the Web of Science, Scopus or similar resources. The second set of papers is more IR oriented: papers mine fulltexts

[3] https://www.nlpcovid19workshop.org/acl2020/

[4] https://sites.google.com/view/scientometrics-si2019-bir

and they use techniques like embeddings and neural networks. The third cluster of papers contains NLP-oriented papers which are, for instance, specialised in summarisation and utilise scholarly documents.

Cluster 1. **SCIM with IR and NLP**:

- Lietz: *Drawing impossible boundaries: field delineation of Social Network Science.*
- Schneider et al.: *Continued post-retraction citation of a fraudulent clinical trial report, eleven years after it was retracted for falsifying data.*
- Kreutz et al.: *Evaluating semantometrics from computer science publications.*
- Haunschild & Marx: *Discovering seminal works with marker papers.*
- Lamirel et al.: *An overview of the history of Science of Science in China based on the use of bibliographic and citation data: a new method of analysis based on clustering with feature maximization and contrast graphs.*

Cluster 2. **IR and Text-mining of scholarly literature**:

- Nogueira et al.: *Navigation-based candidate expansion and pretrained language models for citation recommendation.*
- Greiner-Petter et al.: *Math-word embedding in math search and semantic extraction.*
- Carvallo et al.: *Automatic document screening of medical literature using word and text embeddings in an active learning setting.*
- Saier & Färber: *unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata.*

Cluster 3. **NLP-oriented papers on scholarly literature**:

- Zerva et al.: *Cited text span identification for scientific summarisation using pretrained encoders.*
- La Quatra et al.: *Exploiting pivot words to classify and summarize discourse facets of scientific papers.*
- AbuRa'ed et al.: *Automatic related work section generation: experiments in scientific document abstracting.*
- Jimenez et al.: *Automatic prediction of citability of scientific articles by stylometry of their titles and abstracts.*
- Portenoy & West: *Constructing and evaluating automated literature review systems.*

We hope the selection of papers in this special issue will be interesting and enjoyable for researchers coming from all relevant fields and provides a starting point for future explorations in the field.

Table 1 Overview of the articles in this special issue.

Task	Area of Application	Corpus	Objects	Methods
		Lietz		
field delineation	social network science	Web of Science	metadata (title, abstract, keywords), references	clustering, network analysis
		Schneider, Ye, Hill, & Whitehorn		
analysing citing papers of a retracted study	clinical science	Google Scholar, Web of Science	seed paper, citations, retraction notices	network analysis, citation context analysis, retraction status visibility analysis
		Kreutz, Sahitaj, & Schenkel		
spotting seminal work; classifying papers	computer science	DBLP	fulltext	classification using words, semantics, topics and publication years
		Haunschild & Marx		
spotting seminal work	physics	Microsoft Academic, Web of Science	references, time	reference publication year spectroscopy
		Lamirel, Chen, Cuxac, Al Shehabi, Dugué & Liu		
mapping the evolution of a country's scientific production	Science in China	China National Knowledge Infrastructure database	metadata (title, abstract, authors), dictionary of Chinese names	clustering, topic modelling, network analysis
		Nogueira, Jiang, Cho, & Lin		
ranking citation recommendations	computer science, biomedicine	DBLP, Open Research, PubMed	fulltext	document ranking model, embeddings
		Greiner-Petter, Youssef, Ruas, Miller, Schubotz, Aizawa & Gipp		
discovering mathematical term similarity and analogy and query expansions	mathematics	arXiv	fulltext	embeddings

Table 2 Overview of the articles in this Special Issue (continued).

Task	Area of Application	Corpus	Objects	Methods
		Carvallo, Parra, Lobel, & Soto		
paper screening for evidence-based medicine	medicine	CLEF eHealth, Episte-monekos	fulltext	document ranking model, query expansion, embeddings
		Saier & Färber		
dataset creation	fields of arXiv preprints	arXiv, Microsoft Academic Graph	fulltext, in-text citations, linked data	data integration, descriptive statistics
		Zerva, Nghiem, Nguyen, & Ananiadou		
paper summarization (from citations)	natural language processing	CL-SciSumm	fulltext, in-text citations	neural networks
		La Quatra, Cagliero, & Baralis		
discourse facet summarization	natural language processing	CL-SciSumm	fulltext, in-text citations	neural networks
		AbuRa'ed, Saggion, Shvets, & Bravo		
citation sentence production	text summarization	ScisummNet, Open Academic Graph, Microsoft Academic Graph, RWSData	fulltext	neural networks
		Jimenez, Avila, Dueñas, & Gelbukh		
citation forecasting	the scientific literature	Scopus	metadata (title + abstract)	statistics, stylometry
		Portenoy & West		
generation of a literature review of a field	community detection in graphs, misinformation studies, science communication	Web of Science	references, paper titles	text similarity, supervised learning, embeddings

Acknowledgements We wish to thank all contributors to this special issue: The researchers who submitted papers, the many reviewers who generously offered their time and expertise, and the participants of the BIR and BIRNDL workshops [4].[5]

References

1. Atanassova, I., Bertin, M., Mayr, P.: Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics. Frontiers in Research Metrics and Analytics **4**(2) (2019). 10.3389/frma.2019.00002
2. Brainard, J.: New tools aim to tame pandemic paper tsunami. Science **368**(6494), 924–925 (2020). 10.1126/science.368.6494.924
3. Cabanac, G., Frommholz, I., Mayr, P.: Bibliometric-enhanced information retrieval: preface. Scientometrics **116**(2), 1225–1227 (2018). 10.1007/s11192-018-2861-0
4. Cabanac, G., Frommholz, I., Mayr, P.: Bibliometric-Enhanced Information Retrieval 10th Anniversary Workshop Edition. In: J.M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M.J. Silva, F. Martins (eds.) Advances in Information Retrieval, *LNCS*, vol. 12036, pp. 641–647. Springer International Publishing (2020). 10.1007/978-3-030-45442-5_85
5. Fraser, N., Brierley, L., Dey, G., Polka, J.K., Pálfy, M., Nanni, F., Coates, J.A.: Preprinting the COVID-19 pandemic. bioRxiv (2020). 10.1101/2020.05.22.111294
6. Giles, C.L.: Scholarly big data. In: CIKM'13: Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management, p. 1. ACM, New York, NY (2013). 10.1145/2505515.2527109
7. Kwon, D.: How swamped preprint servers are blocking bad coronavirus research. Nature **581**(7807), 130–131 (2020). 10.1038/d41586-020-01394-6
8. Mayr, P., Frommholz, I., Cabanac, G., Chandrasekaran, M.K., Jaidka, K., Kan, M.Y., Wolfram, D.: Introduction to the Special Issue on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). International Journal on Digital Libraries **19**(2–3), 107–111 (2018). 10.1007/s00799-017-0230-x
9. Mayr, P., Scharnhorst, A.: Combining bibliometrics and information retrieval: preface. Scientometrics **102**(3), 2191–2192 (2015). 10.1007/s11192-015-1529-2
10. Palayew, A., Norgaard, O., Safreed-Harmon, K., Andersen, T.H., Rasmussen, L.N., Lazarus, J.V.: Pandemic publishing poses a new COVID-19 challenge [Comment]. Nature Human Behaviour **4**(7), 666–669 (2020). 10.1038/s41562-020-0911-0
11. Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E., Wang, L.L., Hersh, W.R.: TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19. Journal of the American Medical Informatics Association **27**(9), 1431–1436 (2020). 10.1093/jamia/ocaa091

[5] Since 2016 we maintain the "Bibliometric-enhanced-IR Bibliography" https://github.com/PhilippMayr/Bibliometric-enhanced-IR_Bibliography/ that collects scientific papers which appear in collaboration with the BIR/BIRNDL organizers.