Extending a Research-Paper Recommendation System with Scientometric Measures

Sophie Siebert[1], Siddarth Dinesh[2], and Stefan Feyer[3]

[1] Otto-von-Guericke Unitversity, Magdeburg, Germany,
sophie.siebert@st.ovgu.de,
[2] Birla Institute of Technology and Science, Goa 403726, India
f2012519@goa.bits-pilani.ac.in
[3] University of Konstanz, Germany
stefan@feyer.de

Abstract. **Keywords:** research paper recommender syste, scientometric, bibliometric, altmetric, citations, readership
In recent years the number of academic publication increased strongly. As this information flood grows, it becomes more difficult for researches to find relevant literature effectively. To overcome this difficulty one can use recommendation systems, which often use text similarity to find related documents. To improve those systems we add scientometrics as a ranking measure for popularity into these algorithms. In this paper we analyse whether and how scientometrics are useful in a recommender system.

1 Introduction

The number of academic publications doubles aproximately every ten years [6]. As a result it becomes more difficult for researchers to find relevant literature. It is nearly impossible to read all literature of an academic field to find the most important and relevant documents. Even if one has profound knowledge about his academic field it is difficult to follow the latest news and filter them for relevance [13].

To handle this information flood, recommender systems come into account. They identify the informational needs of researchers and recommend the best fitting literature. Unfortunately, neither the automatic identification of the informational needs, nor the search for relevant literature are trivial tasks. The extent of this challenge can be derived from the amount of research in this area. In the last sixteen years 90 methods were developed and investigated by around 300 academic researches and published in over 200 publications [2].

Since it is important to recommend papers which are relevant, it is necessary to improve recommender systems. In this paper we focus on the use of scientometrics to rank the recommendations.

Scientometrics are introduced as a 'quantitative study of science and technology' [11]. They are most generally classified into bibliometrics and altmetrics. Bibliometrics are a tool for the 'measurement of texts and information' [8]. The term is often used to describe the statistical analysis based on citation data.

There are many ways to use citation data to calculate metrics for the popularity of a paper, author or journal. A list of 108 bibliometrics is published by Wildgaard and Schneider [16], including normalizations, h-index and many others.

On the contrary, altmetrics, which takes its name from alternative metrics, use information from social media, blogs etc. [7]. The count of readers is also an altmetric and correlate with bibliometrics. [15].

Until now recommendation systems rank their recommendation only with respect to content similarity. The assumption in this paper is that a paper with a good reputation is more worth reading, thus should be recommended. To measure the reputation we will use scientometrics. The scientific question is which scientometrics and their combination with the similarity ranking is most liked by the user and thus the best. To measure how much users like a recommendation the Click-Through-Rate (**CTR**) will be analyzed.

2 Related Work

2.1 How to rank papers

Papers can be ranked by different criteria. Lewandowski and Behnert [4] divided these criteria in six fields: 'text statistics', 'popularity', 'freshness', 'locality and availability', 'content properties' and the 'user's background'.

The text statistics can describe how similar two documents are based on the content. A famous measure is for example TF-IDF. Text statistics also can focus on the document length or anchor text and emphasized text. The popularity pays respect to the usage of a document, how often is it read, cited, downloaded or bought, and how good are they rated. The freshness ensures that one gets recent documents. The locality and availability considers, if the user currently has access to the document, since he would not want to get recommendations he can not read. It also considers costs, since free documents are easier to access. Also one can use content properties to rank papers, which includes the file format, the language and the amount of meta data. People like to read their own language and use common file formats like PDF. The last category is the user's background, where the user is analyzed for example which academic background he has and which field of study he belongs to, as a computer scientist most likely will not read a document about history. [4]

In this study we focus on the popularity category to rank papers. We use scientometrics which we assume indicates the popularity and rank the paper accordingly.

2.2 Scientometrics

Scientometrics are used to calculate the reputation of journals, authors, institutions or papers. After that people can for example identify authors with high or low reputation. Scientometrics can help to identify patterns which are highly cited [8] and thus can be used to optimize further citations.

It is also common to rank results in search engines based on their popularity. This concept is for example used in Google's Page Rank Algorithm, where the popularity is expressed by hyperlinks [12]. Scientometrics are also used in search engines for research paper [14]. Bethard and Jurafsky added citations, h-index and recency into their search engine for research paper. They found pure citations improve the ranking, but h-index and recency degrade it. [5]. In our work we will use scientometrics to measure popularity and rerank results in a research-paper recommendation system.

3 Methods

3.1 Our system and data

Mr. DLib (Machine Readable Digital Library, http://mr-dlib.org/) is a research-paper recommender system and for academic purposes only [1,3]. It recommends papers similar to a given input paper. Mr. DLib has a database where the data is gathered and indexed using Solr. We cooperate with GESIS, who provides a framework for the user, as well as our data [10] The data consists of 9.5 million documents, from which 5.3 million are english and 2 million are german. As soon as a document is requested in GESIS, they are requesting recommendations from Mr. DLib, which are displayed at their website as shown in figure 1. Everytime a user clicks on a recommendation he is forwarded from Mr. DLib to GESIS, while Mr. DLib logs the click. With this procedure we can measure the Click-Through-Rate (**CTR**).

$$CTR = \frac{\text{Number of clicked recommendations}}{\text{Number of delivered recommendations}}$$

We gathered data from 17th October 2016 to 8th February 2017. Currently we display six recommendations. All in all we analyse 38,740,893 recommendations with 53,441 clicks, which corresponds to a CTR of 0.138%.

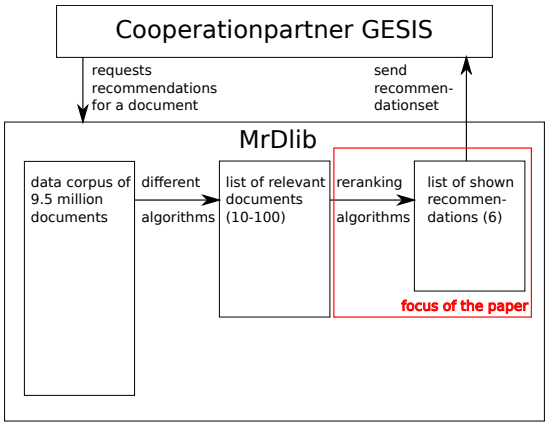

Fig. 1. Simplified visualisation of Mr. DLib and its components

3.2 Algorithms and ranking approaches

Figure 1 shows how the recommendations are generated. First as relevance algorithm is executed, providing us with a list of at most 100 documents, where each document has an attached relevance score. After this we choose which reranking algorithm is executed. It chooses randomly from the following variables: 1) how many documents are considered for the reranking, 2) how much influence the relevant score has and 3) which bibliometric scheme is applied.

There are currently five different algorithms, which creates a list of documents to be reranked. Out of this five algorithms only two provide a relevance score we can usefully combine with the bibliometrics. Since these two algorithm are more interesting, our random generator picks them more often. Roughly 90% of the recommendations are produced by them.

Since it was already shown that readership usually correlate with citations [15], one can use readership data instead of citation as a base for all presented scientometrics. In this paper we use readership data from Mendeley to rank the documents. We got readership data for 1.694.373 documents, which is a coverage of 17.82%. Currently we use the absolute count of readers, the count of readers normalized by the age of the paper, and the count of readers normalized by the number of authors.

4 Results

In this section we present our test results of the reranking approaches. They are split with respect to the three different attributes. All data is available at http://datasets.mr-dlib.org/.

4.1 Reranking method

The reranking method describes how the scientometric data and the text relevance score are combined with each other. We ranked the recommendation considering text relevance (**TR**) only or scientometrics only as well as with different weightings. While the Text Relevance is within a certain value range set by Solr, the scientometric values can range from 0 to several 1000. To balance the impact of the text relevance and the scientometric, the scientometric is set into logarithm and root. Each combination was sorted ascendingly and descendingly to ensure there is actually a difference in sorting them according to the metrics.

As one can see in figure 2, the ascendingly ordered scores are always lower than the descendingly sorted. This was expected since we assume that scientometrics are a measure for popularity and a popular publication has higher quality and more interesting findings thus is more worth reading. Also the ranking with the scientometric scores higher than the rext relevance only approach. One weird occasion is the good performance of the scientometric only ascending ordering. Its performance was nearly identical to the text relevance only descending ordering.

The statistically significance is given for the associated descending and ascending orders, except for the scientometric only. Unfortunately there is no statistic significance between the different descending orders.

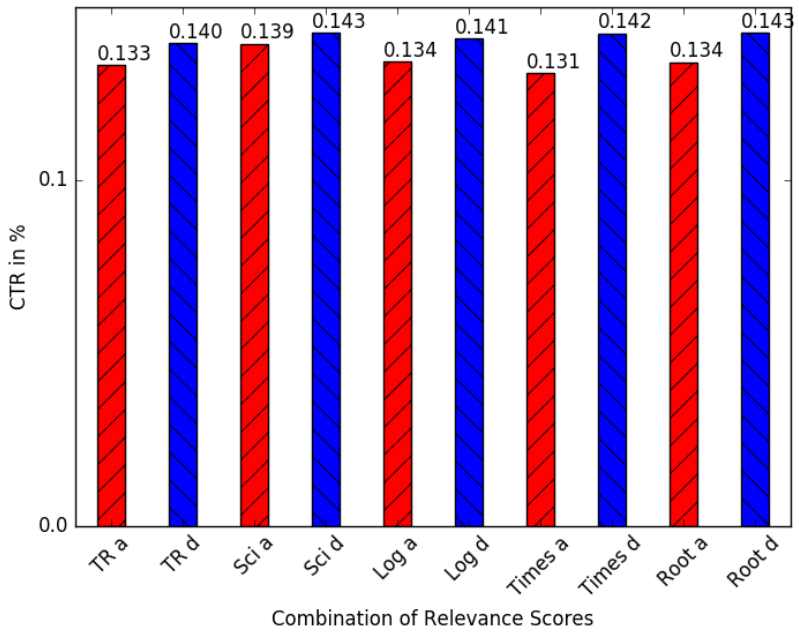

Fig. 2. Visualisation of the Click-Through-Rate (CTR) for different weightings of text relevance and scientometrics. The red bars as well as the letter 'a' indicates ascending order. The blue bars as well as the letter d indicates descending order. TR = Text Relevance, Sci = scientometric, Log = TR * log(Sci), Times = TR * Sci, Root = TR * root(Sci)

4.2 Metrics

By metrics, we refer to the scientometric indicator that we used for ranking recommendations. We analyzed absolute readership count R_{count}, readership count normalized by the age of the paper in years R_{age} and readership count normalized by the number of authors R_{auth}. The formula for readership normalized by the age of the paper R_{age} is

$$R_{age} = \frac{R_{count}}{Year_{now} - Year_{published} + 1}$$

This is expected to show better performance, since good papers need time to become famous.

The formular for readership normalized by the number of authors R_{auth} is

$$R_{auth} = \frac{R_{count}}{\#authors}$$

This normalization is expected to show also better performance. Paper with many authors are likely to be more famous, because of they are wider spread.

We considered in our experiments primarily those metrics which were easy to calculate while incorporating as many different aspects of the scientometrics in the literature review. Another important criteria was that the metrics would only use the data that we had. By evaluating this subset of metrics, we would have an idea of which groups of metrics are performing the best in a real-world setting. With this information we can focus on this group in the future.

Surprisingly the absolute readership count R_{count} scores highest. All of this results are statistically significant.

The discrepancy of the CTR to the reranking method results from not including text relevance only data.

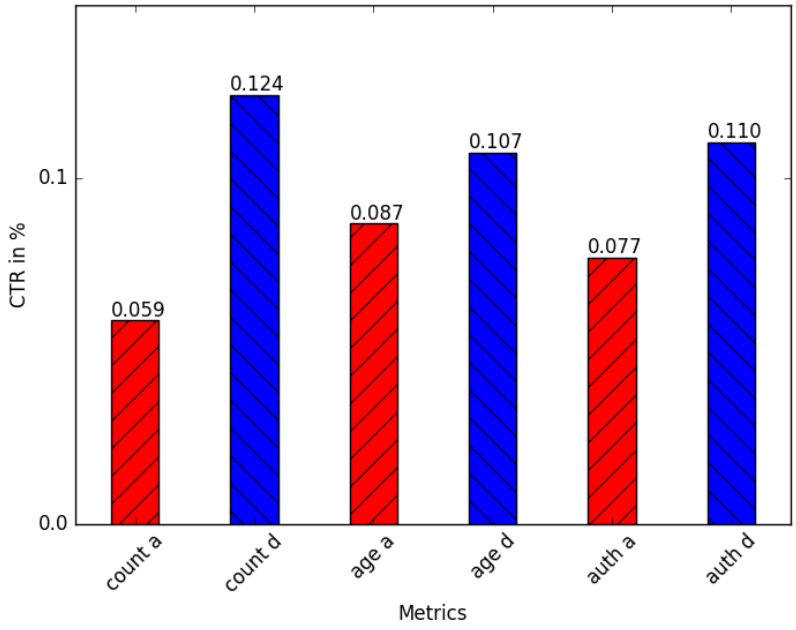

Fig. 3. Visualisation of the Click-Through-Rate (CTR) for different scientometric indicators. The red bars as well as the letter 'a' indicates ascending order. The blue bars as well as the letter d indicates descending order. count = absolute count of readership, age = normalized by the age of the paper, auth = normalized by the number of authors

Reranked candidates The reranked candidates are the number of candidates which will be picked from an algorithm to rerank them according to the ranking parameters. The more candidates are considered for a reranking, the less the text relevance is considered. It seems that more candidates leads to a better performance, although a medium sized list decreased it. The statistically significance is given for the associated descending and ascending orders, except for 30 and 100 candidates. Between the different descending orders, roughly half of the combinations are statistically significant.

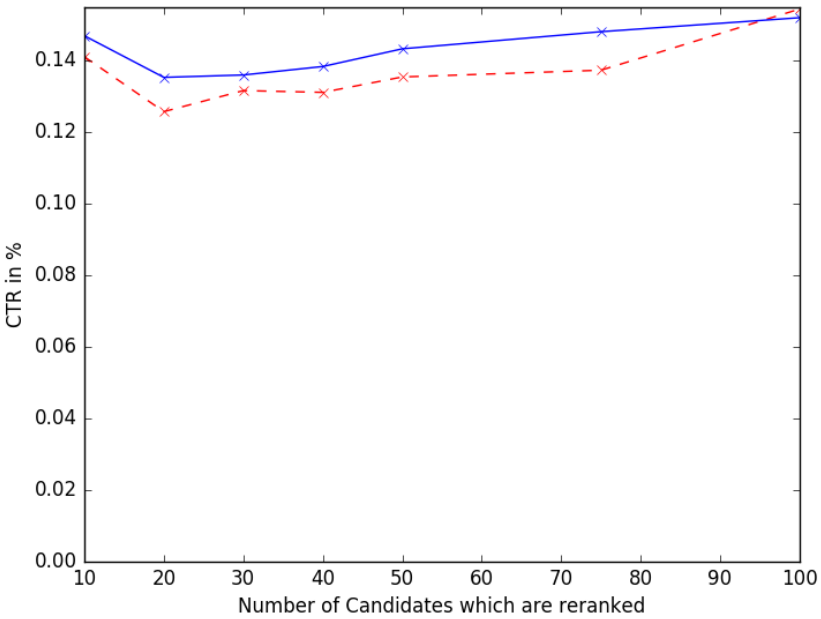

Fig. 4. Visualisation of the Click-Through-Rate (CTR) of the number of recieved candidates from the algorithm. The red lines indicates ascenging order. The blue line indicates descending order.

5 Conclusion and Future Work

In this paper we evaluated different reranking approaches to see whether and how scientometrics can improve academic paper recommendations systems. With our current data we can conclude that scientometrics do improve the ranking of documents in a recommendation system compared to a text relevance only approach. This is shown in 2, where the CTR of the descending rankings with a scientometric scores higher than the text relevance only approach. However this

improvement is rather small. Furthermore smaller list of reranking candidates seem to lead to a better CTR. The metric which achieved the best score is the absolute count without normalization.

However the good scoring of the scientometrics only ascending was not expected. This might be due to the low coverage of 17.82% of the scientometric data. If too many documents of the pre-generated list do not have associated readership data, the rerank will not effect the sorting. Thus descending and ascending order will have the same sorting and same performance.

To achieve a better coverage we will calculate author metrics and calculate them back to the papers by building a sum or the average. This will lead to a coverage of 46,27%. Furthermore there will be a fallback mechanic, which will choose a higher coverage metric, if the current metric has to less data.

Another reason might be the style of evaluation. When the recommendations are displayed only title and author are shown. The user decides only from the content, if the recommendation might be useful. However our approach takes popularity in account, which is an assumption for quality. To overcome this issue we will log if the user used the document after taking a look at it. For example if he clicked several links like cite, export, favorite or search. This measure might be more fitting for evaluation a popularity approach.

Another next step is the gathering and calculation of citation metrics. Furthermore we will soon be able evaluate the scientometrics in JabRef and collect more data [9]. In addition the scientometric rankings can also be evaluated together with the different algorithms, to find out if they are stable and which different combinations of algorithm and ranking approaches work best together.

6 Acknowledgements

This work was supported by a fellowship within the FITweltweit programme of the German Academic Exchange Service (DAAD). Moreover we want to thank Joeran Beel, Martin Glauer and Christoph Doell for the support.

References

1. Beel, J. ; Gipp, B. ; Aizawa, A. : Mr. DLib: Recommendations-as-a-Service (RaaS) for Academia. In: *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2017
2. Beel, J. ; Gipp, B. ; Langer, S. ; Breitinger, C. : Research Paper Recommender Systems: A Literature Survey. In: *International Journal on Digital Libraries* (2015), S. 1–34. `http://dx.doi.org/10.1007/s00799-015-0156-0`. – DOI 10.1007/s00799–015–0156–0. – ISSN 1432–5012
3. Beel, J. ; Gipp, B. ; Langer, S. ; Genzmehr, M. ; Wilde, E. ; Nrnberger, A. ; Pitman, J. : Introducing Mr. DLib, a Machine-readable Digital Library. In: *Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL'11)*, ACM, 2011, S. 463–464. – Available at http://docear.org

4. Behnert, C. ; Lewandowski, D. : Ranking search results in library information systems-considering ranking approaches adapted from web search engines. In: *The Journal of Academic Librarianship* 41 (2015), Nr. 6, S. 725–735
5. Bethard, S. ; Jurafsky, D. : Who should I cite: learning literature search models from citation behavior. In: *Proceedings of the 19th ACM international conference on Information and knowledge management* ACM, 2010, S. 609–618
6. Bornmann, L. ; Mutz, R. : Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. In: *Journal of the Association for Information Science and Technology* 66 (2015), Nr. 11, S. 2215–2222
7. Brigham, T. J.: An introduction to altmetrics. In: *Medical reference services quarterly* 33 (2014), Nr. 4, S. 438–447
8. Daim, T. U. ; Rueda, G. ; Martin, H. ; Gerdsri, P. : Forecasting emerging technologies: Use of bibliometrics and patent analysis. In: *Technological Forecasting and Social Change* 73 (2006), Nr. 8, S. 981–1012
9. Feyer, S. ; Siebert, S. ; Gipp, B. ; Aizawa, A. ; Beel, J. : Integration of the Scientific Recommender System Mr. DLib into the Reference Manager JabRef. In: *Proceedings of the 39th European Conference on Information Retrieval (ECIR)*, 2017
10. Hienert, D. ; Sawitzki, F. ; Mayr, P. : Digital Library Research in Action Supporting Information Retrieval in Sowiport. In: *D-Lib Magazine* 21 (2015), Nr. 3/4. `http://dx.doi.org/10.1045/march2015-hienert`. – DOI 10.1045/march2015–hienert
11. Hood, W. ; Wilson, C. : The literature of bibliometrics, scientometrics, and informetrics. In: *Scientometrics* 52 (2001), Nr. 2, S. 291–314
12. Langville, A. N. ; Meyer, C. D.: *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011
13. Naisbitt, J. : *Megatrends*. Warner Books, 1988
14. Sugiyama, K. ; Kan, M.-Y. : Scholarly paper recommendation via user's recent research interests. In: *Proceedings of the 10th annual joint conference on Digital libraries* ACM, 2010, S. 29–38
15. Thelwall, M. : Why do papers have many Mendeley readers but few Scopus-indexed citations and vice versa? In: *Journal of Librarianship and Information Science* (2015), S. 0961000615594867
16. Wildgaard, L. ; Schneider, J. W. ; Larsen, B. : A review of the characteristics of 108 author-level bibliometric indicators. In: *Scientometrics* 101 (2014), Nr. 1, S. 125–158