

Relations Between Relevance Assessments, Bibliometrics and Altmetrics

Timo Breuer¹[0000-0002-1765-2449]*, Philipp Schaer¹[0000-0002-8817-4632]*, and
Dirk Tunger^{1,2}[0000-0001-6383-9194]*

¹ TH Köln (University of Applied Sciences), 50678 Cologne, Germany
firstname.lastname@th-koeln.de

² Forschungszentrum Jülich, 52425 Jülich, Germany
d.tunger@fz-juelich.de

Abstract. Relevance assessment in retrieval test collections and citations/mentions of scientific documents are two different forms of relevance decisions: direct and indirect. To investigate these relations, we combine arXiv data with Web of Science and Altmetrics data. In this new collection, we assess the effect of relevance ratings on measured perception in the form of citations or mentions, likes, tweets, et cetera. The impact of our work is that we could show a relation between direct relevance assessments and indirect relevance signals.

Keywords: Relevance assessments · bibliometrics · Altmetrics · citations · information retrieval · test collections.

1 Introduction

One of the long-running open questions in Information Science in general and especially in Information Retrieval (IR) is on what constitutes relevance and relevance decisions. In this paper, we would like to borrow from the idea of using IR test collections and their relevance assessments to intersect these explicit relevance decisions with some implicit or hidden relevance decisions in the form of citations. We see this in the light of Borlund's discussion of relevance and its multidimensionality [2]. On the one hand, we have the test collection's relevance assessments that are *direct relevance decisions* and are always based on a concrete topic and the corresponding information need of an assessor [19]. On the other hand, the citation data gives us a hint on a distant or *indirect relevance decision* from external users. These external users are not part of the design process of the test collections, and we do not know anything about their information need or retrieval context. We only know that they cited a specific paper - therefore, this paper was somehow relevant to them. Otherwise, they would not have cited it.

* Listed in alphabetical order. Data and sources are available at Zenodo [3].

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). BIR 2020, 14 April 2020, Lisbon, Portugal.

A test collection that incorporates both direct and indirect relevance decisions is the iSearch collection introduced by Lykke et al. [15]. One of the main advantages of iSearch is the combination of a classic document collection derived from the arXiv, a set of topics that describe a specific information need plus the related context, relevance assessments, and a complementing set of references and citation information.

Carevic and Schaer [4] previously analyzed the iSearch collection to learn about the connection between topical relevance and citations. Their experiments showed that internal references within the iSearch collection did not retrieve enough relevant documents when using a co-citation-based approach. Only very few topics retrieved a high number of potentially relevant documents. This might be due to the preprint characteristics of the arXiv, where typically, a citation would target a journal publication and not the preprint. This information on external citations is not available within iSearch.

To improve on the known limitations of having a small overlap of citations and relevance judgments in iSearch, we expand the iSearch document collection and its internal citation data. We complement iSearch with external citation data from the Web of Science. Additionally, we add different Altmetric scores as they might introduce some other promising insights on relevance indicators. These different data sources will be used to generate a dataset to investigate whether there is a correlation between intellectually generated direct relevance decisions and indirect relevance decisions incorporated through citations or mentions in Altmetrics.

Our expanded iSearch collection allows us to compare and analyze direct and indirect relevance assessments. The following research questions are to be addressed with the help of this collection and a first data evaluation:

- RQ1** Are arXiv documents with relevance ratings published in journals with a higher impact?
- RQ2** Are arXiv documents with a relevance rating cited more highly or do they receive more mentions in Altmetrics?
- RQ3** In the literature, a connection between Mendeley readerships and citations is described. Is there evidence of a link between Mendeley readerships and citations in the documents with relevance ratings?

The paper is structured as follows: In Section 2, we describe the related work. Section 3 is about the data set generation and on the intersections between arXiv, Web of Science, and the Altmetrics Explorer. In Section 4, we use this new combined data set to answer the previous research questions. We discuss our first empirical results in Section 5 and draw some first conclusions.

2 Related Work

Borlund [2] proposed a theory of relevance in IR for the multidimensionality of relevance, its many facets, and the various relevance criteria users may apply in the process of judging the relevance of retrieved information objects. Later,

Cole [5] expanded on this work and asked about the underlying concept of information needs, which is the foundation for every relevance decision. While these works discuss the question of relevance and information need in great details, they lack a formal evaluation of their theories and thoughts.

White [20] combined relevance theory and citation practices to investigate the links between these two concepts further. He described that based on the relevance theory, authors intend their citations to be optimally relevant in given contexts. In his empirical work, he showed a link between the concept of relevance and citations. From a more general perspective, Heck and Schaer [9] described a model to bridge bibliometric and retrieval research by using retrieval test collections³. They showed that these two disciplines share a common basis regarding data collections and research entities like persons, journals, et cetera - especially with regards to the desire to rank these entities. These mutual benefits of IR test collections and informetric analysis methods could advance both disciplines if suitable test collections were available.

To the best of our knowledge, Altmetrics has not been a mainstream topic within the BIR workshop series. Based on literature analysis of the Bibliometric-enhanced-IR Bibliography⁴ only two papers explicitly used Altmetrics-related measures to design a study, Bessagnet in 2014 and Jack et al. in 2018. The reason for this low coverage of Altmetrics-related papers in BIR is unclear, as the inherent advantages in comparison to classic bibliometric indicators are apparent. One of the reasons Altmetrics has been approached is the time lag caused by the peer review and publication process of journal publications: It takes two years or more until citation data is available for a publication and thus, something can be said about its perception. The advantage of Altmetrics can, therefore, be a faster availability of data in contrast to bibliometrics.

On the other hand, there is no uniform definition of Altmetrics, and therefore no consensus on what exactly is measured by Altmetrics. A semantic analysis of contributions in social media is lacking for the most part, which is a major issue making the evaluation of Altmetrics counts so difficult. Mentions are mostly counted based on identifiers such as the DOI. However, it is not possible to mass evaluate which mentions should be deemed as positive and which should be deemed as negative, which means that a “performance paradox” develops. This problem exists in a similar form in classical bibliometrics and must be considered as an inherent problem of the use of quantitative metrics [10]. Haustein et al. [8] found that 21.5 % of all scientific publications from 2012 available in Web of Science were mentioned in at least one Tweet, while the proportion of publications mentioned in other social media was less than 5 %. In Tunger et al. [18], the share of WoS publications with at least one mention on Altmetric.com is already 42 %. It becomes visible that the share of WoS publications referenced in social media is continuously increasing. Among the scientific disciplines, there are also substantial variations concerning the coverage at Altmetric.com: publications

³ IR test collections consist of three main parts: (1) a fixed document collection, (2) a set of topics that contain information needs), and (3) a set of relevance assessments.

⁴ https://github.com/PhilippMayr/Bibliometric-enhanced-IR_Bibliography

from the field of medicine are represented considerably more often than, for example, publications from the engineering sciences. Thus, the question arises, to what extent the statements of bibliometrics and Altmetrics overlap or correlate.

3 Data Set Generation: Intersections between arXiv, Web of Science, and Altmetrics Explorer

This chapter describes the databases and the procedure for combining arXiv data with Web of Science and data from the Altmetrics Explorer.

The iSearch test collection includes a total of 453,254 documents, consisting of bibliographic book records, metadata records, and full-text papers [15]. The metadata records and full texts are taken from the arXiv, a preprint server for physics, computer science, and related fields. We exclude the bibliographic book records since no identifiers are available for retrieving WoS or Altmetric data. As shown in Figure 1, we focus on a total of 434,813 documents, consisting of full-text papers or abstracts. For all considered documents the arXiv-ID is available. With the help of this identifier, we query the arXiv-API⁵ and retrieve the DOI, if available.

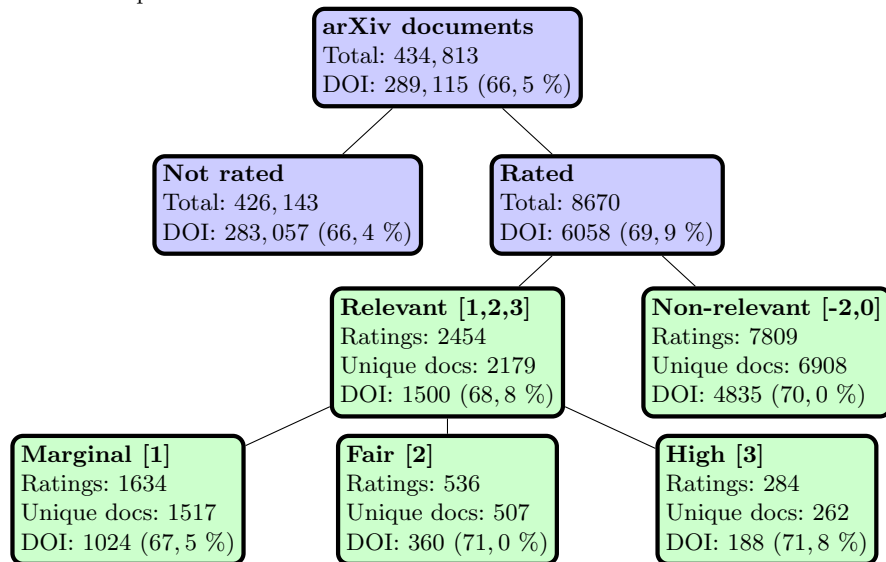
iSearch and WoS data are matched via DOI. iSearch and Altmetric data are matched via DOI or arXivID. This means that we end up not having data from the Web of Science or Altmetric Explorer for all iSearch documents, because the iSearch documents may only be covered in one of the other two databases, or in some cases neither. From WoS, *citation data* are added to the iSearch data sets and, generated via ISSN obtained from WoS data, a classification with the *science classification scheme* according to Archambault et al. [1], as well as the *Journal Impact Factor (JIF)* and the *Research Level*. From the Altmetrics Explorer the frequencies for the individual Altmetric document types on *tweets*, *news mentions*, *Wikipedia mentions*, *patent mentions*, and *Mendeley readership* are added to the iSearch data.

As illustrated in Figure 1, documents can be grouped with regard to their availability and level of relevance. For 8670 different documents ratings are available, and for 69,6 % of these documents, we were able to retrieve DOIs. This percentage of DOI coverage is slightly higher compared to that of documents that are not rated (66,4 %).

Furthermore, it is possible to group rated documents by their corresponding relevance level. Ratings are made with a graded relevance scale of five levels. Whereas relevant documents are rated from marginal-1, fair-2 to high-3. Non-relevant documents can either be rated as 0 or *explicitly* as non-relevant with -2. There are 10,263 relevance ratings for the considered 8670 unique documents. Some documents are rated multiple times across different topics with different levels of relevance. Figure 1 provides an overview of how relevance levels are distributed across different documents. For the sake of simplicity, we reduce the relevance levels by treating ratings of -2 and 0 as *non-relevant*. Since several

⁵ <https://arxiv.org/help/api>

Fig. 1. Document classification with regard to the availability and level of relevance. Blue colored nodes are based on document counts. Green colored nodes are based on ratings. The sum of *relevant* and *non-relevant* ratings is not equal to the number of rated documents, as there are documents with multiple ratings across different topics. Likewise, the sum of unique documents across the three relevance groups (*low*, *marginal*, *high*) is not equal to the number of unique documents that are *relevant* in general, because of duplicates.



documents are rated twice or more, we filtered out duplicates before investigating the DOI coverage of the documents with different relevance levels. As it can be seen, the percentage of DOI coverage increases with higher relevance for documents being rated as *relevant*. Besides, the percentage DOI coverage of marginally (71,0 %) and highly (71,8 %) relevant documents is slightly higher compared to that of documents that are rated non-relevant (70,0 %).

In sum, there are 1228 (out of 8670) documents with two or more ratings across different topics. In the following, we consider the relevance rating on a binary scale by treating ratings of -2 and 0 as *non-relevant* and ratings of 1 , 2 and 3 as *relevant*. 136 (11,1 %) documents are exclusively rated as *relevant*, 698 (56,8 %) documents are exclusively rated as *non-relevant*, and the remaining 394 (32,1 %) documents are rated both *relevant* and *non-relevant* across different topics. Therefore we only have a small intersection of contradicting relevance assessments for the same document which is under 5 % of the total count of judged documents (394 out of 8670).

Category	Documents	RA	CPP _{RA}	CPP _{-RA}
Nuclear & Particles Physics	30.5 %	10.0 %	42.8	34.6
Fluids & Plasmas	23.1 %	34.0 %	47.9	35.5
General Physics	18.8 %	19.4 %	76.9	56.5
Astronomy & Astrophysics	14.6 %	11.0 %	61.9	46.4
Applied Physics	3.1 %	11.5 %	46.8	34.7

Table 1. Percentage of documents per WoS category. RA is the percentage of documents per category that got a relevance assessment. CPP_{RA} is the number of cites per paper for all documents that got a relevance assessment. CPP_{-RA} is the same for all documents without a relevance assessment.

In the next step, we combine the arXiv data with WoS and the Altmetrics Explorer⁶ to obtain statements on the impact of these publications both in the scientific world and beyond in social media. The matching between arXiv and WoS is carried out via DOI. For 4061 out of 10,263 ratings, WoS data is available. Of the documents with DOI that were matched with the WoS, the publications in arXiv can essentially be assigned to four major categories, as shown in Table 1. The distribution of relevance ratings by category shows a slightly different picture, which indicates a small shift and shows that the category with the largest number of articles is not automatically the category with the most relevance assessments.

4 Results

4.1 Relevance and Journal Impact Factor

RQ1 focuses on the relationship between positive relevance assessments and the perception of documents. Or in other words: Whether a publication with a positive relevance rating can achieve a higher scientific perception or a higher perception in Altmetrics. The question is, therefore, whether there is a connection between high relevance rating and high perception.

If we look at Table 2, we see that the unrated documents, which form the vast majority, have a lower citation rate than the relevant-rated documents, on average, about 41 citations per document. The documents with relevance assessments reach higher citation rates than the group of documents without relevance assessment. The highest citation rate is achieved by the documents that have been grouped into the highest relevance level 3: With a citation rate of 76.2, they achieve a citation rate almost twice as high as documents without a relevance rating. The group of documents with relevance rating is small compared

⁶ Altmetrics is also topic of the project UseAltMe (funding code 161F1107): On the way from article-level to aggregated indicators: understanding the mode of action and mechanisms of Altmetrics https://www.th-koeln.de/en/information-science-and-communication-studies/usealtme_68578.php

Relevance	Documents	Citations	CPP	JIF	RL
Non-relevant	2990	157880	52.8	4.4	3.7
Marginal (1)	683	45849	67.1	4.2	3.7
Fair (2)	257	16896	65.7	4.4	3.6
High (3)	131	9976	76.2	4.7	3.6
Not rated	136172	5610332	41,2	4.5	3.8

Table 2. Groups of relevance with corresponding Web of Science (WoS) data. For each relevance group, the sum of citations (Citations), the average citation rate (CPP), the average journal impact factor (JIF), and the average research level are included. The number of documents results from the availability of WoS data retrieved by DOI.

to the group without relevance ratings. Nevertheless, from the authors' point of view, the group size is sufficient for them to be able to read rough trends from it using bibliometric methods and to find answers to the research questions. It is, therefore, not so much the small deviations that are important here, but rather the more specific trends.

Table 1 shows the citation rates for the five subject categories, which together account for 90 % of the arXiv documents, for the documents with and without relevance assessment. It can be seen that the citation rates for publications with a relevance assessment are significantly higher for all categories shown than for publications without a relevance rating in the same category (Pearson = 0.997). The presentation of the citation rates in Table 2 for the individual groups of documents with and without a relevance rating goes in the same direction: One of the trends also lies in the observation that all publications with a relevance rating have a higher citation count than the group whose documents were not selected as relevant.

If there is a correlation, it can be demonstrated at other points: When looking at the results, it is noticeable that the citation rate changes depending on the degree of relevance assessment. In the highest level of relevance assessment, level 3, the citation rate of 76.2 is almost twice as high as for the non-assessed documents. The citation rate increases continuously from the first assessment level 0 to level 3. Level 2 is an exception, where the citation rate is lower than in level 1, but still higher than the citation rate of the non-evaluated documents.

Are there differences in the composition of the groups that would explain the differences in citation rates described above? The Journal Impact Factor (JIF) does not show a difference for any of the groups. It only differs by fractions of a decimal point. The documents without relevance rating have an average JIF of 4.5, the documents with relevance rating have an average JIF between 4.4 and 4.7. This shows that there is no significant difference in the composition of the individual groups as to whether they publish more in high- or low-impact journals. The structure of all groups is the same in terms of average journal impact. Thus, RQ1 can be answered to the effect that the impact of a journal has no influence on the decision about the relevance of a document.

4.2 Relevance and Citation Rates/Altmetrics Mentions

If we look at **RQ2**, we can say that arXiv documents with relevance ratings achieve a higher perception in terms of citation rate than unrated documents. This observation cannot be directly transferred to Altmetrics, where there is no difference in the average number of tweets or news articles. The only measurable difference refers to the documents with a relevance rating of 3, here an above-average number of tweets or mentions in patents can be observed. This effect is the result of a skewed distribution and is presumably favored by two publications: A publication for which 253 tweets exist and another publication for which 50 mentions in patents are recorded. Without these two publications, there are no significant differences between documents with and without relevance rating. Thus, RQ2 can be answered to the effect that documents with a relevance rating receive a higher number of citations on average, but with the exception of Mendeley, there is no effect on Altmetrics.

An explanation of why we do not see an effect for Altmetrics in the arXiv data may be in the year of publication: The majority of the publications were published between 2003 and 2009. During this time, social media were already being used actively in society, but not yet in science. This changed only slowly towards 2008, with the Altmetric Explorer, for example, being founded by Euan Adie in 2011. It is not known to what extent publications prior to the founding year of Altmetric Explorer were retroactively re-indexed and to what extent this is technically possible at all. This is because, in contrast to scientific journal publications, communication in social media is fast and can also be deleted before it has been indexed. These effects have to be taken into account when dealing with altmetrics, as well as the fact that the publications originate from several publication years, so some had more time to generate attention than others. If it is necessary, “that older articles are compensated for lower altmetric scores due to the lower social web use when they were published” [17], is a question that is legitimate but not the focus of this publication. But overall it could be shown, that there is a connection between Citations and Altmetric counts, as also shown by Holmberg et al. [11].

4.3 Mendeley Readerships and Citation Rates

In the literature, the question of whether there is a correlation between citations of scientific publications in Web of Science, Scopus, or Google Scholar to Mendeley readerships (**RQ3**) has often been examined. Li & Thelwall [14] have investigated whether there is a correlation between citations from the three mentioned databases and whether there is a connection between the number of citations and the number of bookmarks of a publication on Mendeley. The result of this investigation was a perfect correlation between the citation counts from the three citation databases Web of Science, Scopus, and Google Scholar. This result is also less surprising because Web of Science and Scopus are about 95 % overlapping, and there is also a big overlap between these two databases and

Relevance	News	Twitter	Patent	Wikipedia
Non-relevant	0.07	0.7	0.3	0.2
Marginal	0.02	0.6	0.3	0.2
Fair	0.3	0.5	0.9	0.2
High	0.2	1.3	4.7	0.3
Not rated	0.05	0.2	0.6	0.3

Table 3. Altmetric scores: News mentions, Twitter mentions, Patent mentions and Wikipedia mentions.

Google Scholar. A correlation between citations from the three mentioned citation databases to Mendeley was also measurable by Li & Thelwall, but it is much worse than the correlation between the citation databases. This is not surprising since Mendeley readerships are not scientific citations. Bookmarking of publications takes place for other reasons than citing a scientific paper. But also, Costas et al. [6] found out, that “Mendeley is the strongest social media source with similar characteristics to citations in terms of their distribution across fields”.

The present study is based on about 435,000 arXiv publications. However, it was not possible to determine citations or Mendeley readerships counts for all of these documents: For 32,081 documents out of the total number, both citation data and data from Altmetric Explorer are available. This quantity contains 1037 documents with a relevance rating. This is where RQ3 comes into account: Is there evidence of a link between Mendeley readerships and citations in the arXiv-documents with relevance ratings? A correlation coefficient, according to Pearson of 0.83 was determined for the total quantity of 32,081 documents, which indicates an existing correlation, even if the value is not perfect. Instead, the result indicates that Mendeley and Web of Science are not entirely overlapping in terms of the perception of the documents.

The result is not significantly different if one takes the 1037 documents with relevance rating out of this set; Pearson is about at the same level at 0.8 (see Table 4). It should be noted that reasons for bookmarking a publication can be different from reasons for a citation: There are publications that are roughly equal in both data sources, for example, 10.1103/PhysRevE.67.026126, which receives 1068 citations and 1192 Mendeley reads. There are also examples of unequal perception: 10.1088/0067-0049/182/2/543 has 3151 citations but is bookmarked “only” 405 times on Mendeley or 10.1088/0954-3899/33/1/001 has 3903 citations but is bookmarked only 24 times, or in the opposite direction 10.1142/S0218127410026721 is cited only eight times but is bookmarked 106 times on Mendeley. So outliers can occur in both directions, which can lead to distortions in the measured correlation.

It becomes interesting if we take individual groups out of the entire group of 1037 documents with relevance assessments: Both for the group of 786 publications, which were singled out as possibly relevant but then rated 0 as non-relevant. For the group of 35 publications, which were rated 3 and thus placed

Relevance	Documents	Pearson M
Non-relevant	786	0.85
Marginal	154	0.72
Fair	62	0.76
High	35	0.89
Not rated	32081	0.83

Table 4. Number of documents with Mendeley readership and citations within the WoS. The Pearson correlation is calculated WoS citations

in the group of the most relevant documents, we get an even better correlation: For the 786 publications rated 0 we get a Pearson correlation of 0.85 and for the 35 publications rated 3 we get a correlation of 0.89. For the groups rated 1 (Pearson = 0.72) or 2 (Pearson = 0.76), we get a worse correlation value in each case because, in these two groups, the number of outliers is higher. From our point of view, the result is to be understood in such a way that a clear relevance decision filters out the papers that receive roughly the same perception on both sides, citation and Mendeley. Decisions in the middle of the relevance scale, on the other hand, filter out papers where the perception may tend to be more to one side. Thus, two things can be observed concerning RQ3: there is a link between citation data and Mendeley readerships. This becomes all the more visible if the paper is part of a set of documents that have previously been subjected to a corresponding relevance assessment.

5 Discussion and Outlook

The results of our small study on the intersection of relevance assessments within the iSearch collection and corresponding citation counts show that direct relevance decisions of a single assessor and indirect decisions of many external authors citing this work are related. What sounds intuitive and like common sense is not fully backed by the literature as the connection between citations and relevance is not undisputed. Ingwersen [12] explains that “citations are not necessarily good markers of relevance, because impact and relevance might not always be overlapping phenomena.” While this might be true sometimes, in other situations, references have been shown to improve retrieval quality as additional keys to the contents [7]. One general conclusion of this contradiction is that citations more represent the general *popularity* or *perception* of a document, which is not the same as a relevance judgment.

Another thing we have to notice about our work is that the differentiation of direct and indirect relevance decisions is not an established concept in information theory. While relevance is often described as multidimensional, layered, et cetera, the terms *direct* and *indirect* are suggestions of the authors of this paper. A concept that is aligned to the different levels and forms of relevance might be the principle of poly representation [16]. Polyrepresentation might be the

common ground where the general popularity of a document measured through citations and the concrete relevance come together.

When looking at JIF, Kacem and Mayr [13] describe that users are not influenced by high impact and core journals while searching. This is in line with our results, as we cannot measure a significant difference in the JIF of unrated, non-relevant, or relevant documents. However, we have to keep in mind that judging on static document lists to generate a test collection might be different from interactive search sessions, which were the basis of the studies of Kacem and Mayr. Regarding the connection between citations and Mendeley readerships, the literature is confirmed. We can clearly reproduce the correlation between these two entities. If one follows the implications of Altmetrics described at the beginning, such a result also appears desirable, because this means that Mendeley data also contain additional information that is not contained in the Web of Science. This follows the goal of Altmetrics also to provide new information and not just to be faster bibliometrics. We must conclude that the reasons for citation and bookmarking are similar but not the same. A bookmark is a reference to a publication that is believed to be of interest to others. This does not necessarily imply that you have read the publication yourself.

The impact of our work is that we could show a relation between direct relevance assessments and indirect relevance signals originating from bibliometric measures like citations. This relation is visible but not fully explainable. There seems to be something inherent in relevant documents that let them gather a higher number of citations. We are sure that it is not the impact of the corresponding journal. Otherwise, there would be no uncited documents within Nature. Popularity alone seems not to explain this effect. Maybe citations in relation to relevance assessments are a marker for “quality”, although we are aware that this term is highly controversial in the bibliometrics community.

It remains a future work to evaluate and investigate the phenomena that are the reason for the relationship we have seen. The principle of poly representation might be an excellent framework to bring together these different factors originating from relevance theory, bibliometrics, and Altmetrics. Additionally, it might help to design a retrieval study to follow these open questions further.

References

1. Archambault, É., Beauchesne, O.H., Caruso, J.: Towards a multilingual, comprehensive and open scientific journal ontology. In: Proc. of ISSI 2011. pp. 66–77. Durban South Africa (2011)
2. Borlund, P.: The concept of relevance in IR. *Journal of the American Society for Information Science and Technology* **54**(10), 913–925 (Aug 2003). <https://doi.org/10.1002/asi.10286>
3. Breuer, T., Schaer, P., Tunger, D.: Relations Between Relevance Assessments, Bibliometrics and Altmetrics (Mar 2020). <https://doi.org/10.5281/zenodo.3719285>
4. Carevic, Z., Schaer, P.: On the connection between citation-based and topical relevance ranking: Results of a pretest using isearch. In: Proc. of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with ECIR 2014. pp. 37–44 (2014)

5. Cole, C.: A theory of information need for information retrieval that connects information to knowledge. *Journal of the American Society for Information Science and Technology* **62**(7), 1216–1231 (Jul 2011). <https://doi.org/10.1002/asi.21541>
6. Costas, R., Zahedi, Z., Wouters, P.: The thematic orientation of publications mentioned on social media: Large-scale disciplinary comparison of social media metrics with citations. *Aslib Journal of Information Management* **67**(3), 260–288 (May 2015). <https://doi.org/10.1108/AJIM-12-2014-0173>
7. Dabrowska, A., Larsen, B.: Exploiting citation contexts for physics retrieval. In: Proc. of the Second Workshop on Bibliometric-enhanced Information Retrieval co-located with ECIR 2015. pp. 14–21 (2015)
8. Haustein, S., Costas, R., Larivière, V.: Characterizing Social Media Metrics of Scholarly Papers: The Effect of Document Properties and Collaboration Patterns. *PLOS ONE* **10**(3), e0120495 (Mar 2015). <https://doi.org/10.1371/journal.pone.0120495>
9. Heck, T., Schaer, P.: Performing Informetric Analysis on Information Retrieval Test Collections: Preliminary Experiments in the Physics Domain. In: Proc. of ISSI 2013. vol. 2, pp. 1392–1400. Vienna, Austria (2013)
10. Holbrook, J.B., Barr, K.R., Brown, K.W.: We need negative metrics too. *Nature* **497**(7450), 439–439 (May 2013). <https://doi.org/10.1038/497439a>
11. Holmberg, K., Bowman, T., Didegah, F., Lehtimäki, J.: The Relationship Between Institutional Factors, Citation and Altmetric Counts of Publications from Finnish Universities. *Journal of Altmetrics* **2**(1), 5 (Aug 2019). <https://doi.org/10.29024/joa.20>
12. Ingwersen, P.: Bibliometrics/Scientometrics and IR A methodological bridge through visualization (Jan 2012), <http://www.promise-noe.eu/documents/10156/028a48d8-4ba8-463c-acbc-db75db67ea4d>
13. Kacem, A., Mayr, P.: Users are not influenced by high impact and core journals while searching. In: Proc. of the 7th International Workshop on Bibliometric-enhanced Information Retrieval co-located with ECIR 2018. pp. 63–75 (2018)
14. Li, X., Thelwall, M.: F1000, mendeley and traditional bibliometric indicators. In: Proc. of the 17th International Conference on Science and Technology Indicators. pp. 541–551 (2012)
15. Lykke, M., Larsen, B., Lund, H., Ingwersen, P.: Developing a test collection for the evaluation of integrated search. In: Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010. Proceedings. pp. 627–630 (2010). https://doi.org/10.1007/978-3-642-12275-0_63
16. Skov, M., Larsen, B., Ingwersen, P.: Inter and intra-document contexts applied in polyrepresentation for best match IR. *Information Processing & Management* **44**(5), 1673–1683 (Sep 2008). <https://doi.org/10.1016/j.ipm.2008.05.006>
17. Thelwall, M., Haustein, S., Larivière, V., Sugimoto, C.R.: Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE* **8**(5), e64841 (May 2013). <https://doi.org/10.1371/journal.pone.0064841>
18. Tunger, D., Meier, A., Hartmann, D.: Altmetrics Feasibility Study. Tech. Rep. BMBF 421-47025-3/2, Forschungszentrum Jülich (2017), <http://hdl.handle.net/2128/19648>
19. Voorhees, E.M.: TREC: Continuing information retrieval’s tradition of experimentation. *Communications of the ACM* **50**(11), 51 (Nov 2007). <https://doi.org/10.1145/1297797.1297822>
20. White, H.D.: Relevance theory and citations. *Journal of Pragmatics* **43**(14), 3345–3361 (Nov 2011). <https://doi.org/10.1016/j.pragma.2011.07.005>