Scientometrics (2022) 127:2455–2478
https://doi.org/10.1007/s11192-022-04319-4

Relevance assessments, bibliometrics, and altmetrics: a quantitative study on PubMed and arXiv

Timo Breuer[1] · **Philipp Schaer**[1] · **Dirk Tunger**[1,2]

Received: 14 June 2021 / Accepted: 18 February 2022 / Published online: 6 May 2022

Abstract
Relevance is a key element for analyzing bibliometrics and information retrieval (IR). In both domains, relevance decisions are discussed theoretically and sometimes evaluated in empirical studies. IR research is often based on test collections for which *explicit* relevance judgments are made, while bibliometrics is based on *implicit* relevance signals like citations or other non-traditional quantifiers like altmetrics. While both types of relevance decisions share common concepts, it has not been empirically investigated how they relate to each other on a larger scale. In this work, we compile a new dataset that aligns IR relevance judgments with traditional bibliometric relevance signals (and altmetrics) for life sciences and physics publications. The dataset covers PubMed and arXiv articles, for which relevance judgments are taken from TREC Precision Medicine and iSearch, respectively. It is augmented with bibliometric data from the Web of Science and Altmetrics. Based on the reviewed literature, we outline a mental framework supporting the answers to our research questions. Our empirical analysis shows that bibliometric (*implicit*) and IR (*explicit*) relevance signals are correlated. Likewise, there is a high correlation between biblio- and altmetrics, especially for documents with *explicit* positive relevance judgments. Furthermore, our cross-domain analysis demonstrates the presence of these relations in both research fields.

Keywords Relevance assessments · Polyrepresentation · Altmetrics · Relevance theory · Information retrieval · Test collections

✉ Philipp Schaer
Philipp.Schaer@th-koeln.de

Timo Breuer
Timo.Breuer@th-koeln.de

Dirk Tunger
Dirk.Tunger@th-koeln.de

[1] Faculty of Information Science and Communication Studies, Institute of Information Management, TH Köln - University of Applied Sciences, Cologne, Germany

[2] Forschungszentrum Jülich GmbH, Project Management Jülich, Center of Excellence "Analyses, Studies, Strategy", Cologne, Germany

Introduction

The broader concepts of relevance and scientific communication (in the meaning of bibliometric and altmetric perception) are interconnected as shown on a theoretical (White 2011) and empirical level (Larsen and Ingwersen 2002; Breuer et al. 2020). Both are fundamental in bibliometrics and information retrieval (IR) but are shallow concepts that are interrelated but do not allow us to directly observe these connections. We can only measure visible entities, like citations and relevance judgments, and common observable patterns, like skewed, power law-like distributions (White 2017).

One representation of these visible entities in the IR context are test collections and their relevance assessments. These assessments are based on the description of an underlying information need that is usually formulated in a specific topic to guide the assessors (Voorhees 2007). While this does not explain each single relevance assessment, it gives us a brief understanding of the reasons for the assessors' judgments. This topical focus and information need is totally missing for citation data, which is the corresponding visible entity for scientific communication. We have no further knowledge on why someone cited another work, but we know to a certain degree why someone judged a document relevant or not. This discrepancy makes it hard to compare the two different types of relevance decision processes: One is a direct way of expressing and judging relevance, and therefore (1) an *explicit relevance signal*, while the other is only (2) an *implicit relevance signal*. Explicit relevance decisions are labeled by assessors with a reasonable level of context information while implicit relevance signals originate from users who are not part of the design process of the test collections. We do not know anything about the information need or retrieval context for these external users. We only know that they cited a specific paper - therefore, we have to assume this paper to be somehow relevant to them as a cumulative process of a group. Otherwise, they would not have cited it. Both types of relevance decisions and signals are connected through an underlying socio-cognitive selection process that assessors and citers go through (Garfield 1996, 1998).

Decision processes regarding relevance in relation to search results and decision processes leading to citing a reference are different. A searcher evaluating single documents in a result list may have a specific goal and information need. An anonymous mass of citers and their observable cumulative relevance signals in the form of citations can only be seen as a conglomerate of many different motivations. Nevertheless, due to the inherent information overload, assessors and citers are both confronted with the issue of discriminating one piece of information against the other and might share not all but some common elements in this decision process.

We conduct a large-scale data analysis on two IR test collections to test and evaluate our assumptions regarding the common elements of relevance decision processes and citation practices. The main goal of this evaluation is to verify whether there is a link between the mechanisms of the underlying concepts of relevance and scholarly communication, as we assume. Therefore, we will focus on two main elements in our evaluation: (1) References of scientific works in the form of citations or altmetric indicators like social media mentions, and (2) relevance judgments by assessors.

IR test collections provide a proven empirical research environment to investigate some previously mentioned entities but usually lack scientific references. To the best of our knowledge, only very few collections contain scientific documents and additional implicit relevance signals in the form of reference information. We investigate the connections by looking at the intersections of IR test collections and external citation/usage data in expanded test collections that we will introduce later in this work. For our setting, IR test collections and their relevance assessments represent explicit, intellectual relevance decisions based on topical criteria, while citations and other usage data are some forms of implicit, hidden, or cumulative relevance decisions.[1]

A collection that incorporates both is the iSearch collection introduced by Lykke et al. (2010). It combines a classic document collection derived from arXiv.org, 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. Another suitable and more recent test collection is the TREC Precision Medicine (TREC PM) collection. It includes biomedical articles in the form of article abstracts derived mostly from MEDLINE/PubMed. Although iSearch also includes citation data, one known limitation is the small overlap of citations and relevance judgments in iSearch since only internal citations within the corpus were counted (Carevic and Schaer 2014). The TREC PM collection does not contain any citation data. Therefore, we expand these two collections by complementing them with external citation data from Web of Science (WoS) and social media usage data from Altmetrics Explorer.

Our expanded test collections allow us to compare and analyze the results of explicit and implicit relevance decision processes. With the help of these collections and metrics from bibliometrics and altmetrics we will address the following research questions:

RQ1 Is there a correlation between documents from an IR test collection with a relevance score and the corresponding citations of these scientific publications or their social media mentions?
RQ2 The literature describes a correlation between some altmetrics indicators, such as Mendeley Readerships, and citations. Does this correlation differ between documents with relevance assessments and documents without relevance assessments?
RQ3 It is known that publication and citation habits vary between scientific fields. Is there a difference in the effects measured in RQ1 and RQ2 between test collections from dissimilar scientific fields?

The paper is structured as follows: In Sect. Related work, we describe the related work on the connections between scientific communication and relevance and propose a first outline of an integrated framework in Sect. Connections between relevance and scientific communication to bridge these domains, especially in IR contexts. Section Experimental evaluation data set is about the experimental design, the data set generation, and our empirical study on the intersections between relevance judgments and citation/usage data extracted from WoS and the Altmetrics Explorer. In Sect. Results, we use this new combined data set to answer the previous research questions. We discuss our empirical results in Sect. Discussion and draw our conclusions in Sect. Conclusion.

[1] The idea and first results on the topic of this publication were presented at the BIR Workshop 2020 (Breuer et al. 2020). This publication is a significant extension of this initial idea, containing more data sets, an extension to other scientific domains, and a generalization of the approach.

Related work

The concepts of relevance and scientific communication share common elements on different levels. These common elements are shallow, as theories at large are not observable or measurable. In the social sciences, we name these latent constructs. At the intersection of scientometrics and IR (Mutschke et al. 2011) coined the term "litmus test" to use methods (and in our case data sets) from the corresponding discipline to examine for adequacy of the underlying models and theories. In the following, we will describe the related work on the concepts of relevance and scientific communications with special regards to the scientific perception of others' work, in the form of citations or social media mentions.

There are many different kinds of relevance rooted in human cognition (Mizzaro 1997). Cosijn and Ingwersen (2000) assigned the following attributes to relevance: Relevance always implies a relation, e.g., in communication or exchange, which means that relevance is never static. This relation involves intentions such as objectives and expectations. Users never start from zero but always combine their knowledge and experience with their information need and their expectations, what they would like to find with their search queries (White 2011). This kind of intention is always integrated into a context (Mizzaro 1997) and is always directed towards this context and towards interaction, which aims inference to be the assessment of the effectiveness of a given relation (Ingwersen 1992).

Cosijn and Ingwersen (2000) argue that affective relevance is not a discrete category or part of a linear scale. It should rather be viewed as part of the subjective types of relevance that can be classified as topical, cognitive, situational, and socio-cognitive relevance. The last one is measured in terms of the relation between the situation, work task, or problem. In this kind of relevance, intention signifies the strategy or tactical decisions of a group of people, a network, or an organization. An example of topical relevance is found during a peer-reviewing process for a conference or a journal. One central criterion in the review process is whether a paper is on the conference's or journal's topics or not.

Relevance theory consists of two basic principles. The first or *cognitive principle* of relevance says that human cognition tends to be geared toward maximization of relevance (MacKenzie 2002; Wilson and Sperber 2004; Sperber and Wilson 2001). Relevance maximization is rooted in how our cognitive system has evolved: selection pressure towards higher efficiency has led to this system automatically taking in potentially relevant stimuli and our processing system automatically drawing relevant conclusions from them. The second or *communicative principle* of relevance says that every expression of information is (1) relevant enough for it to be worth the addressee's effort to process it, and (2) the most relevant one compatible with the communicator's abilities and preferences (White 2011). Cognition is considered an individual process, while communication is a group or collective process. The communicative relevance principle states that utterances always create the expectation of being relevant. Speakers encourage the audience to assume that their utterance is relevant, and scientists always assume that their paper is interesting for the scientific community (White 2011). Otherwise, they would not have submitted it to a journal. Here we have the connection to scientific communication: This results in the scientists' expectation that colleagues may cite their papers. That this is not always the case is shown by the fact that there are many uncited or low-cited publications also in highly cited journals like Nature or Science.

Scientific communication generates scientific publications that aim for relevance and underlie the rules of relevance theory, as described before. These publications are not only part of scientific communication but also part of IR. Publications are part of IR systems

(such as the former Science Citation Index (SCI) as a book edition and today's WOS as a search engine) to be found if they are relevant (Garfield 1964).

The basic idea of Garfield when developing SCI was to select the journals according to their relevance for the respective field area (Garfield 1972): the most relevant journals from each scientific field were to be covered in SCI. These journals were called core journals. In this selection process, a sort of relevance assessment is carried out based on the number of citations that publications in a journal can receive on average, which is called impact: "The citing papers one retrieves from a citation index search are assumed to have a subject relevance to the idea symbolized by the cited item targeted for the search" (Moed 2005). Langham (1995) states about citations: "The function of any citation-signaller is to alert the reader to some kind of association between the citing text and the cited text. Citation-signallers may additionally, by using page-references or chapter numbers, single out a particular part of the cited text as especially relevant. This additional information can be very useful if it is genuinely the intention of the citing author that the cited text be read by his or her own readership."

As described by Ingwersen (2012, minute 8), the number of citations is not a relevance characteristic: "Citations do not signify relevance! But the number of citations signifies utility in a particular work/context". A high number of citations is not an indication of topical relevance, but of usefulness in a particular situation. For example, by citing a publication, a negative example can be given. In this case, the chosen citation would indicate a wrong source, which objectively does not contain any relevant information. According to Ingwersen, a citation index can, therefore, only be used to say something about "social utility" or "academic (re)cognition". He clearly separates this from relevance. "The motives for giving (or not giving) a reference to a particular article may vary considerable" as van Raan (2005) stated. The reason for citing a publication is to refer to scientific results that are relevant for your own publication. However, sometimes not the most appropriate publications are cited, but those whose author was known or those the reviewer asked for. A clue on what preference criteria for information sources might be, is provided by Fisher and Naumer (2006) who identify several criteria in their study. Among these are trustworthiness, contact, access or convenience, inexpensiveness, and ease of use. Ashford (1986) mentioned that to the very end the source selection is based on the relation of information quality and effort of seeking the source. This means, that not necessarily the best possible source is used.

Summarizing, vanRaan (2005, p. 8) argues, that "[...] these 'reference motives' are not so different or 'randomly given' to such an extent that the phenomenon of citation would lose its role as a reliable measure of impact". Glänzel (2008) added to this discussion: "In spite of their statistically evidenced correlation with quality related aspects, citations in general, and impact factors in particular are and remain primarily indicators of reception of scientific information." These points of view show that the question of what significance a citation has in a scientific publication is not conclusively clarified. According to Garfield (1966) citations mark a flow of knowledge: "Theoretically, if two different papers contain the same list of 'references', then they are essentially the same".

In contrast to the long history of bibliometrics, altmetrics is a considerably young discipline. On the one hand, the visibility and presence of altmetrics are quite impressive; several hundred publications on the subject have appeared, and there are conferences and workshops dedicated solely to altmetrics. On the other hand, there is no uniform definition of the term, and therefore no consensus on what exactly is measured by altmetrics and what conclusions can be drawn from their results (Tunger et al. 2017; 2018). Among the scientific disciplines, there are also substantial variations concerning the coverage at Altmetric.

com: publications 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. Ultimately, it is still unclear what exactly is measured with altmetrics and what basic statements about science concerning authors, institutions, and other stakeholders, can be derived from respective data (Butler 2017).

The previous discussion of relevance were aligned to the judgment of relevance by the means of citations, mentions and such. In IR the concept of relevance is described by Borlund (2003). She proposed a theory of relevance in IR with respect to 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 (2011) expanded on this work and asked about the underlying concept of information needs, which is the foundation for every relevance decision. These works discuss relevance and their underlying information needs and cognitive aspects in great detail.

Ingwersen (1994) distinguishes between four different types of how these entities can be represented and align these to the principle of polyrepresentation. At its very core, polyrepresentation relies on intentional redundancy. Different representations of the same entity correspond to likewise different cognitive structures, but figuratively speaking these cognitive structures can overlap if they coincide with each other. It is this overlap that is perceived by the recipient as an indicator of relevance. Larsen et al. (2006) see citations as author-generated representations and the recipient's cognitive space is essential for the retrieval of information and thus for relevance decisions. On the questions how these overlaps can be put into operation (Heck and Schaer 2013) outlined a conceptional framework that pointed out the mutual benefits of IR test collections and informetric analysis methods. In a similar vein, Carevic and Schaer (2014) investigated the connection between topical relevance assessments and document recommendations originating from a co-citation analysis. They found some intersections of relevant documents and the co-citation recommendations but were limited because of the low number of internal references within the iSearch collection used in the experiments.

Connections between relevance and scientific communication

As shown in the description of the related work in Sect. Related work, we have seen that relevance is dynamic and can be assigned in the context of tasks in research assessments and the associated individual assessment of publications by peers. In the same way, relevance can be understood as an overall assessment of a group and can be assigned collectively and additively in the overall view of a publication set and can be derived from data on the perception and use of these publications. This requires cognitive processing of a publication, either in a specific IR task but also in the process of writing scientific publications. The form in which the relevance of a publication is subsequently expressed publicly can vary. It can consist of relevance ratings and citations or altmetric rewards.

Thus, all publications in the WOS have a relevance assessment, which is not carried out intellectually and on a topical basis as in the IR relevance assessment process, but cumulatively for the field as a whole by using the measurable results of the scientific communication process. Since scientific communication is no longer limited to citations, and there are no fundamental differences between bibliometrics and altmetrics we also consider altmetrics as another relevance assessment source. Their structure is very similar and

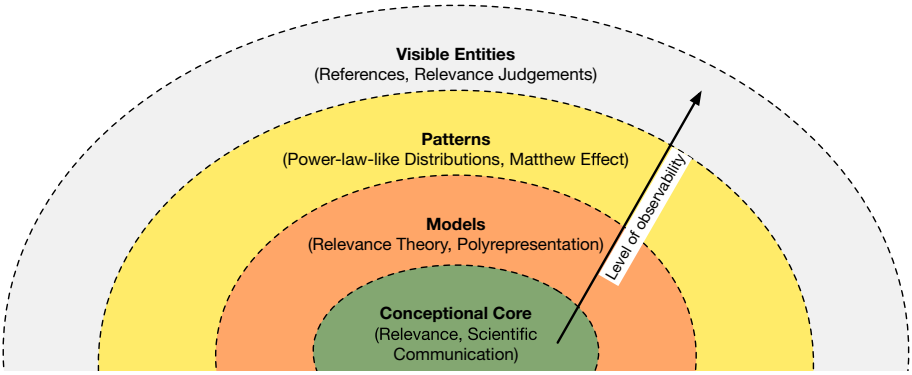

Fig. 1 Different levels of the connection between the concepts of relevance and scientific communication, ranging from the measurable and visible entities, like reference lists and relevance judgments, to the conceptional core that includes the concepts of relevance and scientific communication itself

the distribution of altmetrics is just as skewed as the distribution of citations on publications. We will consider references in social and web-media, the citations here are known as tweets, likes, news items, mentions, shares, reads etc.

After we have seen apparent connections and dependencies between relevance and scientific communication, we built a very brief mental framework that allows us to better map and align experiments in the intersection of both domains. This framework borrows from Cole (2011) in the form that we define the two central concepts of relevance and scientific communications as immeasurable and as a black box. Although Cole was investigating the concept of information need, we see the same pattern as it was coined by Taylor (1968): At the deepest level, these concepts tend to be "inexpressible in linguistic terms". To overcome this inherent issue Cole used so-called surrogates or adjacent concepts to create testable propositions for his central assumptions.

The framework itself is visualized (see Fig. 1) in the form of different layers around a nucleus (green layer) that consists of the two core concepts relevance and scientific communication, where previous work described latent but apparent overlaps (Ingwersen 2012; White 2011). The further we move away from the core concepts themselves we observe more visible phenomena and measurable entities. In the second tier (orange layer), which contains the theoretical modeling, the theory behind the concept of relevance is illuminated, and the principle of polyrepresentation is brought in, which we identify as a connecting element. The third tier (yellow layer) contains the observable patterns, above all, recurring highly skewed power-law distributions, but also domain-specific publication and citation behavior. The last tier (gray layer) focuses on the visible entities: the references (like citations in bibliometrics, links, likes, or shares in altmetrics) and relevance judgments. They are the most tangible and therefore measurable unit, which are repeatedly the subject of evaluation and partly overlap. References and relevance judgments are the two main instantiations of the two main concepts from the conceptional core that are put into application.

We will use this framework in the following experimental evaluation as a mental model of the interconnection between the two main concepts of this work: Relevance and scientific communication.

Experimental evaluation data set

In the following, we describe the process of data set generation that includes compiling data from arXiv, PubMed, WOS, and the Altmetrics Explorer (Sect. The iSearch and TREC precision medicine collections). We use this data set to analyze the intersections between matching documents with biblio- and altmetrics in Sect. Matching relevance assessments with biblio- and altmetrics.

The iSearch and TREC precision medicine collections

The iSearch test collection includes a total of 453,254 documents from the domain of physics, consisting of bibliographic book records, metadata records, and full-text papers (Lykke et al. 2010). The metadata records and full texts are taken from arXiv.org - a preprint server for physics, computer science, and related fields. The bibliographic book records of the iSearch collection were excluded since no identifiers are available for retrieving WOS or altmetric data.

In order to lower the level of abstraction, the underlying topics of this test collection are based on real-world search tasks. Conventionally, assessors or domain experts are provided with pre-defined topics for which relevance assessments have to be made. Assessors judge the relevance based on descriptions and narratives added to the topics to better understand the underlying information need. The assessors in iSearch were asked to contribute their own topics based on their current search tasks. They extensively described these search tasks for a better transparency and understanding why the corresponding assessments were made. Additionally, the assessor provided information on their demographics and their level of domain and retrieval expertise. This means the relevance assessments comprise judgments made with different levels of expertise ranging from MSc students to lecturers. More details on the procedure can be found in the corresponding publication by Lykke et al. (2010). In total, there are 8670 judgments available which account for approximately 2 % of all considered documents. Since we excluded book records these are 434,813 in total.

The TREC PM track is dedicated to finding specific treatments for individual patients and follows a history of different biomedical retrieval attempts, including Genomics, Medical Records, and finally Clinical Decision Support (Roberts et al. 2019). In this study, we investigate the relevance judgments taken from the Clinical Decision Support Track 2014-2016 and the TREC PM 2017-2019. All of the test collections are based on PubMed records. The investigated snapshot from mid-December 2018[2] contains 29,138,916 MEDLINE abstracts. In this case, the corresponding topics are either synthetically created or de-identified patient data. Here, assessors were physicians like oncologists, biomedical informatics students or other medical experts (Roberts et al. 2019). In contrast to the iSearch collection, the assessors were provided with explicit information and requirements regarding the topics/patient data. This includes information on the patients' disease and demographics. In total, there are 116,437 judgments available, which account for 0.4 % of all considered documents.

When comparing both relevance judgments processes, two differences are apparent. First, with regard to the topic selection, the iSearch assessors could basically decide which

[2] http://www.trec-cds.org/2019.html.

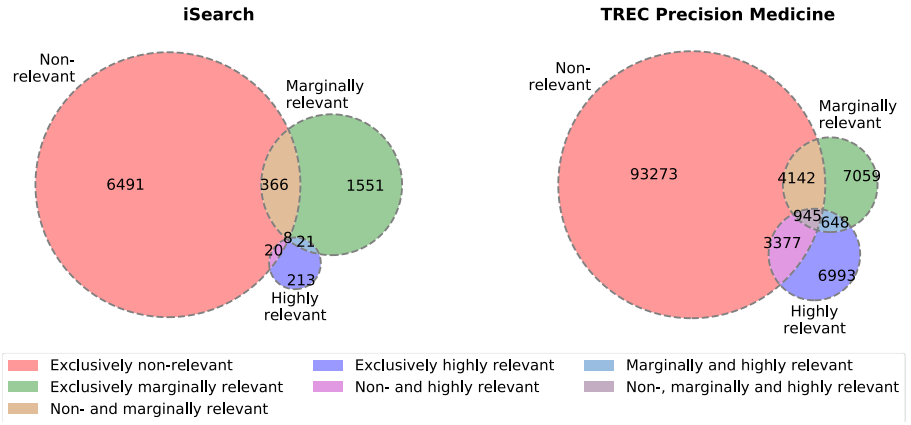

Fig. 2 Intersections between non-, marginally and highly relevant documents

topics to judge, whereas, in the case of TREC PM, the assessors were provided with predefined topics (or patient data). Second, when making the relevance judgments, iSearch assessors were also free to choose what constitutes their personal relevance. In contrast, TREC PM assessors had very clear criteria given by the patient data when judging the relevance of specific documents.

With regard to the set of judged documents in a test collection, some of the documents are judged multiple times across different topics. This leads to documents being judged with possibly different levels of relevance for several topics. Figure 2 shows set diagrams with relevance judgments of both test collections. For both, the set of documents judged exclusively non-relevant is the largest. For the sake of consistency, we combined *marginal* and *fair* to *marginally relevant* judgments of the iSearch collection. This is why the most prominent difference between the two set diagrams is the relative amount of documents exclusively judged marginally relevant. For iSearch, these account for approx. 17.9 %, whereas for TREC PM, these account for 6.1 % of all judged documents in the collection.

Matching relevance assessments with biblio- and altmetrics

For all documents of iSearch, the arXiv-ID is available. With the help of this identifier, we query the arXiv-API[3] and retrieve the Digital Object Identifier (DOI), if available, then iSearch and WOS data are matched via DOI. iSearch and altmetric data are matched via DOI or arXiv-ID. This means that we end up not having data from WOS or Altmetric Explorer for all iSearch documents because the iSearch documents may only be partly covered in the databases, or the original document does not have a DOI. For 69, 6 % of the rated documents, we were able to retrieve the DOI. In comparison, this coverage is slightly higher compared to that of documents that are not rated (66, 4 %).

From WOS, citation data is added to the iSearch collection. Also, more additional data is generated by matching the iSearch collection with WOS and altmetrics: via ISSN, a classification with the science classification scheme according to Archambault et al. (2011),

[3] https://arxiv.org/help/api.

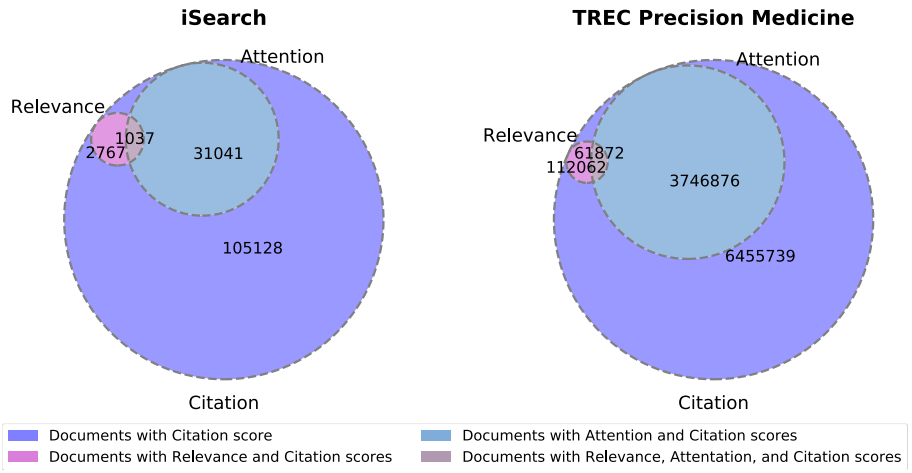

Fig. 3 Intersections between documents with Relevance, Attention, and Citations

Table 1 iSearch: 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. $CPP_{\neg RA}$ is the same for all documents without a relevance assessment

Category	Documents	RA	CPP_{RA}	$CPP_{\neg RA}$
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

as well as the Journal Impact Factor (JIF), and the Research Level are added. From the Altmetrics Explorer, the Altmetric Attention Score, as well as 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. The same is done for TREC PM, with the difference, that the documents in this collection have the PubMed-ID as an identifier. The WOS instance we use also contains a table that converts PubMed-IDs to WOS accession numbers.

Figure 3 provides two set diagrams illustrating the intersections between relevance judgments, citations, and altmetrics. The underlying data is limited to documents for which at least citation data is available. As can be seen, only citation data is available for most documents. For 0.7 % of iSearch and 0.6 % for TREC PM, respectively, all three indicators relevance judgments, altmetrics, and citations are available.

From the documents with DOI that were matched with WOS, the publications in iSearch can essentially be assigned to the four major categories *Nuclear & Particles Physics*, *Fluids & Plasmas*, *General Physics*, and *Astronomy & Astrophysics* which make out 87 % of all documents in the corpus (see Table 1). Furthermore, Table 1 illustrates the share of documents with relevance assessments as well as the citation rates for documents with and without relevance assessments.

Table 2 TREC PM: Percentage of documents per WOS category

Category	Documents	RA	CPP_{RA}	$CPP_{\neg RA}$
Clinical Medicine	48.3 %	73.4 %	42.2	19.6
Biomedical Research	15.9 %	15.4 %	49.1	27.0
Chemistry	7.5 %	1.3 %	22.0	30.7
General Science & Technology	4.7 %	3.1 %	152.3	36.5
Public Health & Health Services	4.6 %	2.1 %	23.9	14.1
Psychology & Cognitive Sciences	2.5 %	1.4 %	37.7	23.3
Agriculture, Fisheries & Forestry	2.0 %	1.6 %	15.9	14.0

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.

$CPP_{\neg RA}$ is the same for all documents without a relevance assessment

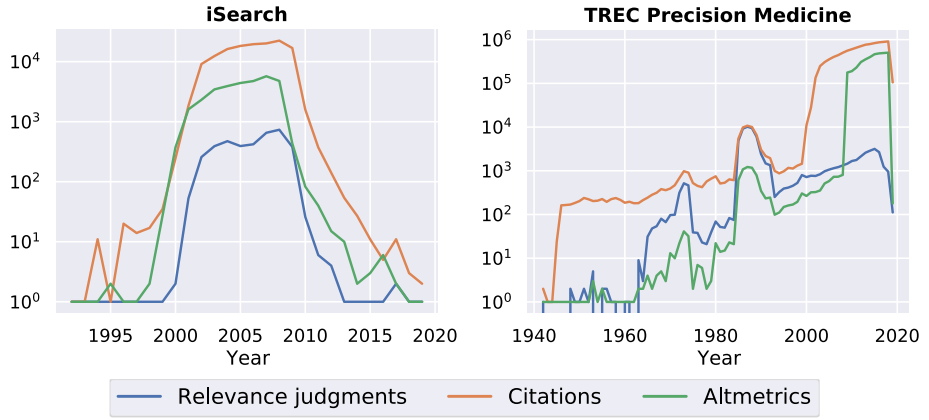

Fig. 4 Distributions of relevance judgments, citations, and altmetrics over the years for both test collection iSearch and TREC Precision Medicine

In TREC PM, two categories contain the most publications: *Clinical Medicine* with almost half of all documents and *Biomedical Research* with about 16 %. All other categories remain below 10 %. The picture is even clearer when it comes to the distribution of relevance assessments between these two categories: here, *Clinical Medicine* accounts for almost 75 % of all relevance assessments, while in *Biomedical Research*, the proportion of relevance assessments roughly corresponds to the proportion of document, as shown in Table 2. The distribution of relevance ratings by category indicates a small shift and shows that the category with the largest number of articles is not necessarily the category with the most relevance assessments.

The proportion of documents with a relevance assessment is different: In iSearch, most of the documents with a relevance assessment are in the categories *Fluids & Plasmas* and *General Physics*. These two categories together account for about 55 % of the documents with a relevance rating. *Nuclear & Particle Physics*, which accounts for one-third of the iSearch collection documents, accounts for only 10 % of the documents with relevance assessment in this category. A similar picture, i.e. a shift in the technical focus of the collection and the relevance assessed documents, can be seen with TREC PM: While *Clinical*

Table 3 Groups of relevance with corresponding Web of Science (WoS) data. For each relevance group, the number of documents (P), the sum of citations (Citations), the average citation rate (CPP), the average journal impact factor (JIF), and the average research level (RL) are included for both data collections: iSearch and TREC PM. The number of documents results from the availability of WOS data retrieved by DOI

Relevance	iSearch				TREC PM			
	P	CPP	JIF	RL	P	CPP	JIF	RL
Non-relevant	2926	52.8	4.2	3.7	98,781	38.9	4.7	2.4
Fair or Marginal	765	63.7	4.1	3.7	10,767	61.2	5.6	2.4
High	113	77.6	4.6	3.6	10,359	56.4	5.5	2.4
Not rated	136,169	41.2	4.3	3.8	18,805,400	27.4	4.2	2.5

Medicine accounts for about half of the documents in the collection, almost 75 % of the documents with relevance assessments are in this category.

Tables 1 and 2 also show the citations per paper (CPP) of the documents (with and without relevance assessment) in one category. In all but one case, the citation rates of documents with relevance assessment are higher than the citation rates of documents without relevance assessment. The only exception is *Chemistry* in the TREC PM: This may also be due to a bias caused by the fact that only 1.3 % of the documents in *Chemistry* have a relevance assessment, while the category represents 7.5 % of the documents in the collection.

Figure 4 shows the distribution of the publications for both search collections over publication years, seperated for publications with or without relevance assessments, and for documents with citation data from WOS and with data from Altmetrics Explorer. Please note, that both diagrams have a logarithmic scale. It is interesting to note the difference in the distribution of publications over publication years between documents with relevance assessments and those without relevance assessments in the two collections: While in the iSearch Collection, the distribution over time of documents with and without relevance assessment is quite similar, with publications limited to the period from 2000 to 2011 and peaking in 2008, the situation is quite different in TREC PM: The period covered by TREC PM is much longer than that of iSearch; it started back in the 1970s. The largest peak for documents with relevance rating is between 1984 and 1993, with a second peak in 2015, while documents without relevance rating do not increase significantly until 2002, and continue to do so until the end of the analysis in 2018.

Since there is no bibliographical information on the publication years for either data collection, the diagrams do not represent the complete collection, but rather the part of publications that could be matched with WOS and is listed there. The documents with altmetrics (green line) usually reproduce the course of the documents with relevance assessments, so that it can be said that documents with altmetrics are not underrepresented in the entire study, but always make up a decisive part of it. This is despite the fact, that the Altmetrics Explorer will only be fully available from around 2008. This shows very well that even publications from the longer past are represented in the Altmetrics Explorer to a not inconsiderable extent, even if they are already decades old because their content is still up-to-date and these publications are still being talked about in net media. From 2008 on, TREC PM will undergo a drastic change: Altmetrics Explorer is fully available from this point, which increases the number of documents with altmetrics in TREC PM to a level that includes almost all documents with citation data. This change is not visible in iSearch because the number of documents in this collection has already passed its peak at that time.

Table 4 Groups of relevance with corresponding altmetric data. For each relevance group, the mean of the Attention Score per Paper is given

Relevance	iSearch Attention Score / P	TREC PM Attention Score / P
Non-relevant	2.9	10.0
Fair or Marginal	2.9	13.2
High	8.8	11.4
Not rated	2.3	9.9

Results

In the following, we describe and analyze the data generated in Sect. Experimental evaluation data set. The results of the data analysis are presented here in detail and compared between the two search collections and assigned to the research questions. When we talk about citations in scientific publications and the Attention Score in altmetrics, both are forms of perception. In the following we will focus more on the perception of scientific literature (Ball and Tunger 2006; Glänzel 2008; Haustein 2011) and will outline a connection to papers with high relevance assessments.

Relations between relevance judgments and bibliometric & altmetric perception

In the following section we investigate the relationship between (positive) relevance assessments and the perception of documents. In Table 3 we see that the unrated documents (not rated), which form the vast majority, have a lower citation rate than the group of rated documents (non-relevant, fair, marginal, and high relevant). On average, the rated documents receive about 41 citations per document in iSearch and 27 in TREC PM. In iSearch, the highest citation rate is achieved by the highly relevant documents: With a citation rate of 77.6, they achieve a citation rate almost twice as high as documents without a relevance rating. In TREC PM, documents with a relevance assessment also achieve a citation rate twice as high as documents without any relevance assessment. There is no significant difference between documents with a high relevance assessment to documents with a relevance assessment that could be classified as fair or medium: Both citation rates are between 56.4 and 61.2 citations per paper.

Tables 1 and 2 show the citation rates for the five scientific fields, which together account for 90 % of the iSearch collection and the seven fields, that do the same for the TREC PM documents, also for those with and without relevance assessment. The citation rates for publications with a relevance assessment are higher for all categories shown than for publications without a relevance rating in the same category.

Are there differences in the composition of the groups that would explain the differences in citation rates described previously? The JIF does not show a difference for any of the groups. It only differs by fractions of a decimal point for iSearch: documents without relevance rating have an average JIF = 4.3, the documents with relevance rating have an average JIF between 4.1 and 4.6, non-relevant papers are between these values with JIF of 4.2. In TREC PM, JIF is only 1 citation per paper higher for documents with a fair (JIF =

Table 5 Number of relevance-rated documents (P) with Mendeley readership and citations within the WoS. The Pearson correlation (r) is calculated for the correlation between WoS citations and Mendeley readerships

Relevance	iSearch		TREC PM	
	P	r	P	r
Non-relevant	786	0.85	29,841	0.59
Fair or marginal	216	0.72	4820	0.85
High	35	0.89	5148	0.83
Not rated	32,081	0.83	3,808,747	0.63

5.6) or high (JIF = 5.5) relevance ranking instead to the documents that are non-relevant (JIF = 4.7) or do not have a relevance assessment (JIF = 4.2).

To investigate whether it is possible to transfer the relationship between relevance assessments and citations to altmetrics we can look at the individual altmetric scores (e.g. news items, tweets, Facebook posts, etc.). These scores are used very differently. It is better to use a summarized and appropriately weighted value, like the Altmetric Attention Score, in which all activities of a paper in the different document types are considered with an appropriate weighting.[4] Thus the value can be used like a number of citations, which is given in Table 4 as an average Attention Score per document in the respective relevance rating.

Looking at the result for iSearch, we see that unrated documents get a score of 2.3, quite similar to non-relevant or marginally relevant paper with 2.9. There is a very large gap to high relevant documents, which average at 8.8. The explanation for the large gap to the other relevance scores is the low number of documents to which this relevance score applies: These are 60 papers and one of these having an Attention Score of 338.

The situation is similar for TREC PM , although the level of the values is already higher, unrated or non relevant paper receive a value of 9.9 or 10.0. documents classified as marginally or highly relevant receive a significantly higher value in this search collection than non-relevant or unrated documents with 13.2 for marginally relevant paper and 11.4 for highly relevant paper. Relevant papers therefore have a higher rating than non-relevant ones, even if the lower rating level achieves the higher value.

RQ1 has to be answered as follows: On average, documents with a relevance rating receive a higher number of citations and also a higher Attention Score in the Altmetrics Explorer.

Correlation between relevance judgments, citations, and altmetrics

The majority of the publications were published a long time before the Altmetrics Explorer has been put online. A large part of the time, social media has not been used in society, and certainly not in science. This changed only slowly towards 2008, with the Altmetrics Explorer being founded by Euan Adie in 2011. It is not known to what extent publications prior to the founding year of Altmetrics Explorer were retroactively re-indexed and to what extent this is technically possible at all. For our given collections we can see in Fig. 4, that also before 2011, publications that are decades old are covered by Altmetrics Explorer today. So, we can cover the altmetrics and attention during this time.

[4] https://help.altmetric.com/support/solutions/articles/6000233311.

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, document 10.1103/PhysRevE.67.026126, which receives 1068 citations and 1192 Mendeley reads. There are also examples of unequal perception: document 10.1088/0067-0049/182/2/543 has 3151 citations but is bookmarked "only" 405 times on Mendeley. Document 10.1088/0954-3899/33/1/001 on the other hand has 3903 citations but is bookmarked only 24 times, or in the opposite direction document 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.

In Table 5 we report the results of a Pearson correlation analysis between WOS citations and Mendeley readerships. We only consider documents that were relevance-rated and have WOS citations and Mendeley readerships and we differentiate between the three relevance ratings non-relevant, fair or marginal, highly relevant, and the unrated documents. This results in a total of 1037 documents for iSearch and 39,809 for TREC PM.

For iSearch 786 publications were rated non-relevant, which leaves only a small set of 251 documents with a positive relevance ranking. For the group of 35 publications, which were rated highly relevant, we see a high correlation of $r = 0.89$. The second highest correlation is for the non-relevant publications with a Pearson correlation of $r = 0.85$. The lowest correlation is for the group rated fair or marginal with a correlation of $r = 0.72$. The control group of unrated documents gets a comparable high correlation with $r = 0.83$. These values for iSearch should be used carefully as we see a number of outliers that have a high impact due to the small sample size. If we evaluate TREC PM, we get more valid results, since they are based on a much higher number of cases compared to iSearch. Unrated documents with $r = 0.63$ and non-relevant documents with $r = 0.59$ get a much lower correlation between citation data and Mendeley than the documents from the other two classes with a marginal or high rated relevance: These documents receive a Pearson correlation of $r = 0.85$ and $r = 0.83$, respectively, which can be considered identical.

Overall, for **RQ2** it could be shown, that there is a correlation between citations and altmetric counts in the two collections, as also shown by Holmberg et al. (2019). This is especially true for the TREC PM collection if they are part of a set of documents that have a positive relevance rating and if we compare against the control group of unrated documents.

Comparison of iSearch and TREC PM

We have applied the same evaluations to two test collections, which have very different thematic focuses: While iSearch is limited to topics in physics, TREC PM is focused on medicine. The publications vary in size, age, and the development over time, as well as the ratio of unrated documents to relevance-rated documents, are completely different.

Previous research questions addressed measurable effects between relevance-assessed documents on the one hand and citation or altmetrics data on the other hand in research collections, and whether these effects are more random or whether the results are repeated in different contexts. We did not measure this across the entire collection, but separately for individual fields to which we assigned the publications using Archambault's classification (see Tables 1 and 2). With one exception (Chemistry in TREC PM), this applies to all relevant fields of both collections.

We then looked at the citation rates (CPP) in the individual relevance classes, and found for both collections that generally relevance-rated documents achieve a higher citation rate

than unrated ones. The higher the relevance rating, i.e. the more relevant a document is from the viewpoint of the assessors, the higher the citation rate. The JIF and the research level are almost completely unchanged in Table 3. On the contrary, this means that it is not necessary to search for documents in journals with a higher JIF to find a relevant document. Thus, the impact of a journal does not play a role in the relevance rating. The research level is also completely independent of it. In both test collections, the research level is almost unchanged at the same level for all relevance classes in the respective research collection.

Regarding the Attention Score from the Altmetrics Explorer, broken down to the individual relevance classes, the same picture emerges: relevance-rated documents receive more attention than non-rated documents and that this increases with higher Attention Scores.

The final analysis was to calculate the Pearson correlation with regard to the connection between citations and Mendeley readerships whether a difference in the individual relevance classes is also discernible here. This means that documents with a (higher) relevance score are more likely to have a higher citation score at the same time, coupled with a higher number of bookmarks on Mendeley for this publication. Thus, **RQ3** on the question whether measured effects in both test collections are comparable, can be answered with yes. We still should be aware that the relationships shown are unidirectional: A paper with a high relevance assessment usually also has a high number of citations or a high perception in altmetrics. Not vice versa.

Discussion

At the beginning of the paper, we asked whether there is a connection between data that has received a relevance assessment in a socio-cognitive process and citations whose relevance is more likely to be expressed in a cumulative process. To find out whether this relationship exists, we enriched data from two test collections whose publications contain relevance ratings with citations. In this way, we merged the two previously described characteristics of relevance and combine both in one dataset. From this, we performed analyses that allowed us to answer the three research questions. Against this background, we now would like to discuss and interpret these results.

Relevance and scientifc communication

We measured the co-occurence of relevance assessments and citations/mentions for scientific documents in the context of IR test collections. We can clearly state from this analysis that documents that receive a relevance rating are more likely to also be highly cited. This is an unidirectional connection as we don't see any evidence that this connection is bidirectional that would allow to draw the inverted conclusion. This is in line with previous results from Larsen (2004), who clearly showed that citation data is no proxy or substitution for relevance in an actual retrieval process and that the expected "boomerang effect" for retrieval purposes, while at first promising, was in the end only hardly operationalizable. Just presenting highly cited documents in a retrieval process would not retrieve a reasonable amount of (topically) relevant documents. This might be the reason that popular academic search engines like Google Scholar incorporate (among others) citation-based and text/topic-related ranking factors in a combined way (Beel and Gipp 2009) and not solely rely on citation counts only to rank the result lists.

If we look beyond the individual perception of a document, the JIF constitutes as an agglomerating score of perception. We could not measure any substantial 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. This means, that the impact of a journal has no influence on the decision about the relevance of a document. When looking at the JIF, Kacem and Mayr (2018) 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. Searching and judging are two different things.

The results of our study on the intersection of relevance assessments within the iSearch collection, TREC PM, and corresponding citation counts show that explicit relevance decisions of a single assessor and implicit 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 (compare with the discussion above about the meaning of a citation). Ingwersen (2012) 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 (Dabrowska and Larsen 2015) but not as a substitute for topical content. 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.

This argument has to be further deconstructed, as the analysis of citation practices show some markers that support the deep relations that lie beneath the concept of popularity. Shadish et al. (1995) analyzed citations in psychology journals and focus on the judgments by the authors who cite them. The three dimensions that influenced the judgments most were quality, time aspects, and creativity of the cited works. They found that in general, highly cited scholarly works are rated as exemplars and as being of higher quality, although there were differences between older and newer works in these ratings. Another factor was the creativity of the cited work. While some works rated as highly creative gathered a lot of citations some creative works fit poorly into existing conceptual or methodological structures, and so are used less. The question why some works are cited more and gain more popularity and perception than others is still unsolved (Cronin 1981; Leydesdorff 1998). Linked to this is the fundamental question of why the currency in a reward system (citations or social media mentions in our case) generally leads to a skewed distribution, as this is a fundamental observation that applies to all reward systems. With regard to publications, the relevance to one's own scientific work is one approach to why some publications are preferred cited.

These three aspects are intuitive factors that might also influence a relevance decision in a retrieval task but they are for sure not the only factors. This is also true for relevance decisions and while relevance is often described as multidimensional or layered another concept is aligned to the different levels and forms of relevance. This might be the principle of polyrepresentation (Ingwersen 1996; Skov et al. 2008). Polyrepresentation might be the common ground where the general popularity of a document measured through citations and the concrete relevance come together as it hypothesises that overlaps between different cognitive representations of the information need as well as documents can be exploited for reducing the uncertainties in IR (Larsen and Ingwersen 2002).

All in all we have to record the fact that topical relevance is the easiest to assess. Matching keywords or using checklist-based assessment criteria like in TREC PM is something that even non-experts like crowd workers can achieve. Getting an understanding for the Œuvre of a scientific work is something different and needs a lot of experience, expertise and domain knowledge. From an IR perspective other relevance features that might be employed to make use of the later are desirable but are most often not available in practice. This is not the case with our two created test collections. In iSearch we have subjective relevance assessments from the original people who designed the topics. The whole context of the corresponding topic, the background knowledge and expertise of the assessor are integrated in the relevance judgments. The iSearch assessor also described their task and context and let us reconstruct their current situation and information need. They are therefore not as (strictly) objective as the assessors of TREC PM where a very objectified way of document assessment was chosen. Despite the differences in the relevance assessment process both collections show comparable effects and results – across different scientific domains. The relations between relevance assessments and perception of documents is clearly visible in both collections, although we only based our investigations on the small subset of judged documents in iSearch (2 % of total collection) and TREC PM (0.4 % respectively).

Citations and altmetrics

In the literature, the question of whether there is a correlation between citations of scientific publications in WOS, Scopus, or Google Scholar to Mendeley readerships has often been examined. Li and Thelwall (2012) 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 a publication received in one of these databases and the number of bookmarks of this publication on Mendeley. The result of this investigation was a perfect correlation between the citation counts from the three citation databases WOS, Scopus, and Google Scholar. This result is also less surprising because WOS and Scopus are about 95 % overlapping, and there is also a big overlap between these two databases and Google Scholar. A correlation between citations from the three mentioned citation databases to Mendeley was also measurable by Li and Thelwall (2012), but it is much worse than the correlation between the citation databases. As mentioned above, the correlation of Mendeley readerships and citations will be stable at 0.6 over some years (Maflahi and Thelwall 2016). This is not surprising since Mendeley readerships are not scientific citations. Bookmarking of publications takes place for other reasons than citing a scientific paper. Costas et al. (2015) found out, that anyway "Mendeley is the strongest social media source with similar characteristics to citations in terms of their distribution across fields". Zahedi et al. (2017) emphasizes "the ability of Mendeley readership to identify highly cited publications and its role as a potential evaluative tool".

In a study on the citation behavior of publications, where the question is answered about the ideal citation window for bibliometric indicators, it could be shown, that publications that were already heavily cited at the beginning are also heavily cited in the maturing process up to the peak of citation, and publications that are cited infrequently receive relatively little attention in later periods (Clermont et al. 2021). This means, that the author of a paper is making a decision about the relevance of the cited literature, that e.g. can be reflected in bibliometric or altmetric indicators. An author would not cite a paper, that is not relevant for his or her work, if we assume that an

author is free in his or her decision about a citation. Since an author can only consider relevant publications in his reference list of cited publications, those ones that he can find with common effort, e.g. that are listed in a database such as the WOS, have an advantage: finding a publication in this database requires not such a high effort than in less visible sources. This explains, why it is a big advantage, if a paper is listed in highly visible databases: If a relevant paper would have no visibility, the effort would be too high, to identify this paper as relevant. This means, that it is not enough to have relevant information for somebody, this information also has to be as visible as possible to be found and considered as relevant to be used. In science, the use of information is marked with a citation. In the end, it is not a single citation that is important, but the aggregate of them: Citation indicators are very stable after a few years and significant changes, e.g. in rankings, are very seldom after this time (Clermont et al. 2021). The same applies to altmetrics: a paper is bookmarked at Mendeley because it has a certain relevance to the own work, because you want to remember it for a later point of time or because you think it is so relevant that you want to share it with others, for whose work it could be equally relevant.

Even if a citation is made for other reasons than a bookmark, there is a permanent correlation between both: 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, but it also is a kind of relevance assessment. Regarding this connection, 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 WOS. This follows the goal of altmetrics also to provide new information and not just to be faster bibliometrics.

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 altmetrics scores due to the lower social web use when they were published" (Thelwall et al. 2013), is a question that is legitimate but not the focus of this publication.

It can be seen that documents with relevance assessment achieve a higher average perception in altmetrics than documents without relevance assessment. It also shows that the publications of the iSearch collection on average have a lower perception in altmetrics than publications in TREC PM. This is also in accordance with expectations and corresponds to previous studies and literature (Tunger et al. 2017), which show that life science topics are significantly higher to appear in social media than engineering or physics topics.

The correlation of Mendeley readership and citations differ between highly relevant documents and less or irrelevant documents. If we assume that a citation of a publication, the bookmarking of a publication on Mendeley or the intellectual relevance rating of a paper are all manifestations of a relevance rating, then it can be said that the higher the relevance of a paper, the more extreme is the manifestation of the relevance statement in all three data sources simultaneously. This results in a reduced scattering as could be shown and explains, why the correlation between Mendeley readerships and citations is stronger for documents with a high relevance assessment.

Another observation is the common appearance of power-law distributions in the scientific publication context. We found out, that these result from a process of

maximizing relevance. We only see the visible entities: References and relevance judgments are the measurable units. In many different applications, precisely this form of distribution is used because it is skewed and can therefore contribute much more easily to the grouping of classes than if it were normally distributed. Thus many applications of bibliometrics and relevance assessment only become possible through a skewed distribution.

Conclusion

High relevance ratings for scientific documents correlate with high citation counts and altmetrics Attention Scores. This pattern is clearly visible in the two test collections, we created for this study. It tends to be the same entities that have highly relevant or highly cited publications or whose publications achieve high perceptions in altmetrics. The frequency and regularity with which this happens is striking and leads to the conclusion that this goes far beyond random events and we are eager to conclude that none of this can be a coincidence. We should be aware of the fact that by citing a scientific publication, the author is making a decision about its relevance that can be reflected in relevance assessments, bibliometric indicators or altmetrics. All three are based on decisions about the relevance of a publication to a certain topic. This means that once a publication is relevant to a topic, it remains so. It could not be explained otherwise that there is a permanent correlation between publications with intellectual relevance assessment and biblio- or altmetric indicators.

We observe that it is not only the impact of the corresponding journal, the institution of an author is working for or the author himself that are responsible for the impact of a publication: Popularity alone seems not to explain this effect, as shown in Table 3. Otherwise, uncited papers from high-impact journals like Nature or Science would be highly unlikely – which is not the case. We suggest to understand citations or altmetric attention in relation to relevance assessments as an implicit marker for "quality", although we are aware that this term is highly controversial in the bibliometrics community, but used anyway (Knorr-Cetina 1981). Ashford (1986) explains that "quality" does not always mean the best possible source, but that e.g. "decision makers will use accessible sources rather than other sources providing higher quality information because of the cost in effort involved in seeking out those more informative sources."

The aforementioned work by White (2011) was mainly motivated by theoretical concepts and only included a small-scale evaluation. Likewise, previous work on polyrepresentation by Ingwersen (1996) has analyzed the relations between relevance and the corresponding implicit signals from a theoretical point of view and has only been picked-up in few empirical studies, like Skov et al. (2008) or Larsen and Ingwersen (2002). In this sense, the impact of this work is that we could show a relation between explicit relevance assessments and implicit relevance signals originating from biblio- or altmetric measures like citations or bookmarks on a large-scale analysis. It could be shown that these correlations appear in different collections, regardless to size of the data set or scientific domain. Thus it can be assumed that the results are not purely random, but that they are more profound.

If both explicit and implicit relevance assessments are related and correlate with each other, then the same facts are described with different characteristics, i.e. if these are only different sides of the same coin, then we can speak of polyrepresentation at this point. We still need to evaluate and investigate the reasons for the relationship we have seen.

The brief framework introduced in Sect. 3 is a first step in this direction that still requires a more elaborated understanding of the interconnections between the different concepts of relevance, what will be done as part of our future work. This work, however, has a focus on the visible entities of relevance and the empirical observations that can be derived from their correlations and interactions. The principle of polyrepresentation 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, e.g. based on very recent collections like TREC-COVID.

In this publication, we have dealt with the question of whether relevance assessments and citations are connected on a theoretical and empirical level. We have been able to show correlations between publications with a high relevance rating, a high citation count and a high perception in altmetrics. We could show that they are indeed two sides of the same coin and that the relevance of publications can have different characteristics and we are thus dealing with polyrepresentation. From a conceptional perspective, we only argued with the visible entities. What is still missing is an understanding of the underlying processes: When and why do we perceive a publication as relevant? What influence do individual factors have on the citation behavior of scientists? These further questions still need to be clarified in the future; relevance theory is a crucial basis for this. In future work we will use our layered concept model to design experiments along these layers and to sharpen the expensiveness of analysis and discussion.[5]

Acknowledgements This study was funded by a grant from the German Federal Ministry of Education and Research (BMBF), grant number 16IFI107 (UseAltMe), and partly by DFG (German Research Foundation) project no. 407518790. We would like to thank the Competence Center for Bibliometrics for the access to WOS. The center was founded in 2008 to allow the scientific community to perform comprehensive and reliable bibliometric analyses by providing adequate data infrastructure and is funded by the German Federal Ministry of Education and Research (BMBF) under reference no. 01PQ17001.

Author Contributions Philipp Schaer and Dirk Tunger designed the research. Dirk Tunger retrieved the alt- and bibliometrics data, processed and analyzed the data. Timo Breuer contributed visualizations in the form of Venn diagrams. Philipp Schaer and Dirk Tunger wrote down the introduction, related work, and framework design. They analyzed the results and wrote the discussion. Timo Breuer described the existing test collections, whereas Dirk Tunger provided descriptions of the enriched datasets. All authors provided input for writing and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The data generated for this study and the scripts of the experiments are available in a public Zenodo archive at https://doi.org/10.5281/zenodo.5883400.

[5] http://www.bibliometrie.info.

References

Archambault É, Beauchesne, O.H., Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In *Proc. of ISSI 2011, Durban South Africa*, pp. 66–77.

Ashford SJ (1986) Feedback-Seeking in Individual Adaptation: A Resources Perspective. Academy of Management Journal 29(3):465–487. https://doi.org/10.2307/256219.

Ball, R., & Tunger, D. (2006). Bibliometric analysis - A new business area for information professionals in libraries? *Scientometrics, 66*(3), 561–577.

Beel, J., Gipp, B. (2009). Google Scholar's Ranking Algorithm: An Introductory Overview. In: Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09), International Society for Scientometrics and Informetrics, vol 1, pp. 230–241.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology, 54*(10), 913–925. https://doi.org/10.1002/asi.10286

Breuer, T., Schaer, P., Tunger, D. (2020). Relations between relevance assessments, bibliometrics and altmetrics. In: Cabanac G, Frommholz I, Mayr P (eds) Proceedings of the 10th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 42nd European Conference on Information Retrieval, BIR@ECIR 2020, Lisbon, Portugal, April 14th, 2020 [online only], CEUR-WS.org, CEUR Workshop Proceedings, vol 2591, pp 101–112. http://ceur-ws.org/Vol-2591/paper-10.pdf.

Butler, J. S., Kaye, I. D., Sebastian, A. S., Wagner, S. C., Morrissey, P. B., Schroeder, G. D., Kepler, C. K., & Vaccaro, A. R. (2017). The evolution of current research impact metrics: from bibliometrics to altmetrics? *Clinical spine surgery, 30*(5), 226–228.

Carevic, Z., Schaer, P. (2014). On the Connection Between Citation-based and Topical Relevance Ranking: Results of a Pretest using iSearch. In: Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval, Amsterdam, The Netherlands, CEUR Workshop Proceedings, vol 1145, pp 37–44. http://ceur-ws.org/Vol-1143/paper5.pdf.

Clermont, M., Krolak, J., Tunger, D. (2021). Does the citation period have any effect on the informative value of selected citation indicators in research evaluations? Scientometrics Submitted for publication.

Cole, C. (2011). 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. https://doi.org/10.1002/asi.21541

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Inf Process Manag, 36*(4), 533–550.

Costas, R., Zahedi, Z., & Wouters, P. (2015). 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. https://doi.org/10.1108/AJIM-12-2014-0173

Cronin, B. (1981). The need for a theory of citing. *Journal of Documentation, 37*(1), 16–24.

Dabrowska, A., Larsen, B. (2015). 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.

Fisher KE, Naumer, C.M. (2006). Information Grounds: Theoretical Basis and Empirical Findings on Information Flow in Social Settings. In: Spink A, Cole C (eds) New Directions in Human Information Behavior, vol 8, Springer-Verlag, Berlin/Heidelberg, pp 93–111, 10.1007/1-4020-3670-1_6, http://link.springer.com/10.1007/1-4020-3670-1_6, series Title: Information Science and Knowledge Management.

Garfield, E. (1964) ."Science Citation Index"–A New Dimension in Indexing. Science 144(3619):649–654, 10.1126/science.144.3619.649. https://www.sciencemag.org/lookup/ https://doi.org/10.1126/science.144.3619.649.

Garfield, E. (1966). Patent citation indexing and the notions of novelty, similarity, and relevance. *Journal of chemical documentation, 6*(2), 63–65.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471–479. https://doi.org/10.1126/science.178.4060.471

Garfield, E. (1996). When to Cite. The Library Quarterly: Information, Community, Policy 66(4):449–458. http://www.jstor.org/stable/4309157, publisher: University of Chicago Press.

Garfield, E. (1998). Random thoughts on citationology its theory and practice. *Scientometrics, 43*(1), 69–76.

Glänzel, W. (2008). Seven Myths in Bibliometrics About facts and fiction in quantitative science studies. *Collnet Journal of Scientometrics and Information Management, 2*(1), 9–17.

Haustein, S. (2011). Taking a multidimensional approach toward journal evaluation. In: Noyons E, Ngulube P, Leta J (eds) 13th Conference of the International-Society-for-Scientometrics-and-Informetrics (ISSI), pp. 280–291.

Heck, T., Schaer, P. (2013). Performing Informetric Analysis on Information Retrieval Test Collections: Preliminary Experiments in the Physics Domain. In: Proc. of ISSI 2013, Vienna, Austria, vol 2, pp. 1392–1400.

Holmberg, K., Bowman, T., Didegah, F., & Lehtimäki, J. (2019). The Relationship Between Institutional Factors, Citation and Altmetric Counts of Publications from Finnish Universities. *Journal of Altmetrics, 2*(1), 5.

Ingwersen, P. (1992). Information retrieval interaction. Taylor Graham, London, oCLC: 832020263.

Ingwersen P (1994) Polyrepresentation of Information Needs and Semantic Entities Elements of a Cognitive Theory for Information Retrieval Interaction. In: Croft BW, van Rijsbergen CJ (eds) SIGIR '94, Springer London, London, pp 101–110.

Ingwersen, P. (1996). Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. *Journal of Documentation, 52*(1), 3–50. https://doi.org/10.1108/eb026960

Ingwersen, P. (2012). Bibliometrics/Scientometrics and IR – A methodological bridge through visualization. http://www.promise-noe.eu/documents/10156/028a48d8-4ba8-463c-acbc-db75db67ea4d, presentation

Kacem, A., Mayr, P. (2018). 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.

Knorr-Cetina, K. (1981). The manufacture of knowledge: an essay on the constructivist and contextual nature of science. Pergamon international library of science, technology, engineering, and social studies, Pergamon Press, Oxford ; New York.

Langham, T. (1995). Consistency in Referencing. *Journal of Documentation, 51*(4), 360–369.

Larsen, B. (2004). References and citations in automatic indexing and retrieval systems - experiments with the boomerang effect. *PhD Thesis, Department of Information Studies, Royal School of Library and Information Science*, Copenhagen, Denmark, http://pure.iva.dk/files/31034810/birger_larsen_phd.pdf.

Larsen, B., Ingwersen, P. (2002). The boomerang effect: retrieving scientific documents via the network of references and citations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press*, pp. 397–398.

Larsen, B., Ingwersen, P., Kekäläinen, J. (2006). The polyrepresentation continuum in IR. In *Proceedings of the 1st international conference on Information interaction in context - IIiX, ACM, New York, NY, USA*, pp 88–96, 10.1145/1164820.1164840, http://portal.acm.org/citation.cfm?doid=1164820.1164840.

Leydesdorff, L. (1998). Theories of citation? *Scientometrics, 43*(1), 5–25.

Li, X., Thelwall, M. (2012). F1000, mendeley and traditional bibliometric indicators. In *Proceedings of the 17th International Conference on Science and Technology Indicators*, pp. 541–551.

Lykke, M., Larsen, B., Lund, H., Ingwersen, P. (2010). 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, 10.1007/978-3-642-12275-0_63.

MacKenzie, I. (2002). Paradigms of reading: Relevance theory and deconstruction. Palgrave Macmillan, New York. http://www.dawsonera.com/depp/reader/protected/external/AbstractView/S9780230503984, oCLC: 312746582.

Maflahi, N., & Thelwall, M. (2016). When are readership counts as useful as citation counts? S copus versus m endeley for lis journals. *Journal of the Association for information Science and Technology, 67*(1), 191–199.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48*(9), 810–832.

Moed, H. F. (2005). Citation analysis in research evaluation. Information Science and Knowledge Management, vol. 9. Springer-Verlag, Berlin/Heidelberg. 10.1007/1-4020-3714-7. http://link.springer.com/10.1007/1-4020-3714-7.

Mutschke, P., Mayr, P., Schaer, P., & Sure, Y. (2011). Science models as value-added services for scholarly information systems. *Scientometrics, 89*(1), 349–364.

Raan van, A. F. (2005). Measuring science. In Moed HF, Glänzel W, Schmoch U (eds) Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems. Springer Netherlands, Dordrecht, (pp. 19–50). 10.1007/1-4020-2755-9_2. https://doi.org/10.1007/1-4020-2755-9_2.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., Lazar, A. J., Pant, S., Meric-Bernstam, F. (2019). Overview of the TREC 2019 precision medicine track. In Voorhees EM, Ellis A (Eds.) Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, 13–15 Nov, 2019. National Institute of Standards and Technology

(NIST), NIST Special Publication, vol. 1250. https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.PM.pdf.

Shadish WR, Tolliver D, Gray M, Sen Gupta SK (1995) Author Judgements about Works They Cite: Three Studies from Psychology Journals. Social Studies of Science 25(3):477–498, 10.1177/030631295025003003, http://journals.sagepub.com/doi/10.1177/030631295025003003.

Skov, M., Larsen, B., & Ingwersen, P. (2008). Inter and intra-document contexts applied in polyrepresentation for best match IR. *Information Processing & Management, 44*(5), 1673–1683. https://doi.org/10.1016/j.ipm.2008.05.006

Sperber, D., & Wilson, D. (2001). *Relevance: communication and cognition* (2nd ed.). Cambridge, MA: Blackwell Publishers, Oxford.

Taylor RS (1968) Question-Negotiation and Information Seeking in Libraries. College & Research Libraries 29(3):178–194, 10.5860/crl_29_03_178, http://crl.acrl.org/index.php/crl/article/view/12027.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE, 8*(5), e64841. https://doi.org/10.1371/journal.pone.0064841

Tunger D, Meier A, Hartmann D (2017) Altmetrics Feasibility Study. Tech. Rep. BMBF 421-47025-3/2, Forschungszentrum Jülich, http://hdl.handle.net/2128/19648.

Tunger D, Clermont M, Meier A (2018) Altmetrics: State of the art and a look into the future. In: Scientometrics, InTech, pp 123–134

Voorhees, E. M. (2007). TREC: Continuing information retrieval's tradition of experimentation. *Communications of the ACM, 50*(11), 51. https://doi.org/10.1145/1297797.1297822

White, H. D. (2011). Relevance theory and citations. *Journal of Pragmatics, 43*(14), 3345–3361. https://doi.org/10.1016/j.pragma.2011.07.005

White HD (2017) Relevance theory and distributions of judgments in document retrieval. Information Processing & Management 53(5):1080–1102, 10.1016/j.ipm.2017.02.010, https://linkinghub.elsevier.com/retrieve/pii/S0306457316307130

Wilson, D., & Sperber, D. (2004). Relevance Theory. In G. Ward (Ed.), *Horn L* (pp. 607–632). Blackwell: The Handbook of Pragmatics.

Zahedi, Z., Costas, R., & Wouters, P. (2017). Mendeley readership as a filtering tool to identify highly cited publications. *Journal of the Association for Information Science and Technology, 68*(10), 2511–2521.