On the Connection Between Citation-based and Topical Relevance Ranking: Results of a Pretest using iSearch

Zeljko Carevic and Philipp Schaer

GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, 50667 Cologne, Germany
{firstname.lastname}@gesis.org

Abstract. In Information Retrieval a lot of work and effort was put into the exploitation of textual information. Alternative approaches like citation search strategies make use of structural information that can be found in digital libraries or scientific information systems. One dataset containing structural information and everything needed for an IR evaluation is the iSearch test collection. Given this test collection we investigate its suitability with a special interest in bibliometric methods like citation analysis. By using co-citation recommendations we look for a topical connection between the seed document (which is known to be relevant to the given topic) and the recommended documents.

Keywords: Co-citation, Information Retrieval, Topical Relevance

1 Introduction

Traditional Information Retrieval (IR) research systems have their shortcomings as shown by Buckley [2]. While in recent years a lot of work and effort was put into the exploitation of textual information i.e. in the form of full text analysis or natural language processing a lot of structural information is available for scientific documents. Digital libraries and scientific information systems hold many of these structural information like citation and reference information. To complement or even surpass the previously mentioned shortcomings alternative IR approaches were suggested. One of these alternatives are so-called citation search strategies that typically require the user to supply a known document as starting point for the citation search. In this search strategy similar documents to the given and relevant seed document are retrieved with the help of co-citation analysis. The underlying idea is that there is some kind of semantic relationship between the citing and the cited documents or authors that can be exploited [7].

This is in line with one of the original ideas behind the Science Citation Index that was introduced to help improve retrieval performance. As shown by Pao [6] performance can be significantly raised by including citation information into the retrieval system. In the study by Pao citation searching added an average of 24% recall to traditional subject retrieval. In today's scientific retrieval systems like Google Scholar citations are primarily used to influence the document ranking as shown by Beel [1].

Although first results were promising and real-world systems like Google Scholar already exist the scientific community lacks a large and robust IR evaluation corpus to thoroughly investigate the influences of citations for scientific information retrieval. In 2010 Lykke et al. [4] presented the iSearch test collection that contains everything needed for an IR retrieval evaluation: a document corpus of 453,254 documents, a list of 65 topics and 200 graded relevance assessments per topic and additionally more than 3.7 million internal references.

Given this new test collection we want to test the suitability of iSearch as a test collection for IR research with a special interest in bibliometric methods like citation analysis. We think that it is a valid approach to have a further look on the usage of citation data within the iSearch test collection by posing the following questions:

- Can we apply citation analysis techniques like co-citation analysis on the iSearch collection?
- It is generally known that by using co-citation analysis we can find semantically similar documents to a given seed document. By combining this information about semantic relatedness/similarity with the test collections information about document relevancy to a given topic, can we find any significant overlaps? Are co-cited documents relevant to the underlying topic that the seed document comes up from?

This paper is meant as a proof of concept paper to the general feasibility and to learn more about the dynamics and features of iSearch. In this study we try to reproduce some of the previously described experiments by White and to apply them on the iSearch data set that is, to the best of our knowledge, the only IR test collection that includes citation data. The last research question is a good example for a question that can only be quantitatively analyzed and answered with a thoroughly designed IR test collection. By matching co-cited documents with the given relevance assessment we hope to rate the semantic relatedness of the seed document to the co-cited documents.

2 Materials and Methods

Of course other data sets with scientific documents and citation information exist (like metadata from PubMed) but these corpora lack a given set of topics and the needed relevance assessments. Using these corpora for an experimental setup would require to manually evaluate every single result generated by the co-citation ranking. Another IR test collection comparable to iSearch is Datacite [3] which was collected from CiteULike and CiteSeerX. Compared to iSearch it is small and contains 81,433 articles. We did an initial literature study analyzing 23 papers that used or mentioned the iSearch collection. The list was compiled by looking at all articles listed in Google Scholar that cite the central iSearch paper by Lykke et al. After data cleaning and removing duplicates and presentation slides 17 papers remain from which only one paper by Norozi et al. [5] actually made use of the available citation data in iSearch. Compared to this the DataCite collection was cited 6 times and only used once for retrieval experiments.

Table 1. Example of available citation data within iSearch.

ID	1
arXivURI	http://arxiv.org/abs/0704.2164
refNum	1
internalRef	http://arxiv.org/abs/hep-ph/9304232
externalRef	null
authors	J.Levelt:P.J.Mulders
fullRefText	[] J. Levelt and P.J. Mulders, Phys. Rev. D 49, 96 (1994) [arXiv:hep-ph/9304232].
journal	PHYS.REV.D
year	1994

2.1 The iSearch test collection

The iSearch test collection consists of the three standard parts of an IR test collection: (1) a corpus of documents, (2) a set of topics, and (3) relevance assessments. The corpus consists of monographic records that were extracted from the Danish National Library and full text and metadata sets that were crawled from the arXiv.org open-access/preprint repository. The set of 65 topics and their relevance assessments (~200 per topic) were extracted from 23 lecturers, PhD and MSc students from three physics departments. The collection additionally contains 12,727,716 references (33.6 per paper) from which 3,768,410 are linked to iSearch.

A record contains a unique identifier (ID) as well as an URI (arXivURI) pointing to the online resource on arXiv. Each record has up to several corresponding citations that are numbered by a running counter (refNum) and an URI pointing either to an internal (internalRef) resource within the citation dataset or an external resource (externalRef). Furthermore, each citation comprises information about the authors (authors), publication year (year), and publication venue or journal (journal). All these information were extracted from the actual reference text (fullRefText). For an example containing one citation see table 1.

2.2 Implementing Co-Citation Analysis Methods using iSearch

Reference data can be used to apply bibliographic co-citation analysis which is a popular similarity measure used to establish a subject similarity between two documents. If documents A and B are both cited by document C, they may be related to one another, even though they don't directly reference each other. If A and B are both cited by many other items, they have a stronger relationship. The more items they are cited by, the stronger their relationship is. Co-citation was first proposed in the fields of citation analysis and bibliometrics as a fundamental metric to characterize the similarity between documents.

In 2010 White [7] describes a novel approach for retrieving documents related to a seed by combining methods from bibliometrics and information retrieval. Usually the well-known tf*idf formula is used to rank documents by their degree of relevance to a query [5]. By replacing the query (commonly a term) with author names or document

Table 2. Results for sample seed

ID	Category	Title	tf	df	log_tf	log_df	tf*idf
0704.1800	hep-ph	Phenomology with Massive Neutrinos	58	58	1.76	3.23	5.70
0606054	hep-ph	Neutrino masses and mixings and..	16	121	1.20	2.91	3.51
0706.0399	hep-ph	Confusing Sterile Neutrinos with Deviation from Tribimaximal Mixing at Neutrino Telescopes	7	9	0.84	4.04	3.41
0704.1500	hep-ph	A Search for Electron Neutrino Appearance at the Delta m**2 ~ 1 eV**2 Scale	10	49	1.0	3.30	3.30

titles and using bibliometric analysis techniques like document co-citation or author co-citation White proposes that it can be used to predict semantic relatedness. This relatedness would not express the similarity between a query term and document (like in the IR model) but the semantic distance/relatedness between the seed to the list of corresponding documents and can be used to implement recommender systems.

In our case the seed is a document found in the iSearch corpus. For calculating the tf*idf measure we applied co-citation analysis to the seed document where tf is the citation count of the seed document together with a potential candidate at the same time. Correspondingly df is the overall citation count for the candidate in the whole iSearch corpus. Given a seed document A we first have to retrieve all documents that cite A. To this end we used the internal reference identifiers (see internalRef in table 1) which are available for about 1.7 million records. The list of candidates is then created by taking all citations that are being cited from the documents retrieved in step one.

To demonstrate our proof of concept implementation we picked an example seed document titled *Phenomology with Massive Neutrinos* which was best to fit our needs as it had a large number of citations pointing to documents which were also available in the iSearch corpus: In the first step we retrieved all 58 documents that cite the seed. Next we retrieved all citations within those documents (1076 documents) and use these as a list of potential candidates for being equality relevant to the seed's topic. The tf measure is then calculated as the number of times a candidate is found within this list. The df measure is calculated as the number of times the candidate is overall cited within the corpus. In table 2 you find the three best ranked documents based on the tf*idf measures for the co-citation analysis. Just like our seed document the first three highest ranked results were published in the field of *High Energy Physics - Phenomenology (hep-ph)*. For our seed we found 246 potential candidates that have been co-cited at least once together with the seed. Looking at the top 3 ranked documents within table 2 we can see that there is a topical similarity between the seed and the retrieved results. All 3 candidates deal with *Neutrinos* or *Neutrino Masses / Massive Neutrinos*. We interpret similarity in the title as a topical similarity between documents.

3 Results

We now apply co-citation analysis to all relevant marked documents in the qrel files hoping to answer the two research questions proposed in the introduction. The first one was on the general feasibility of citation analysis using the iSearch test collection. The second one was to investigate if co-cited documents are relevant to the underlying topic that the seed document comes up from. In a first step we had to make sure that all documents within the qrel files were available in the reference file of iSearch. Figure 1 shows the number of documents per topic in the qrel files and the corresponding number of documents found in the iSearch reference list. We were able to find a matching entry within the corpus for 78.1% (8803 documents) of the docu-

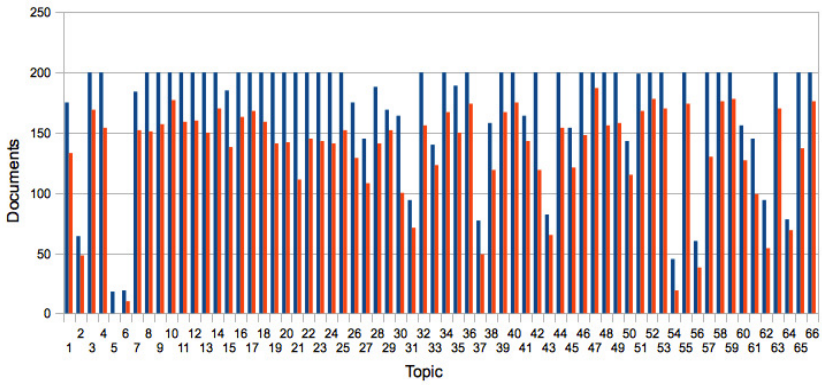

Figure 1. Availability of qrel documents in citation-corpus

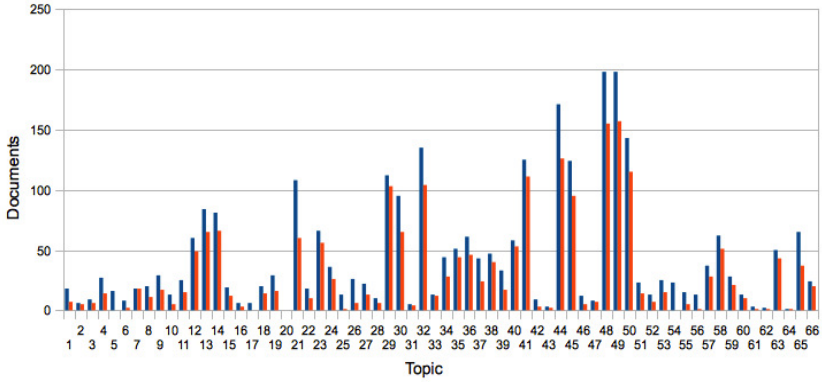

Figure 2. Number of relevant seeds in qrel and their availability in the citation-corpus

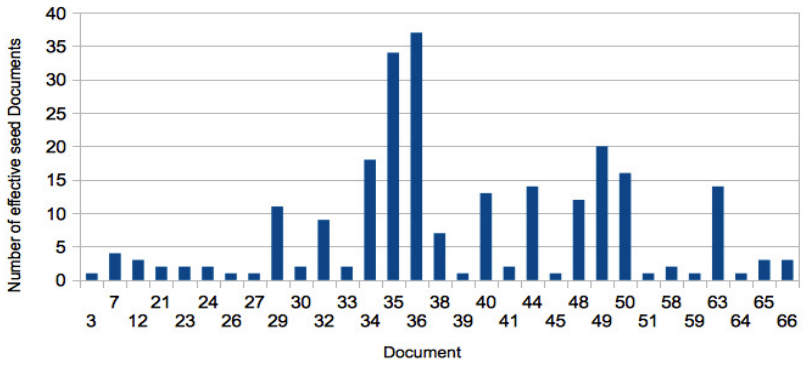

Figure 3. Number of seed documents with at least one potential candidate per topic

ments in the qrel files with an average number of 133.3 documents per topic. The only exception was topic 5 with no matching documents in the corpus. The best results were achieved for topic 47 where 93.5% of the documents in the qrel files were available in the iSearch corpus.

Subsequently, we filtered the documents in the qrel files to those judged relevant by the experts. Figure 2 shows the number of relevant documents per topic and their availability in the reference list. In total we found 2074 relevant documents in the reference list. Each topic contains an average number of 31.42 relevant documents per topic although we find exceptions for certain topics. For example in topic 48 and 49 almost every document (99%) is judged relevant. On the other hand topic 5 only contains about 5% relevant documents.

We then applied co-citation analysis using those documents as a seed that were available in our corpus and that were rated as relevant. The linkage between the seed document and the citations is again constructed by using their identifiers. The results of the co-citation analysis are displayed in figure 3. We found results for 30 topics with an average number of *effective* seed documents of nearly 8. A seed document is considered as effective if it has at least one candidate that is potentially related to the seed. Again we find variations between the topics in this case the number of effective seed documents. Topics 34, 35 and 36 achieve a larger number of effective seed documents with a success rate of nearly 76% whereas the topics 21, 23 and 58 produce weaker results with a success rate of nearly 3.6%.

As a demonstration and to further illustrate our results table 3 shows seven tf*idf ranked candidates for one seed documents taken from topic 48. Our seed document had a relevance rating of 2 (fairly relevant), was titled "*Kinetic exchange vs. room temperature ferromagnetism in diluted magnetic semiconductors*", and published in the field of *Condensed Matter*.

The first row contains the seed document. This is a wanted behavior in our analysis as the seed defines the highest possible tf*idf value. Just like the seed all candidates were published in the field of Condensed Matter. If we find a candidate within our

qrel files this is shown in column 4 together with the source topic and the rating. Column 5 shows that each candidate has a tf score of 2, which means that each candidate was co-cited with the seed two times. Variations between the candidates are found for the df score in column 6 expressing the number of times the candidate was cited in the corpus overall showing a typical tf*idf measure behavior. The more often a document is cited in the overall corpus the less relevant it becomes for a certain seed document. Both log_tf in column 8 and log_idf in column 9 were calculated with log10.

Looking at the titles in column 1 we can see that there is a similarity between the seed document and the ranked list of candidates. A title match between the seed document and the candidates shows that except for document 5 all contain the terms "*ferromagnetism*" or "*semiconductors*" which can be found in the seed document too.

Table 3. Results for seed document from topic 48

ID	Field	Title	Topic/ Rating	tf	df	log_tf	log_df	tf*idf
0201012	cond-mat	Kinetic exchange vs. room temperature ferromagnetism in diluted magnetic semiconductors	48/2	9	9	0.95	4.04	3.86
0309509	cond-mat	First-principles investigation of the assumptions underlying Model-Hamiltonian approaches to ferromagnetism of 3d impurities in III-V semiconductors	31/0	2	2	0.30	4.69	1.41
0201179	cond-mat	Why ferromagnetic semiconductors?	48/1	2	3	0.30	4.52	1.36
0208596	cond-mat	Disorder effects in diluted ferromagnetic semiconductors	-/-	2	4	0.30	4.39	1.32
0208010	cond-mat	Magneto-optical study of ZnO based diluted magnetic semiconductors	48/2	2	5	0.30	4.30	1.29
0302178	cond-mat	Self-interaction effects in (Ga,Mn)As and (Ga,Mn)N	31/0	2	9	0.30	4.04	1.21
0111045	cond-mat	Mean-field approach to ferromagnetism in (III,Mn)V diluted magnetic semiconductors at low carrier densities	50/1	2	10	0.3	4.0	1.20
0111314	cond-mat	Ferromagnetism in (III,Mn)V Semiconductors	-/-	2	36	0.3	3.44	1.03

4 Discussion

In this paper we showed some preliminary results of our experiments on the iSearch test collection and the included internal references. Our long-term goal is to make use of these references in the context of an information retrieval system but for this paper we choose to focus on two general questions. These questions focused on the feasibil-

ity of the iSearch collection with regards to citation analysis and the possible intersection of relevant documents in the qrel files and the co-citation recommendations. We intended to detect a significant overlap to make some statements about the topical connection between the seed document and the recommended documents.

Unfortunately our experiments showed that by only using the internal reference identifiers (see internalRef in table 1) to apply the co-citation analysis did not retrieve a high enough number of documents. Although certain topics like 34, 35 and 36 retrieved a very high number of potential relevant documents the majority of topics only had a rather low number. Additionally the qrel files with a maximum number of 200 rated documents per topics are too sparse for an analysis like this. Therefore a next step would be to expand the co-citation by using a combination of title, authors, journal and publication year to identify citations.

We are totally aware of the fact that the presented pre-study didn't include any retrieval experiments but only made use of the available relevance information to measure an intersection between potentially relevant documents recommended through an co-citation analysis-based system and known relevant documents on a per topic basis. The implementation of citation analysis in an IR system and the evaluation of the recommended documents have to be left open for future work.

To enable further research on the combination of bibliometrics/citation analysis and information retrieval we decided to release our source code. The code is hosted on GitHub[1] and contains the following functionalities: (1) the initial SQL statements to include the reference lists of iSearch into a MySQL database, (2) a bundle of Groovy scripts to perform bibliometric analysis on this reference database.

References

1. Beel, J., Gipp, B.: Google Scholar's Ranking Algorithm: An Introductory Overview. Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09). pp. 230–241 International Society for Scientometrics and Informetrics (2009).
2. Buckley, C.: Why current IR engines fail. Inf. Retr. 12, 6, 652–665 (2009).
3. Harpale, A. et al.: CiteData: A New Multi-faceted Dataset for Evaluating Personalized Search Performance. Proceedings of the 19th ACM International Conference on Information and Knowledge Management. pp. 549–558 ACM, New York, NY, USA (2010).
4. Lykke, M. et al.: Developing a Test Collection for the Evaluation of Integrated Search. In: Gurrin, C. et al. (eds.) Advances in Information Retrieval. pp. 627–630 Springer, Berlin, Heidelberg (2010).
5. Norozi, M.A. et al.: Contextualization from the Bibliographic Structure. CEUR Workshop Proceedings. pp. 9–13 (2012).
6. Pao, M.L.: Term and citation retrieval: A field study. Inf. Process. Manag. 29, 1, 95–112 (1993).
7. White, H.: Some new tests of relevance theory in information science. Scientometrics. 83, 3, 653–667 (2010).

[1] https://github.com/ZCarevic/iSearchCitationAnalysis