Bibliometrics of "Information Retrieval" – A Tale of Three Databases

Judit Bar-Ilan[1]

[1] Bar-Ilan University, Ramat Gan, 5290002, Israel
Judit.Bar-Ilan@biu.ac.il

Abstract. Coverage is an important criterion when evaluating information systems. This exploratory study investigates this issue by submitting the same query to different databases relevant to the query topic. Information was retrieved f from three databases: ACM Digital Library, WOS (with the Proceedings Citation Index) and Scopus. The search phrase was "information retrieval", publication years were between 2013 and 2016. The location of the search phrase was limited to title and abstract (and also keywords for WOS) and the subject area was limited to computer science or information science in WOS, computer science or social science in Scopus. From the ACM Digital Library data were retrieved from the more comprehensive ACM Guide to Computer Literature that includes also non-ACM data and also covers the major journals in information science. Altogether 9050 items were retrieved, out of which 5591 (62%) items were retrieved by a single database only, and only 1059 (12%) items were located in all three databases. There are great variations in the citation counts as well.

Keywords: bibliometrics, information retrieval, citation databases.

1 Introduction

Cyril Cleverdon [2] stated that users judge information retrieval systems by six criteria: 1) coverage 2) recall 3) precision 4) response time 5) presentation and 6) effort. Most evaluations consider precision and recall, but in this paper, we concentrate on the first criterion: coverage by testing three large databases on a test query.

It is well-known that there are differences between the coverage of databases. As a result of which both publication and citation counts can differ greatly (see for example [1]), which influences other indicators, like the h-index, most cited sources and most cited publications as well. In the following we demonstrate this for the term "information retrieval", by comparing three databases that provide citation counts, two of them comprehensive (the Web of Science (WOS), Scopus and one subject specific, the ACM Digital Library (ACM)). Information retrieval is a topic relevant both for computer science and for information science. A priori it was expected that the best coverage in terms of publication counts will be provided by the ACM Digital Library's Guide to Computing Literature, as it claims to be "the most comprehensive bibliographic database focused exclusively on the field of computing"

(http://dl.acm.org/advsearch.cfm), and also because the coverage of papers appearing in proceedings is known to be spotty in Scopus and WOS [1]. The ACM guide to Computer Literature also covers well the major information science sources related to information retrieval. In terms of citation counts there were no special expectations, because each database draws the citations only from the items covered by it, and it was not clear how much interest there is in information retrieval outside the field.

We only found a few articles that assessed information retrieval research, all having a different flavor from what is presented here. For example, Ding, Chowdhury and Foo [3], conducted a journal co-citation study of information retrieval. Another study [4] ranked highly cited researchers in IR by using a weighted PageRank-like algorithm. A more recent study [6] explored the intellectual structure of information retrieval.

2 Methods

2.1 Data Collection

For this study data were collected in May 2017, from three databases, ACM, Scopus and WOS. The search query was identical in all three cases: "information retrieval" as a phrase and so were the publication years, 2013-2016. However, there were slight differences in the search strategies as described below.

The ACM Digital Library allows to search in two sources: the ACM Full Text Collection and the more comprehensive (in terms of meta-data) ACM Guide to Computing Literature. The second option was chosen and we searched for "information retrieval" in the abstract or in the title. After data cleansing (removal of duplicates, items with missing titles or authors), 3849 items remained out of the initially retrieved 4161 items. ACM Digital Library allows to download meta-data, but these do not include citation counts, which had to be added manually.

In Scopus, the searches were also in title and abstract, however in addition to limiting the publication years to 2013-2016, we had to limit the retrieved items to those that were in the area of computer science or social science (to include information science as well). Out of the 5635 items retrieved, 5458 remained after data cleaning.

WOS does not allow to limit the search to abstract only, so we chose topic, which includes title, abstract and keywords. We had to exclude keywords from Scopus because inclusion of keywords added mainly noise (12,931 documents for a keyword search limited to publication years and subject area as above). An examination of a sample of the documents showed that the addition of keywords introduced a lot of noise, while in ACM the keyword search had a huge overlap with the title and abstract search). The search in WOS included the Science Citation Index, the Social Science Citation Index, the Arts & Humanities Citation Index, the Proceedings Citation Indexes and the Emerging Journal Citation Index, the subject areas were limited to computer science and information science and 4265 documents were retrieved.

Next a list of unique documents was created from the items retrieved from the different data sources. This part was rather time consuming, because not all items had DOIs, and occasionally the DOIs were incorrect. Pairwise comparisons were conduct-

ed to discover overlap, and to collect the citation counts of the given item from the three databases. Then for items not matched by DOI, title and publication year were compared. These matches were manually checked, as in several cases the items with identical titles and publication years were published in two different venues. It was impossible to automatically match items using the publication source as well, because there are no uniform naming conventions for proceeding titles (e.g. to publication source for papers in SIGIR 2015, appear as:

- "Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval" in ACM
- "SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval" in Scopus and WOS

and CIKM 2015 appears as

- "Proceedings of the 25th ACM International on Conference on Information and Knowledge Management" in ACM
- "CIKM'16: PROCEEDINGS OF THE 2016 ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT" in WOS

WOS retrieved items from this conference series only in 2016, while Scopus indexed only the 2014 proceedings, and ACM retrieved items from all four years, however the source title for 2013 was slightly different, using & instead of and.

Interestingly for conducting the manual check of items that were paired only by title and publication year the start and end page of the items were most useful. Altogether 9050 unique items were identified.

It should be noted that it was not feasible to use Google Scholar or Microsoft Academic Search. In Google Scholar, one can search in the title, but not in the abstract, and appearance of the term "information retrieval" in the full text cannot serve as evidence that the paper is about information retrieval. In any case, even when conducting a title search Google Scholar reports as of May 2017, about 4,240 results published between 2013 and 2017, and for a general search about 45,400 results. Since Google Scholar does not allow to retrieve more than 1000 results, it was not feasible to include Google Scholar. Microsoft Academic Search reports more than 50,000 results for the time period, and 28,700 results for items published in 2013 alone.

2.2 Data Analysis

Longitudinal publication trends for the whole set of publications and also for the individual databases was charted both in terms of number of publications and in terms of number of citations. The h—index of the topic in each database was computed. Most cited publications were identified.

3 Results

3.1 Longitudinal Trends

Table 1 and Fig. 1 show the longitudinal trends in terms of the number of publications. Interesting to note that while the number of unique publications per year is nearly constant, the numbers are decreasing for ACM and Scopus, while increasing for WOS.

Table 1. Number of publication per year and per database

Year	ALL	ACM	Scopus	WOS
2013	2,330	1,139	1,389	922
2014	2,279	937	1,420	996
2015	2,219	941	1,295	1,235
2016	2,222	831	1,354	1,111
Total	**9,050**	**3,848**	**5,458**	**4,264**

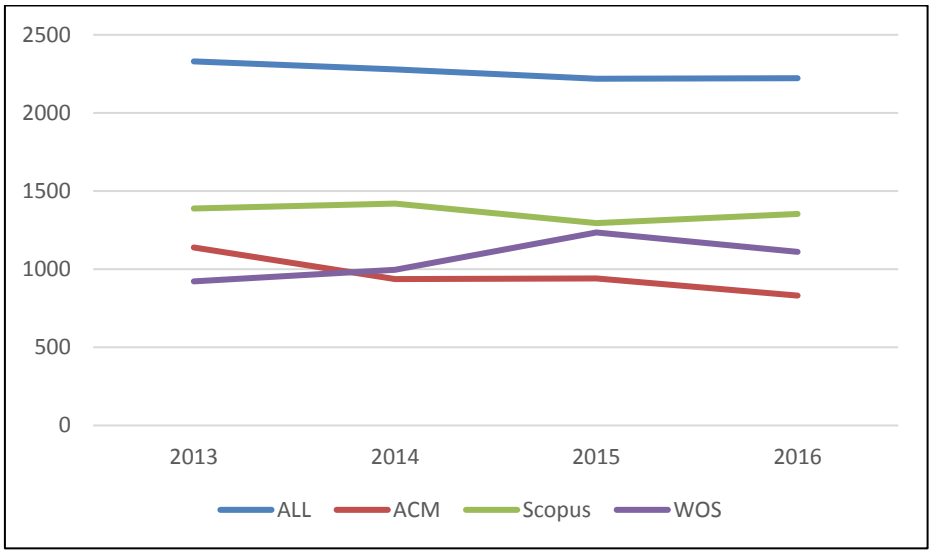

Fig. 1. Number of publication per year and per database

Table 2 shows the number of citations publications received from the time of publication until May 2017 per database. Scopus is highest for all years, WOS is second for documents published in 2013 and third for the rest of the years in terms of average number of citations received per paper. Citations accumulate over the years; thus, it is

not surprising that both total citation and citations per paper decrease as the time between publications and citations decrease.

Table 2. Citations publications received from the time of publication until May 2017 per database, total number of citations and average number of citations

Year	**ACM**		**Scopus**		**WOS**	
Citations	**Total**	**Average per paper**	**Total**	**Average per paper**	**Total**	**Average per paper**
2013	3,524	3.09	5,574	4.01	3,016	3.27
2014	2,141	2.28	3,746	2.64	2,028	2.04
2015	1,049	1.11	2,144	1.66	1,208	0.98
2016	318	0.38	623	0.46	422	0.38
Total	**7,032**	**1.83**	**12,087**	**2.21**	**6,674**	**1.57**

3.2 Overlap

The most interesting finding of this explorative study is the small overlap between the results retrieved by the databases as can be seen in Fig.2. We found only 1,059 documents (12% out of the total number of retrieved documents – 9050) that were retrieved by all three databases. On the other hand, 5,591 documents (62%) were found in a single database only. The largest overlap was between Scopus and WOS, 58% of the documents found by WOS were retrieved also by Scopus, and the smallest overlap was found between WOS and ACM, only 28% of the publication in WOS were found also by ACM.

3.3 Most cited publications

The h-index of the retrieved publications was 24 for ACM, 25 for WOS and 35 for Scopus. Although Hirsch [5] defined the h-index for individuals, it can be easily extended to any data set, where a data set has h-index *h*, if there are *h* publications that received at least *h* citations each, and *h* is maximal.

Last, the set of most cited documents retrieved by each of the databases is displayed in order to highlight the differences in terms of citation between them. The top three documents ranked by citation counts are displayed in Table 3 for ACM, Scopus and WOS respectively. Table 3 shows, that the intersection between the three sets is empty! This finding supports the subtitle: “A tale of three databases”.

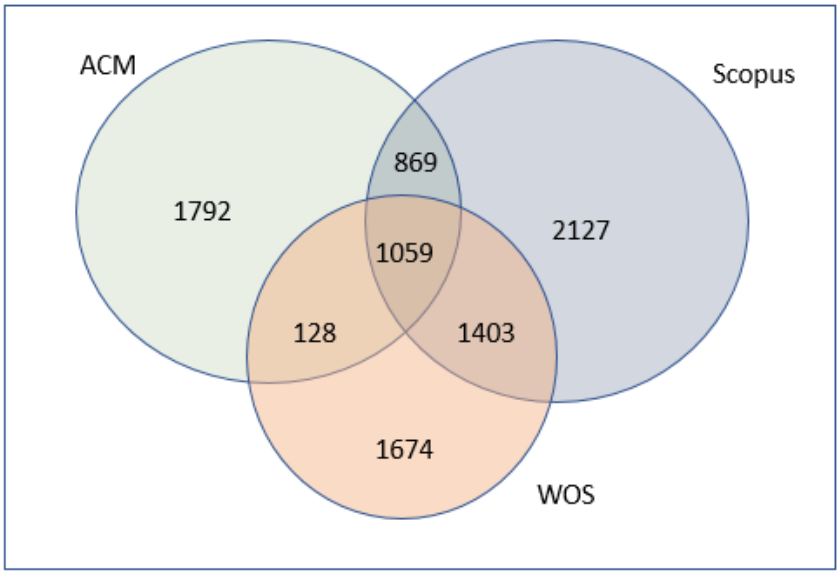

Fig. 2. Overlap between the databases

Table 3: Top-cited documents by database

rank	author	title	source	year	cits_acm	cits_sc	cits_wos
Most cited ACM							
1	Yuan et al	Time-aware Point-of-interest Recommendation	SIGIR	2013	68		
2	Xiao et al.	Expanding the Input Expressiv-ity of Smartwatches …	SIGCHI	2014	52	55	
3	Panichella et al.	How to Effectively Use Topic Models for Software Engineer-ing Tasks?	ICSE	2013	36	73	42
Most cited Scopus							
1	Deng & Yu	Deep learning: Methods and applications	Found.Trends in Signal Proc.	2013	22	145	
2	Hussein et al.	Human action recognition using a temporal hierarchy …	IJCAI	2013	26	89	
3	Brehmer& Munzner	A multi-level typology of abstract visualization tasks	IEEE Tr. Visuali-zation	2013	29	77	53
Most cited WOS							
1	Fu et al.	Enabling Personalized Search over Encrypted…	IEEE TR. PAR. & DIST .SYS.	2016			112
2	Dit et al.	Feature location in source code	J. SOFTWARE-EVOLUTION	2013			111
3	Wei et al.	Operators and Comparisons of Hesitant Fuzzy Linguistic Term Sets	IEEE TR. FUZZY SYSTEMS	2014			65

Fig. 3: The most frequently occurring title words

4 Conclusion

This contribution belongs to the BIR part of BIRNDL, as it explores one of the evaluation criteria of IR systems, coverage, applying bibliometric techniques. It emphasizes the need for searching in multiple databases in order to increase recall. The results highlight the considerable differences between the databases both in terms of the number of results found for the given query, and it terms of the citations these publications receive.

The study is exploratory in its nature and has its limitations. It should be extended to try to understand the meaning of these differences, i.e. why does each database tell us a different story? A single query is not enough for far reaching conclusions, but enough to raise interest to further explore the issue.

References

1. Bar-Ilan, J.: Which h-index? - A comparison of WoS, Scopus and Google Scholar. Scientometrics 74(2), 257-271 (2007).
2. Cleverdon, C. W.: The critical appraisal of information retrieval systems. Retrieved from https://dspace.lib.cranfield.ac.uk/bitstream/1826/1366/1/1968c.pdf
3. Ding, Y., Chowdhury, G., & Foo, S.: Journal as markers of intellectual space: Journal cocitation analysis of information retrieval area, 1987–1997. Scientometrics 47(1), 55-73 (2000).

4. Ding, T., Yan, E., Frazho, A., & Caverlee, J. “PageRank for ranking authors in co-citation networks.” Journal of the American Society for Information Science and Technology, 60(11), 2229-2243 (2009).
5. Hirsch, J. E.: An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America 102(46), 16569-16572 (2005).
6. Rorissa, A., & Yuan, X.: Visualizing and mapping the intellectual structure of information retrieval. Information processing & management, 48(1), 120-135 (2012).