Customers Who Cited this Item also Cited... - Comparing Data from Amazon.com with the new Book Citation Index

Miloš Jovanović[*] and Christian Bauckhage[**]

[*] *milos.jovanovic@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis INT, Appelsgarten 2,
53879 Euskirchen (Germany)

[**] *christian.bauckhage@iais.fraunhofer.com*
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Schloss Birlinghoven, 53757 Sankt Augustin (Germany)

Background

Thomson Reuters' bibliographic databases include citation indexes like the Science Citation Index which is part of the Web of Science (WoS). The WoS is frequently used as a database for bibliometric analyses. But some authors have pointed to the fact that the use of the WoS for bibliometric analyses of the humanities or social science can be problematic, since in those fields publication in books and monographs is at least as common as publication in scientific journals (which make up most of the WoS) (Hornbostel & Ins, 2009, p. 19). With the new Book Citation Index the number of books and monographs in the WoS has been enlarged significantly (if access to this index is available). Two book citation indexes exist, one for science (BKCI-S) and one for social sciences and humanities (BKCI-SSH) but here we will refer to both as the "BCI." The company's website conveniently points to the slogan "Completing the Research Picture" when referring to the new indexes. Up to now, no bibliometric analysis of these new indexes has been published. Since the WoS is a widely used database for bibliometric studies, it is likely that the BCI will be utilized in the same way. With the present paper, we wish to introduce our first findings of the publication and citation structure of the BCI and describe our future research on this data. In addition to analysing the BCI itself we also compare this data with another database that includes millions of books: the online retailer Amazon.com. Since Amazon includes a "Best Sellers Rank" (BSR) for most of its books and a recommendation feature based on what other customers bought ("Customers Who Bought This Item Also Bought") we want to test whether any correlations or links between the data in the BCI and Amazon can be found. More specifically, we use the BCI's citation data for these comparisons. One of the questions we want to answer is whether or not books that have often been bought have also been cited often. A similar analysis, without citation data, has been done by Yoneyama and Krishnamoothy (2009). They concluded that a book with a high sales rank is bought together with other books that are well sold. This finding is interesting since another study describes that recommendation algorithms try to compensate for best-selling items "making less well-known items much more relevant" (Linden, Smith & York, 2003). We wish to know whether a similar correlation exists between the sales rank of a book and its number of citations. We also wish to shed some light on the relationship of books in the "Customers Who Bought…" category with books that cite other books.

Methods

We used a total of 26,171 entries in the BCI. This represents the database in November 2011. In a first step, we divided the data sets into smaller subsets using the BCI's subject categories (SC). Books in the BCI have at least one SC, but can also have multiple ones. We determined the number of existing SC in the BCI (150 were found) and then counted the number of books in each of the categories. Thus, if a book has been classified with multiple SCs it counts once for each of these. The distribution of citations in scientific journals is normally heavily skewed, meaning that relatively few papers have been cited very often while a lot of papers have only been cited a few times or not at all. We analysed whether or not the same is true for books in different SCs in the BCI. In order to be able to compare the data set from the BCI with data available from Amazon.com, we searched Amazon.com for the ISBNs of the books in our database. For each book found this way, we retrieved its corresponding website at Amazon.com. If the books were found in both data sets we paired the corresponding books. For paired books we consider the BSR and the "Customers Who Bought…"-recommendation. In a next step, we will analyse whether the ranking of books by citations correlates with the BSR and whether or not the references or citations of a book also reflect the recommendations on Amazon.com.

Preliminary results

Of the 26,171 Books in the BCI only 10,383 had at least one citation. The remaining 15,788 had not been cited up to November 2011. Of those, 4,735 were published before 2009. In table 1 we show the number of books in the top five SCs. One has to keep in mind that for example a book with three SCs counts once for each of them.

Table 1. Nr. of books in the BCI for the top five SCs.

Subject Category	Number of books
Government & Law	2931
Business & Economics	2836
Computer Science	1524
Mathematics	1397
Education & Edu. Research	1353

The SC with most books is "Government & Law" and the one with the fewest is "Tropical Medicine" with 4 books. Taking a look at the books within "Government & Law," we found that the distribution of citations is, as expected, skewed (see figure 1).

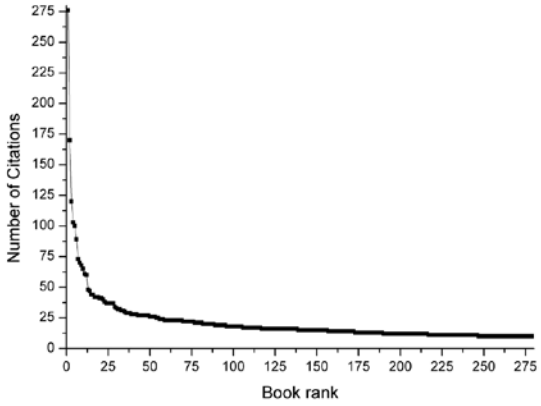

Figure 1. Distribution of citations for books from the SC "Government & Law."

At the end of this distribution are 363 books which have been cited only once and 1,361 books that have no citation. In our crawl of Amazon.com, we found 6,576 books (ca. 25% of all books in the BCI) from the BCI data set for which there is information available as to their BSR. The highest ranking book can be found at rank 52,255, the lowest on rank 11,252,068; the mean and median sales ranks were estimated to be 3,834,831.5 and 3,488,718, respectively. Figure 2 shows a histogram of the distribution of sales ranks. To produce this, we binned the data of the BSR into 75 bins. We note that the distribution is skewed to the right but of a shape that appears to differ from normal, lognormal, or power law distributions which are common for abundance data. In next steps of work, we will apply maximum likelihood methods to fit different statistical distributions to our data so as to study its peculiarities. Already at this point, we note that within the set of 6,576 ranked books, only six items were found to share a sales rank; in particular, we found three sales ranks occupied by two books each. Interestingly, these were high sales ranks and, somewhat counter intuitive, no ties were found among rarely sold items. The peak towards the right hand side of the histogram does not correspond to a collection of books of a shared sales rank.

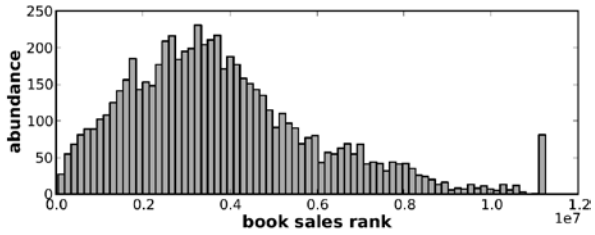

Figure 2. Histogram of the sales rank of books which can be found at Amazon.com and the BCI.

Conclusion

Since this paper represents research-in-progress, further results and conclusions will be shown in the oral presentation of this paper. However, at this early stage of our research we can already draw some preliminary conclusions: There are a number of parallels between the new book citation indexes and the other citation indexes of the Web of Science. For one, it seems as if the citation distribution is very similar. On the other hand, the distribution of those books which can be found in both the BCI and on Amazon.com in the BSR is not a typical one. It could be argued, that buying a book alone does not necessarily mean that the book will also be read and used. Thus, any correlation between the BSR and the citations can only be indicative at best. In a next step, the preliminary results presented in this paper will be further analysed and correlations checked.

References

Hornbostel, S., Klingsporn, B. & Ins, M.v. (2009). Messung von Forschungsleistungen. In G. Schütte & C. Schuh (Eds.), *Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen.* (pp. 14-35). http://www.humboldt-foundation.de/

Linden, G., Smith, B. & York, J. (2003). Amazon.com Recommendations–Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7, p. 76-80.

Yoneyama, T. & Krishnamoothy, M.S. (2009). Observation of Network Structure in Amazon.com. In *IEEE International Conference on Intelligence and Security Informatics* (ISI '09) (pp. 13-18). Dallas.