Evaluating Popularity Data for Relevance Ranking in Library Information Systems

Kim Plassmeier, Timo Borst
German National Library of Economics
k.plassmeier;t.borst@zbw.eu

Christiane Behnert, Dirk Lewandowski
Department of Information
Hamburg University of Applied Sciences
christiane.behnert;dirk.lewandowski@haw-hamburg.de

ABSTRACT
In this poster, we present our work in progress to develop a relevance model for library information systems, which takes non-textual factors into account. Here we focus on popularity data like citation or usage data. These data contain various biases that need to be corrected so as not to degrade the performance of the relevance model. Further, the different data might be to some extent incommensurable. We make use of the Characteristic Scores and Scales method to achieve two goals: first, remove biases from the raw data, and second, establish a common scale for the different data to support weighing the data against each other.

KEYWORDS
Relevance Ranking; Library Information Systems; Discovery Systems; Information Retrieval

INTRODUCTION
Web search engines can be seen as a useful model for other information retrieval systems when it comes to ranking algorithms and results presentation. Nowadays, library catalogs or *discovery systems* have implemented search engine technology to meet users' expectations in terms of searching and finding information (Antelman, Lynema, & Pace, 2006; Breeding, 2006; Connaway & Dickey, 2010; Lewandowski, 2010; Niu & Hemminger, 2010). In particular, relevance factors for ranking search results can be adopted by libraries as their materials contain more and more digital content. Traditional library catalogs usually rank search results according to publication date, whereas the additional integration of popularity-based factors is highly promising to produce valuable benefits.

In this poster, we present, which data can best be used for relevance ranking based on popularity and how these data are corrected for some contained biases and how to establish a common scale for the different data to support weighing the data against each other. We present a sample of three types of data:

- Citation counts for citation-related factors;
- Author metrics for citation- and author-related factors;
- Usage data for usage-based factors.

We describe a work in progress, as the data collection and processing are prerequisites within our research project[1] that aims to systematically evaluate ranking factors by using human relevance assessments. We use a test environment based on EconBiz, which is a search portal for economics hosted by the German National Library of Economics. EconBiz aggregates multiple databases, such as the library's local collection or RePEc[2], an open bibliographic database for economics, which also provides extensive usage and citation data for its collection.

RANKING FACTORS AND DATA SOURCES
The factors suitable for relevance ranking in library information systems can be categorized into six groups, modified as per Lewandowski (2009): Based on *text statistics* (e.g., tf*idf), the *freshness* (e.g., publication date) and *popularity* (e.g., number of clicks) influence the ranking of search results to a great extent. Next to *locality and availability* (e.g., open access), the *user background* (e.g., user group allocation) contains useful information with regard to personal ranking methods, and *content properties* (e.g., language) can be useful as well.

Popularity-based relevance ranking can be determined by different factors. In the following we will focus on citation counts, author metrics, and usage data, while we will also consider other popularity data in our complete relevance model. Some of the data can be retrieved from the library's internal sources, such as number of loans or number of clicks on a record. Other data, like citation data, will, in general, need to be retrieved from external sources.

ASIST 2015,November 6-10, 2015, St. Louis, MO, USA.

[1] http://librank.info

[2] http://repec.org

Data type	Data source (internal or external)	Popularity factor
Citation counts	CitEc (external)	Number of citations for item
	SCImago Journal Rank, CitEc (external)	Citation impact for journal
Author metrics	CitEc (external)	Citation impact for author
Usage data	Web analytics tool (internal), LogEc (external)	Number of record views
	Web analytics tool (internal), LogEc (external)	Number of clicks on full text
	Library's local system (internal)	Number of loans at local library

Table 1: Types and sources of data for popularity based ranking factors (sample).

One might be tempted to refer to standard bibliometric indicators e.g., citation counts or the *h*-index, as appropriate measures for the corresponding factors. In the following, however, we will discuss some of the biases contained in these measures, which might degrade the performance of the relevance model and, therefore, their effect should be reduced. First, however, we will roughly summarize the method of Characteristic Scores and Scales (CSS method), which we will rely on later.

CHARACTERISTIC SCORES AND SCALES

Glänzel & Schubert (1988) proposed the Characteristic Scores and Scales method to find characteristic partitions for citation distributions. They used the method to establish classes of papers that they interpreted as "poorly cited", "fairly cited", "remarkably cited", or "outstandingly cited".

The classes are constructed in the following way: The first class boundary is set to the mean of the distribution, $\beta_1 = \mu$. Then the distribution is truncated at the first boundary and the second boundary is found at the mean of the truncated distribution, $\beta_2 = \text{mean}(\{x_i | x_i \geq \beta_1\})$. Finally, the k-th class boundary is given by

$$\beta_k = \text{mean}(\{x_i | x_i \geq \beta_{k-1}\}).$$

The iteration might be stopped, for instance, if a maximal number of classes has been reached or if the proportion of elements in class k drops below a predetermined threshold.

From these classes, we construct a continuous transformation of the original values to the normalized values. Therefore we map the class boundaries to the interval [0,1]; e.g., $\beta'_k = k/k_{\max}$, and linearly interpolate between the class boundaries.

In our setting, the CSS method serves two purposes: first, it allows the removal of biases from the raw data by normalizing the distributions with respect to some subsets (e.g., publications of the same age). Second, when applying the CSS method also to other popularity data, we expect the common scale established by the CSS method will help to weigh these factors against each other in the overall relevance model. Conceptually, this common scale corresponds to the utility scale used in multi-attribute utility theory (MAUT). Note that the construction of the utility scale is quasi-parameter-free (only the number of classes must be determined). Therefore, unlike other common methods, no training data are required (see e.g., Gerani, Zhai, & Crestani, 2012).

EVALUATION OF DATA FOR POPULARITY FACTORS

Citation counts

Citation counts are affected by various effects (Bornmann & Daniel, 2008). One of the most compelling effects is due to the time dependence of the citation counts. A very common pattern in citation count distributions over time is an increase in citation counts for younger articles and a decline in citation counts for older articles. This effect is known as citation obsolescence and can be explained by the growth of the literature and a decline in cited literature as it ages (Glänzel & Schoepflin, 1995). Therefore, older articles tend to have a lower citation count than middle-aged articles. Nevertheless, to conclude that the older article had less impact or was somehow less important than the younger article would be premature.

The comparability of citation counts is a common problem in scientometrics and various normalization schemes have been developed to address this issue. Recently, percentile or quantile based methods, like PR(6), have been found to be more effective than mean-based methods (e.g., Bornmann, Leydesdorff, & Mutz, 2013). We have chosen the method of Characteristic Scores and Scales to normalize the citation distributions, which is very similar to PR(6). However, the CSS method does not depend on a predefined set of percentiles, like PR(6), but it can adapt to the underlying distribution.

We have collected citation data from CitEc[3]. The CitEc data contain citation information for about 650,000 documents (although this covers only a subset of the records contained in EconBiz). In Figure 1a, we show the counter-cumulative distribution of the raw citation counts for selected years (double logarithmic). The difference in the scale of the citation counts is evident from the plot. Further, the distributions cannot be made congruent by simple linear transformations. In Figure 1b, we show the distributions of the normalized citation counts, which are now very similar, although the raw citation distributions for some years exhibit a higher variance in the tails, for which the CSS method cannot completely compensate (in part, this might be due to the variance of coverage of the

[3] http://citec.repec.org

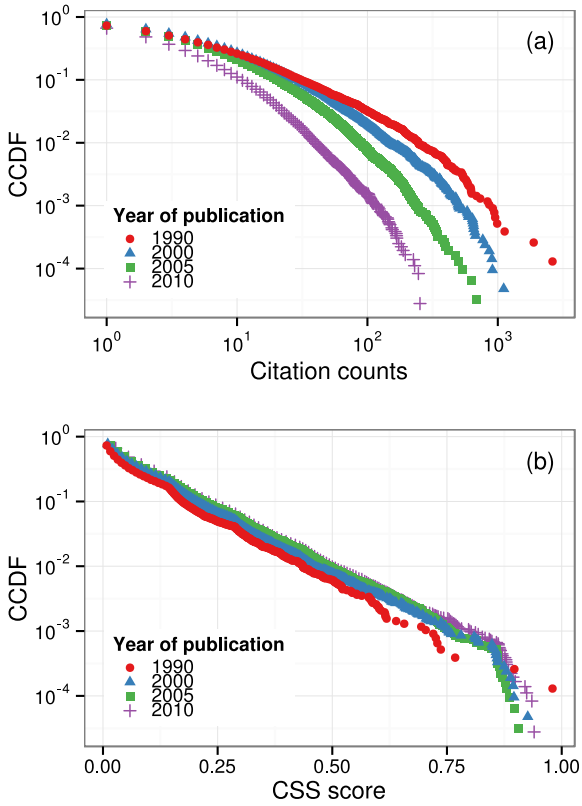

Figure 1: Counter cumulative distribution function (CCDF) of the CSS scores for the citation counts.

citation data per year). This leads to deviations in the normalized citation count distributions for these years – especially in the tails (for instance, for the year 1990 in Figure 1).

Finally, the CSS scores represent the utility values, which enable the relevance model to weigh the citation counts against other factors. Due to the non-linear transformation of the CSS method, the effect of extreme values is reduced.

Author metrics

One of the most prominent indicators to assess the performance of an author is the *h*-index (Hirsch, 2005). The *h*-index of an author is *h* if the author has published *h* publications that each received at least *h* citations. The *h*-index increases monotonically in time; therefore, authors of higher scientific age have a higher expected *h*-index. Theoretical models for the publication and citation process indicate that the *h*-index should be linearly correlated with the scientific age of an author (Burrell, 2007; Guns & Rousseau, 2009). Empirically, slightly non-linear behavior has also been observed (Liang, 2006).

The seniority of an author should not be an (implicit) factor for his or her popularity. Due to the approximately linear relationship, we can instead use the *m* quotient, which is defined as the *h*-index divided by scientific age (Hirsch, 2005). In Figure 2, the counter-cumulative distribution of

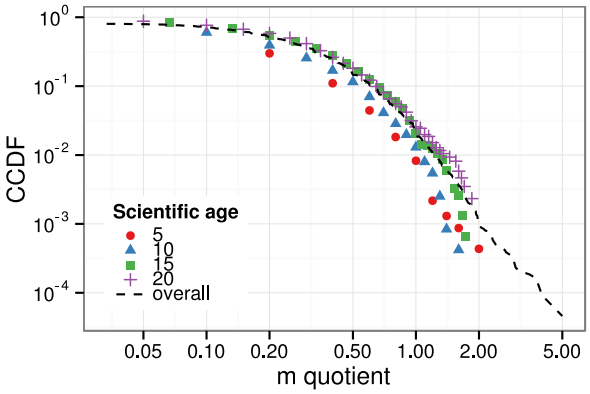

Figure 2: Counter cumulative distribution of the *m* quotient.

the *m* quotient for selected scientific ages is shown. The distributions are reasonably similar, while larger deviations occur for lower seniority, which is in part due to the discreteness of the *h*-index. Further, the mean of the *m* quotient according to age shows a slight age dependence: i.e., on average, authors of lower seniority will have a slight disadvantage, although one could argue that the *m* quotient for scientists with higher seniority is a descriptive measure of the scientist's work; while, for lower seniority, it is more prescriptive and therefore, less informative.

Since the individual distributions are already sufficiently similar, we do not apply further normalization, but use the overall *m* quotient distribution instead. We apply the CSS method, however, here we only use it to calculate the final utility scores. Again, the effect of extreme values is thereby reduced. Finally, to get a score for a specific publication, the author scores for this publication must be aggregated (e.g., by taking only the maximal author score).

Usage data

We have collected usage data from EconBiz and LogEc[4]; i.e., number of views and number of full-text accesses or downloads. The distributions of the two samples are quite different due to a significantly different usage per document ratio. This leads to a more pronounced accumulation of lower number of downloads per document in the EconBiz sample. After applying the CSS method to normalize the number of downloads from the two samples, this artifact is still present in the resulting distribution but the shapes of the two normalized distributions are quite similar (Figure 3). The normalized scores then allow us to merge both samples into an overall factor for the number of downloads. Note that here we only focused on combining usage data from different sources. However, we also plan to normalize for usage obsolescence effects (Kurtz & Bollen, 2010).

[4] http://logec.repec.org

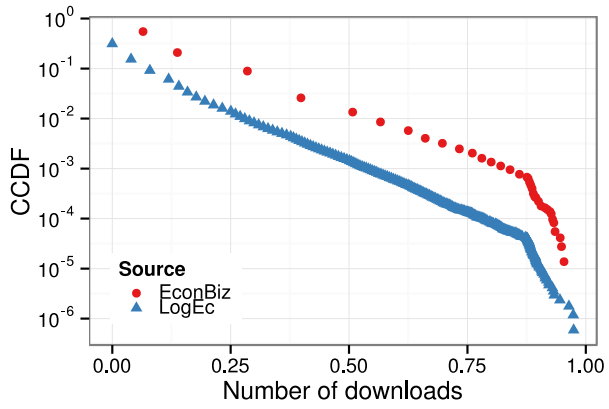

Figure 3: Counter cumulative distribution of the CSS scores for the usage data from LogEc and EconBiz.

CONCLUSION

We evaluated various popularity data for the integration into a relevance model. The raw data needed to be normalized due to various biases contained in the data (e.g., age bias for citation counts or source bias for usage data). Applying the CSS method worked well to normalize the citation distributions. For the usage data, the CSS method was not able to fully compensate for the differences in the raw distributions, although the shapes of the normalized distributions are quite similar. We did apply the same approach to other popularity data, such as number of copies in libraries, number of loans, number of editions/variants, etc., however, only to obtain comparable utility scores. Since the construction of the utility scores is quasi- parameter-free, this method might be especially interesting for LIS systems when no training data are available. However, the effectiveness of CSS scores as utilities in the overall relevance model must still be evaluated in user studies.

Interesting future directions might be to also normalize the citation data for subfield effects or to also consider other variants of the *h*-index, like the *m* index, that are more related to the impact of an author (Bornmann, Mutz, & Daniel, 2008).

ACKNOWLEDGMENTS

The individual results presented in this poster are part of the research project *LibRank* funded by the German Research Foundation.

REFERENCES

Antelman, K., Lynema, E., & Pace, A. K. (2006). Toward a twenty-first century library catalogue. *Information Technology & Libraries*, *25*(3), 128–139.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.

Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics*, *7*(1), 158–165.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, *59*(5), 830–837.

Breeding, M. (2006). Technology for the Next Generation. *Computers in Libraries*, *26*(10), 28–30.

Burrell, Q. L. (2007). Hirsch's h-index: A stochastic model. *Journal of Informetrics*, *1*(1), 16–25.

Connaway, L. S., & Dickey, T. J. (2010). The Digital Information Seeker: report of findings from Selected OCLC, RIN and JISC User behaviour projects. *Behaviour*. OCLC Research.

Gerani, S., Zhai, C., & Crestani, F. (2012). Score transformation in linear combination for multi-criteria relevance ranking. *Advances in Information Retrieval*.

Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, *21*(1), 37–53.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Studies in International Education*, *14*(2), 123–127.

Guns, R., & Rousseau, R. (2009). Simulating growth of the h-index. *Journal of the American Society for Information Science and Technology*, *60*, 410–417.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 102, pp. 16569–16572).

Kurtz, M. J., & Bollen, J. (2010). Usage bibliometrics. *Annual Review of Information Science and Technology*, *44*(1), 1–64.

Lewandowski, D. (2009). Ranking library materials. *Library Hi Tech*, *27*(4), 584–593.

Lewandowski, D. (2010). Using search engine technology to improve library catalogs. *Advances in Librarianship*, *32*, 35–54.

Liang, L. (2006). H-index sequence and h-index matrix: Constructions and applications. *Scientometrics*, *69*(1), 153–159.

Niu, X., & Hemminger, B. M. (2010). Beyond text querying and ranking list: How people are searching through faceted catalogs in two library environments. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–9.