On Publication Usage in a Social Bookmarking System

Daniel Zoller
Data Mining and Information Retrieval Group
University of Würzburg
zoller@informatik.uni-wuerzburg.de

Stephan Doerfel
ITeG,[*] Knowledge and Data Engineering (KDE) Group
University of Kassel
doerfel@cs.uni-kassel.de

Robert Jäschke
L3S Research Center
jaeschke@L3S.de

Gerd Stumme
ITeG, Knowledge and Data Engineering Group (KDE)
University of Kassel
stumme@cs.uni-kassel.de

Andreas Hotho
Data Mining and Information Retrieval Group
University of Würzburg
hotho@informatik.uni-wuerzburg.de

ABSTRACT

Scholarly success is traditionally measured in terms of citations to publications. With the advent of publication management and digital libraries on the web, scholarly usage data has become a target of investigation and new impact metrics computed on such usage data have been proposed – so called *altmetrics*. In scholarly social bookmarking systems, scientists collect and manage publication meta data and thus reveal their interest in these publications. In this work, we investigate connections between usage metrics and citations, and find posts, exports, and page views of publications to be correlated to citations.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

altmetrics, scholarly impact, social bookmarking, collaborative tagging

1. INTRODUCTION

Scholarly impact is traditionally measured in scores computed from counting citations to publications. However, citation counts come with the drawback of being only available long after an article has been published – simply because it takes time to write and publish new articles with a corresponding reference. With the advent of the social web, more and more scholarly communication and parts of the publication process have moved to the web and have thus become observable.

The creation of impact measures from such data has been subsumed under the umbrella term *altmetrics* (alternative metrics). According to the Altmetrics Manifesto [4] the goals of this initiative are to complement traditional bibliometric measures, to introduce diversity in measuring impact, and to supplement peer-review. The manifesto also appeals: "Work should correlate between altmetrics and existing measures, predict citations from altmetrics and compare altmetrics with expert evaluation." In that spirit, we analyze the usage metrics that can be computed in the social web system BibSonomy,[1] a bookmarking tool for publication references [1]. Like other tagging systems, BibSonomy allows its users to create collections of publications and to annotate each publication with a set of tags.

In our investigation, we use six different metrics for a publication's impact:

1. $post(p)$ counts how often a publication p was bookmarked.
2. $view(p)$ denotes how often a publication p has been viewed (e.g., its details page or a page with all posts about this publication from different users).
3. $exp(p)$ denotes the number of times a publication p has been exported into citation formats.
4. $exp_{Bib}(p)$ counts how often a publication p has been exported to BibTeX, the most often requested export format on BibSonomy.
5. $req(p)$ counts all requests to a publication p, exports or otherwise, i.e., it includes the counts of $view(p)$ and $exp(p)$.
6. $tag(p)$ counts for a publication p, how often one of its tags has been used in a search query.

These metrics are computed from user-generated content of BibSonomy (i.e., the bookmarked publication references) and the traces of usage behavior that are stored in the web logs of such a system (see [2] for details). We compute these measures to investigate correlations between them and actual citations, which we gathered from the scholarly search engine Microsoft Academic Search[2] (MAS in the following).

[*] Interdisciplinary Research Center for Information System Design

WebSci '15, June 28 - July 01, 2015, Oxford, United Kingdom
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3672-7/15/06…$15.00
http://dx.doi.org/10.1145/2786451.2786927

[1] `http://www.bibsonomy.org/`

[2] `http://academic.research.microsoft.com/`

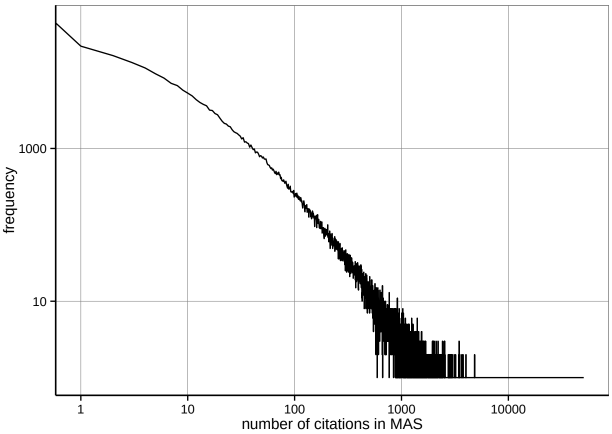

Figure 1: The frequency distribution of the number of citations in MAS to publications in BibSonomy (visualized on a log-log scale).

In our experiments we go beyond previous work in this area (i) by using more behavioral features than just post counts, and (ii) by using a large dataset of more than 250,000 publications instead of choosing only articles from selected high profile venues. Thus, our research question is the following: *Despite our large corpus spanning various disciplines and publications of different quality, can we still detect a usage bias towards highly cited publications?*

2. RESULTS

To get an impression on the dataset, the frequency distribution of the total number of citations in MAS to publications in the BibSonomy dataset is shown in Figure 1. Most publications in BibSonomy have no recorded citation; and publications with only one citation represent the second largest subset in the crawled dataset. The frequency decreases continuously with higher numbers of citations, but also starts to oscillate for citation counts larger than about 100.

We compute correlations between behavioral features and citation counts over all publications in our corpus. In Table 1, we report for each pair of metrics Pearson's correlation coefficient r, as well as Spearman's ranking correlation ρ. The latter has the advantage that it is suitable for non-linear relationships, and we will focus on ρ for the discussion.

Among the six behavioral metrics, the number of posts (*post*) exhibits the strongest correlation with the number of citations. It is lower than in previously reported studies, however this was to be expected since our analyzed corpus is much more inhomogeneous than the publication sets in other experiments. E.g., [3] found correlations between $\rho = 0.304$ and $\rho = 0.603$ between post counts in the bookmarking systems Mendeley and CiteULike and citations on the citation database *Web of Science* for about 800 *Nature* and about 800 *Science* articles. We still observe a small correlation that clearly indicates a bias in the behavior of users towards posting rather highly cited publications more often.

Regarding citations and the other behavioral features, we observe a noticeable bias for exporting publications (*exp*).

Table 1: Correlation between behavioral features in BibSonomy and the number of citations (*cit*) of a publication. The upper right triangle shows Pearson's r, the lower left triangle shows Spearman's ρ. All values are significant at the 0.01-level. Correlations are computed over all publications in the data.

	post	*view*	*exp*	exp_{Bib}	*req*	*tag*	*cit*
post	1	0.64	0.64	0.63	0.45	0.33	0.18
view	0.32	1	0.72	0.71	0.66	0.32	0.09
exp	0.32	0.43	1	0.99	0.74	0.28	0.16
exp_{Bib}	0.33	0.42	0.95	1	0.72	0.28	0.16
req	0.33	0.91	0.66	0.63	1	0.21	0.07
tag	0.28	0.27	0.24	0.24	0.27	1	0.04
cit	0.20	0.10	0.12	0.12	0.10	0.01	1

Hereby, the choice between all exports (*exp*) and BibTeX exports (exp_{Bib}) makes little difference – both features are almost perfectly correlated. This can be easily attributed to the fact that BibTeX is the most often used export format in BibSonomy. While *req* and *view* also show a weak correlation to citations, no real correlation can be observed between the *tag* metric and citation counts. A possible explanation for this lack of correlation is that one tag can occur in many posts and thus the metric is not publication-specific enough.

Finally, apart from *exp* and exp_{Bib}, and *req* and *view*, none of the behavioral metrics is strongly correlated to another one. Particularly between *post* and the other metrics we find medium correlations, indicating, that while these metrics are not completely diverse, they are valuable complements to just counting posts.

Conclusion. In this work we observed small yet noticeable correlations between citations and posting, viewing, and exporting publications. We conclude that the community of all users is indeed biased towards using publications that are relevant already.

Future Work. While the analysis presented here focuses on correlations between usage and citations in general, we will extend this work by investigating citations that occur in the future, i.e., citations that occur after the usage in the bookmarking system. Furthermore, we plan to evaluate actual predictability of citations for publications, based only on usage metrics in the social bookmarking system.

3. REFERENCES

[1] D. Benz, A. Hotho, R. Jäschke, B. Krause, F. Mitzlaff, C. Schmitz, and G. Stumme. The social bookmark and publication management system BibSonomy. *The VLDB Journal*, 19(6):849–875, Dec. 2010.

[2] S. Doerfel, D. Zoller, P. Singer, T. Niebler, A. Hotho, and M. Strohmaier. Evaluating assumptions about social tagging – A study of user behavior in BibSonomy. In *Proceedings of the 16th LWA Workshops: KDML, IR and FGWM, Aachen, Germany*. CEUR-WS.org, 2014.

[3] X. Li, M. Thelwall, and D. Giustini. Validating online reference managers for scholarly impact measurement. *Scientometrics*, 91(2):461–471, 2012.

[4] J. Priem, D. Taraborelli, P. Groth, and C. Neylon. Altmetrics: a manifesto, 2011.