Can Social Reference Management Systems Predict a Ranking of Scholarly Venues?

Hamed Alhoori and Richard Furuta

Center for the Study of Digital Libraries and
Department of Computer Science and Engineering
Texas A&M University, USA
{alhoori,furuta}@tamu.edu

Abstract. New scholarly venues (e.g., conferences and journals) are emerging as research fields expand. Ranking these new venues is imperative to assist researchers, librarians, and research institutions. However, rankings based on traditional citation-based metrics have limitations and are no longer the only or the best choice to determine the impact of scholarly venues. Here, we propose a venue-ranking approach based on scholarly references from academic social media sites, and we compare a number of citation-based rankings with social-based rankings. Our preliminary results show a statistically significant correlation between the two approaches in a number of general rankings, research areas, and subdisciplines. Furthermore, we found that social-based rankings favor open-access venues over venues that require a subscription.

Keywords: Scholarly Venues, Ranking, Digital Libraries, Bibliometrics, Altmetrics, Impact Factor, Readership, Social Reference Management, Citation Analysis, Google Scholar Metrics.

1 Introduction

Rankings play a vital role in daily life. Students use rankings to select top universities, graduate students use rankings to select the best jobs, patients use rankings to select hospitals, and travelers use rankings to plan their vacations. Rankings of scholarly venues are often used in academia and research. Despite the concerns and objections regarding venue rankings, they continue to be used to identify major scholarly hubs. Researchers agree that these hubs should be assessed based on academic quality. The top scholarly venues have an influence on research. Prestigious journals use the rankings for publicity, librarians use them for subscription decisions, researchers use them for publication decisions, and universities use them for academic hiring, promotions, and funding decisions.

The impact of scholarly venues is typically measured using citation analysis. A major measure used in ranking scholarly venues is the controversial ‘impact factor’, which has its own limitations. Moreover, research articles, especially those published in conferences, are limited in terms of length, so authors may not be able to cite all the related references. Various usage-based metrics, such as readership [1], downloads,

This is an author-created version. The final publication is available at:
http://link.springer.com/chapter/10.1007/978-3-642-40501-3_14

comments, and bookmarking statistics, have been proposed and used to measure the impact of articles and journals, and each has its benefits and limitations [2].

Researchers often use social reference management systems to store and discover scholarly articles. By storing references online, researchers can archive their research interests without limits. Therefore, the statistics for these online digital libraries are strong indicators of researchers' interests and may reflect more accurate interests than statistics about downloads or views.

In this study, we propose and investigate a social-based approach to ranking scholarly venues. We compare our method of venue-ranking with various citation-based ranking approaches and find several strong positive relationships. We also investigate the effects of open-access venues on rankings. This paper is structured as follows: We discuss the related work in Section 2. In Section 3, we describe the experiments, data collection, and methodology. In Section 4, we present and discuss our results. In Section 5, we conclude and highlight some of the future work.

2 Related Work

Although the impact factor is a well-known method for ranking scholarly venues, it suffers from citation delay [3], differs according to discipline [4], and may not be available for emergent venues. The Science Journal Ranking (SJR) indicator [5] has been proposed as an alternative to the impact factor. The SJR indicator considers the quantity and quality of citations. A number of journal-ranking approaches have used the PageRank algorithm, including the SJR indicator and Eigenfactor [6]. The h-index, expert survey [7], and publication power approach [8] have also been used to rank venues.

Zhuang et al. [9] used program committee characteristics to discover and rank conferences. Yan et al. [10] defined two approaches to rank academic venues, a seed-based approach that used author meta-data and a browsing-based approach that used both the citation and author meta-data. Martins et al. [11] used a large number of features with machine learning techniques to assess the quality of scientific conferences. Rahm et al. [12] found that conferences could have a higher impact factor than journals. Google Scholar joined the effort to rank venues when it announced Scholar Metrics, which ranks top scholarly venues in several disciplines and languages, ordered by their five-year h-index.

Bollen et al. [13] concluded that "the notion of scientific impact is a multi-dimensional construct that cannot be adequately measured by any single indicator". Alhoori and Furura [14] found that social reference management systems significantly affect the scholarly activities of researchers. Social-based approaches have been used to assist in evaluating the scientific impact in several projects such as Altmetrics[1], Article-Level Metrics[2], and Usage Factor[3]. Li et al. [15] compared Web of Science citation counts and CiteULike/Mendeley readership counts on a limited sample of

[1] http://altmetrics.org/manifesto/
[2] http://article-level-metrics.plos.org/
[3] http://www.projectcounter.org/usage_factor.html

articles published in *Nature* and *Science* and found significant correlations between the two rankings. Kraker et al. [16] found a significant relationship between Mendeley references and SCImago's impact index, which is SCImago's version of the impact factor. They also found differences among disciplines and indicators that results improve with number of references available.

3 Experiments

We crawled CiteULike and downloaded 554,023 files, in which each file contains a reference to an article and the users who have added it to their digital libraries. We used only the files that contained details of either conferences or journals, for a final sample of 407,038 files. We then extracted the details of venues and collected a total of 1,317,336 postings of researcher–article pairs and a total of 614,361 researcher–venue pairs. We defined three social-based metrics and used them in venue-ranking:

1. **Readership:** The number of researchers who have added references from a venue to the social reference management system.
2. **Article Count:** The number of unique articles from a single venue that were added to the social reference management system.
3. **Active Researchers Rating (ARR):** We defined active researchers as those who added twenty or more venues to their digital libraries. We used a weighted sum to increase the importance of newly added references. Equation (1) was used to compute the ARR for venue v.

$$ARR(v) = \sum_{i=1}^{n} \sum_{w=m}^{1} w \log(v_w + 1) \tag{1}$$

The outer summation of the ARR totals the individual ratings for n researchers. In the inner summation, v_w denotes the number of references from a specific venue, that a researcher added to his or her digital library during a particular year, out of all the m years that the researcher followed venue v. Weight w increased the importance of newly added references. The ARR favors researchers who have followed venues for several years over researchers who have added numerous references from venues for a few years. The log minimized the effect of adding many references.

We first compared Google's h5-index with our social-based rankings. Google's Scholar current h5-index includes research articles published between 2007 and 2011 and indexed in Google Scholar as of November 2012. To compare our social-based rankings with Google's h5-index, we selected the articles that were published and added to CiteUlike between 2007 and 2011. Our question was whether a correlation exists between social metrics from CiteULike and Google h5-index for the indicated time span. We repeated the same strategy with the other citation-based rankings. For example, Eigenfactor score, which uses Web of Knowledge citations, was released in 2011 and includes articles published between 2006 and 2010. Therefore, in this instance, we used a dataset of articles that were published and added to CiteULike between 2006 and 2010.

We used Spearman's rank correlation coefficient, ρ(rho), to compare our social-based rankings with different citation-based rankings, such as Google's h5-index, SCImago's h-index, the Thomson Reuters Impact Factor, the Eigenfactor score, and total citations. We began with citation-based rankings and mapped the corresponding values from the social-based rankings.

4 Results and Discussion

We first compared the general citation-based rankings of the top 100 venues with our social-based rankings. We found a strong positive relationships ($p < 0.01$), as shown in Table 1. There was no significant correlation between the social metrics and the impact factor or the impact index.

Table 1. Correlations between citation-based metrics and social metrics for the top 100 venues

Citation-based metric	Readership	ARR	Article count
SCImago h-index	0.581	0.566	0.534
Google h5-index	0.336	0.354	0.349
Eigenfactor Score	0.688	0.669	0.665
Total citations	0.675	0.625	0.632

We then compared the top twenty venues among different research areas using Google's h5-index and social-based metrics. We found significance relationships in some areas, as shown in Table 2. In Tables 2 and 3 below, we used * to represent ($p<0.05$) and ** to represent ($p<0.01$).

Table 2. Correlations between Google 5h-index and social metrics for different research areas

Research area	Readership	ARR	Article count
Health & Medical Sciences	0.647 **	0.672**	0.642**
Humanities, Literature & Arts	0.368	0.471	0.200
Life Sciences & Earth Sciences	0.788 **	0.768 **	0.735 **

We also compared Google's h5-index with the social metrics for some subdisciplines in engineering and computer science, as shown in Table 3.

Table 3. Correlations between Google 5h-index and social metrics for some engineering and computer science subdisciplines

Subdiscipline	Readership	ARR	Article count
Automation & Control Theory	0.567 *	0.382	0.466
Bioinformatics & Computational Biology	0.814 **	0.700 **	0.706 **
Educational Technology	0.575 *	0.512 *	0.374
Library & Information Science	0.761 **	0.769 **	0.754 **
Robotics	0.532 *	0.482	0.460 *

No significant relationships were found between Google's h5-index and social-based rankings in some areas, such as arts and humanities, and some subdisciplines, such as artificial intelligence. However, we found a significant relationship between SCImago's h-index and the readership ranking in arts and humanities ($p < 0.05$) and in artificial intelligence ($p < 0.01$). Surprisingly, and in most cases when compared with the citation-based rankings, the readership rankings had higher correlations than the ARR. The article count usually had weaker correlations than did readership and ARR.

As shown in Table 1, it is clear that social metrics are an effective way to measure the popularity of venues because they have a strong positive correlation with the total number of venue citations. Social metrics can also measure the quality of venues, as they are strongly positively correlated with quality ranking methods, such as Eigenfactor scores. Tables 2 and 3 show differences in correlations among various research areas; these differences could be due to varied levels of online scholarly activity. Moreover, such differences may also relate to unequal distributions of research communities across social reference management systems, or to the existence of research communities that are not active in such online systems. We experimented with two social-based metrics that resemble the impact factor, but we did not find any strong correlation. For the first metric, we divided the readership of a venue by article count, and for the second metric, we divided the ARR by article count.

Finally, we investigated whether the venue-ranking approach (citation-based or social-based) was related to the type of access to venues (subscription or open access). We compared the top 20 venues in Google's h5-index with the top 20 venues in readership and ARR rankings. We included hybrid and delayed access venues in the open-access venue category. There were more open-access venues in the readership and ARR rankings than in the citation-based rankings. We did not find a significant relationship for the readership ranking. However, using the ARR, we found 13 open-access venues but only 6 in Google's h5-index; a Chi-squared test determined there was a significant positive relationship ($X^2 = 4.9123$, $p < 0.05$) between the venue ranking approach and the type of access to venues.

5 Conclusions and Future Work

In this study, we investigated the relationship between ranking methods for scholarly venues that use traditional citation-based metrics and our proposed social-based metrics. We found statistically significant correlations between the two approaches, with disciplinary differences. Our results suggest that social reference management systems have the potential to provide an early intellectual indicator of the influence of scholarly venues, while reducing the limitations of citation-based metrics.

In the future, we will investigate whether there is a set of social-based metrics that can measure the influence of scholarly venues in all research areas, or if each research area needs to define its own metrics. We plan to explore how the data from different social reference management systems differ and whether they measure similar or different impact of research.

Acknowledgement. This work has been supported under the grant ID NPRP 4 – 029 – 1 – 007 for the project entitled “Establishing a Qatari Arabic-English Digital Library Institute”, 2011-2014, which is funded by the Qatar National Research Fund (QNRF).

6 References

1. Darmoni, S.J., Roussel, F., Benichou, J., Thirion, B., Pinhas, N.: Reading factor: a new bibliometric criterion for managing digital libraries. Journal of the Medical Library Association : JMLA. 90, 323–327 (2002).
2. Neylon, C., Wu, S.: Article-Level Metrics and the Evolution of Scientific Impact. PLoS Biology. 7, 6 (2009).
3. Brody, T., Harnad, S., Carr, L.: Earlier Web usage statistics as predictors of later citation impact: Research Articles. Journal of the American Society for Information Science and Technology. 57, 1060 (2006).
4. Rinia, E., Van Leeuwen, T., Bruins, E., Van Vuren, H., Van Raan, A.: Citation delay in interdisciplinary knowledge exchange. Scientometrics. 51, 293–309 (2001).
5. González-Pereira, B., Guerrero-Bote, V.P., Moya-Anegón, F.: A new approach to the metric of journals’ scientific prestige: The SJR indicator. Journal of Informetrics. 4, 379–391 (2010).
6. Bergstrom, C.: Eigenfactor: Measuring the value and prestige of scholarly journals. CRL News. 68, 314–316 (2007).
7. Serenko, A., Dohan, M.: Comparing the expert survey and citation impact journal ranking methods: Example from the field of Artificial Intelligence. Journal of Informetrics. 5, 629–648 (2011).
8. Holsapple, C.W.: A Publication Power Approach for Identifying Premier Information Systems Journals. Journal of the American Society for Information Science. 59, 166–185 (2008).
9. Zhuang, Z., Elmacioglu, E., Lee, D., Giles, C.L.: Measuring conference quality by mining program committee characteristics. Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries. pp. 225–234. ACM (2007).
10. Yan, S., Lee, D.: Toward alternative measures for ranking venues. Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries. pp. 235–244. ACM (2007).
11. Martins, W.S., Gonçalves, M.A., Laender, A.H.F., Pappa, G.L.: Learning to assess the quality of scientific conferences. Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. pp. 193–202. ACM (2009).
12. Rahm, E., Thor, A.: Citation analysis of database publications. ACM SIGMOD Record. pp. 48–53 (2005).
13. Bollen, J., Van de Sompel, H., Hagberg, A., Chute, R.: A principal component analysis of 39 scientific impact measures. PloS one. 4, e6022 (2009).
14. Alhoori, H., Furuta, R.: Understanding the Dynamic Scholarly Research Needs and Behavior as Applied to Social Reference Management. Proceedings of the 15th international conference on Theory and practice of digital libraries. pp. 169–178. Springer Berlin Heidelberg (2011).
15. Li, X., Thelwall, M., Giustini, D.: Validating online reference managers for scholarly impact measurement. Scientometrics. 91, 461–471 (2011).
16. Kraker, P., Körner, C., Jack, K., Granitzer, M.: Harnessing user library statistics for research evaluation and knowledge domain visualization. Proceedings of the 21st international conference companion on World Wide Web - WWW ’12 Companion. pp. 1017–1024. ACM (2012).