Semantometrics

Towards fulltext-based research evaluation

Drahomira Herrmannova
Milton Keynes, United Kingdom
d.herrmannova@gmail.com

Petr Knoth
Milton Keynes, United Kingdom
petrknoth@gmail.com

ABSTRACT

Over the recent years, there has been a growing interest in developing new research evaluation methods that could go beyond the traditional citation-based metrics. This interest is motivated on one side by the wider availability or even emergence of new information evidencing research performance, such as article downloads, views and Twitter mentions, and on the other side by the continued frustrations and problems surrounding the application of purely citation-based metrics to evaluate research performance in practice.

Semantometrics are a new class of research evaluation metrics which build on the premise that full-text is needed to assess the value of a publication. This paper reports on the analysis carried out with the aim to investigate the properties of the semantometric *contribution* measure [1], which uses semantic similarity of publications to estimate research contribution, and provides a comparative study of the contribution measure with traditional bibliometric measures based on citation counting.

Keywords

Research Evaluation, Citation Analysis, Text Mining

Acknowledgements

This work was supported by Jisc under contract no. 3790.

1. INTRODUCTION

We have introduced the idea of Semantometrics in [1] as a new class of metrics for evaluating research. As opposed to existing Bibliometrics, Webometrics, Altmetrics, etc., Semantometrics are not based on measuring the number of interactions in the scholarly communication network, but build on the premise that full-text is needed to assess the value of a publication.

In [1] we have attempted to create the first semantometric measure based on the idea of measuring the progress of scholarly discussion. Our hypothesis states that the added value of publication p can be estimated based on the semantic distance from the publications cited by p to the publications citing p. This hypothesis is based on the process of how research builds on the existing knowledge in order to create new knowledge on which others can build. A publication, which in this way creates a "bridge" between what we already know and something new, which will people develop based on this knowledge, brings a contribution to science [1].

Until recently, it was still technically challenging for us to obtain an evaluation dataset on which properties of the contribution metric could be analysed. In this respect, we are now able to report on the first large-scale analysis of this metric. The goal of our study was to understand the properties and behaviour of the semantometric contribution measure in comparison with established research evaluation metrics. We chose to use citation counts obtained from the Microsoft Academic Graph (MAG) [4], as the representative of Bibliometrics, usage data (readership) obtained from Mendeley[1], as the representative of Altmetrics and research articles aggregated by the Open Access Connecting Repositories[2] (CORE) system as a representative sample for studying the characteristics of the contribution measure.

2. DATASET

Our experiments have been conducted on a dataset obtained by merging data from CORE, MAG and Mendeley. To assemble this dataset, we mapped DOIs of papers from CORE with MAG. Using MAG we then identified DOIs of papers citing and cited by the CORE papers. Finally, we used the DOIs to retrieve metadata, readership counts and primarily the titles and the abstracts using the Mendeley API. By merging these three datasets, we obtained a final dataset containing metadata, citation counts and reader counts of about 1.6 million Open Access papers. Additionally, we obtained metadata, including titles and abstracts of over 10 million papers which cite or are cited by the 1.6 million papers from CORE and are needed to calculate the contribution metric.

3. RESULTS

The main area of interest to us was the relation between the contribution measure and citation counts. The reason for this was the prevalence of use of citation counts in research evaluation. While using metrics based purely on citation counts has been subject to much criticism, these metrics still remain best known and most widely adopted. The aim was not to find a perfect correlation with citation counts, but rather demonstrate how does the contribution measure behave in relation to the well-known metric.

We have first investigated the distributions of the three metrics, these are shown in Figures 1 [c], 1 [f] and 1 [i]. As expected, the citation distribution (Figure 1 [f]) is a long tail (power law) distribution. This is consistent with existing studies [3]. The readership distribution (Figure 1 [c]) exhibits the same properties as the citation distribution. In contrast to the first two metrics, the contribution distribution (Figure 1 [i]) resembles a normal distribution.

[1] http://dev.mendeley.com
[2] http://core.ac.uk

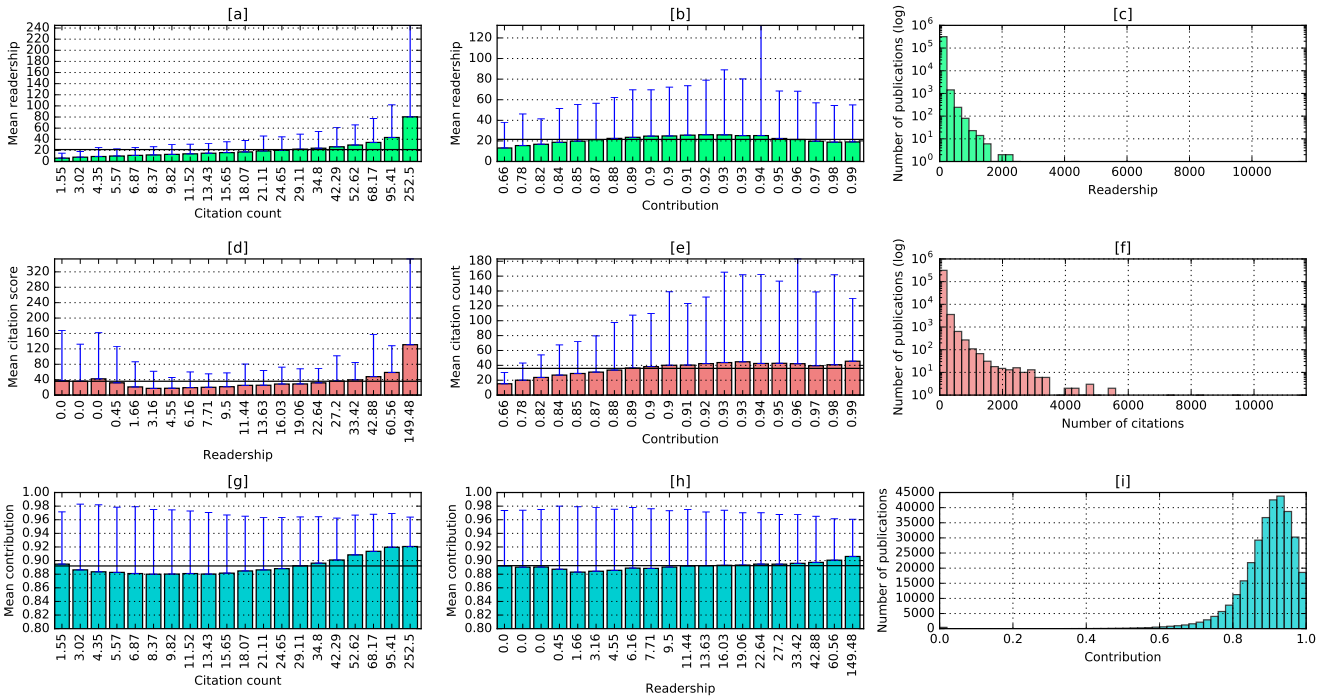


Figure 1: Results of the study. To produce figures [a], [b], [d], [e], [g] and [h], the data were split into 20 equally sized buckets by one of the studied metrics (x-axis). Mean and standard deviation of a second metric (y-axis) was then calculated for each of the buckets. The mean values are represented by the height of the bars, the vertical lines on top of the bars represent the standard deviations. The solid horizontal lines represents the mean value across all buckets.

To confirm our data are consistent with previous studies, we have investigated the relation between the citation and reader counts. We found that the two metrics are slightly correlated with Pearson $r = 0.3584$. A similarly strong correlation has been reported also by [2]. This correlation can also be seen when comparing the averaged values in Figures 1 [a] and 1 [d].

In contrast to the reader counts, we found no correlation between the citation counts and contribution (Pearson $r = 0.0871$). However, according to Figures 1 [g] and 1 [e] we can see that when comparing averaged values the behaviour of the contribution metric is not random, instead it is clearly correlated with citation counts. We can observe that publications with a citation score above a certain threshold achieve on average consistently higher contribution (Figure 1 [g]). Although the standard deviation shows it is not always the case, the results suggest that publications with more than 25 citations are more likely to have higher contribution. However, once a paper receives around 90 citations, higher citation counts do not lead on average to a higher contribution. We think this is an interesting observation that is consistent with our perception of research quality. One possible and highly simplified explanation could be that receiving around 90 citations is typically an indication of quality work. Higher citation counts then typically reflect the size of the target audience community (impact) rather than higher quality of the underlying research work. This leads us to the conclusion that the contribution metric seems to capture different aspects of research performance than citation counts.

Similarly as in the previous case, there is no correlation between the contribution measure and reader counts, which is confirmed by Pearson $r = 0.0444$. Interestingly, while we observed a correlation between the averaged contribution and citation counts, there seems to be no such relation between averaged contribution and reader counts (Figures 1 [b] and 1 [h]).

4. CONCLUSION

We have demonstrated that new measures for assessing publication impact, which take into account the manuscript of the publication, can be developed and presented a comparative study of the semantometric contribution measure with citation and reader counts. The results of our study suggest that the contribution metric captures different aspects of research performance than citation counts. More specifically, we believe that Semantometrics have the potential to capture research quality and contribution rather than research impact.

5. REFERENCES

[1] P. Knoth and D. Herrmannova. Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. *D-Lib Magazine*, 20(11/12), 2014.

[2] C. Schlögl, J. Gorraiz, C. Gumpenberger, K. Jack, and P. Kraker. Are downloads and readership data a substitute for citations? the case of a scholarly journal. *LIDA Proceedings*, 13, 2014.

[3] P. O. Seglen. The Skewness of Science. *JASIS*, 43(9):628–638, oct 1992.

[4] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of WWW 2015*, pages 243–246, Florence, Italy, 2015. ACM Press.