Using the Co-Citation Network to Indicate Article Impact

David Tarrant and Dr Les Carr

School of Electronics and Computer Science, University of Southampton
davetaz@ecs.soton.ac.uk & lac@ecs.soton.ac.uk

Scholarly outputs are growing in number and frequency, driving the requirement to research new early indication metrics. Historically, citations have been used as an independent indication of the significance of scholarly material. However, citations are very slow to accrue since they can only be made by subsequently published material. This enforces a delay of a number of years before the citation impact of a publication can be accurately judged. Existing early indication metrics, such as download metrics and web based link analysis, have obtained correlation results. Brody [1] finds a good correlation between download metrics and subsequent impact by citation, while Thelwall [2] finds very little correlation between Google's PageRank and the number of links (or citations) to a web site, suggesting neither is a good surrogate indicator for the other. While valid studies, neither take account of the quality assessment factor of peer-review citation. This work presents an investigation into new metrics which utilize the co-citation network in order to rate a publications impact.

A co-citation represents a relationship, that is established indirectly via citations, between two articles both cited by the same publication. With each publication citing a good number of others, a single direct citation to a paper will establish a good number of co-citation links to this same paper. Further, these relationships will exist with older material whose citation impact may already be established. Figure 1 demonstrates the significant difference between the citation network and co-citations networks for a single publication (p). On the left is the citation network of our publication (p), while on the right is the co-citation network of the same publication. In both cases the y-axis represents time.

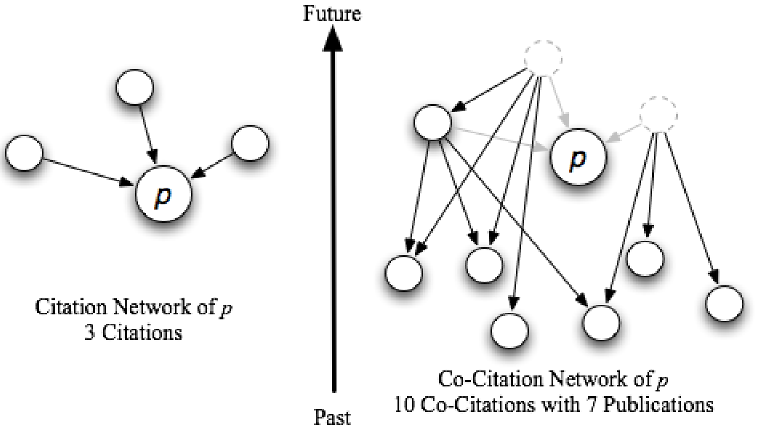

Fig. 1. Example Citation and Co-Citation Network for Publication p

Co-citation analysis is traditionally used to relate two objects together by saying that they are strongly linked in some way. As an example, if two publications are frequently co-cited then they can often be classified into the same research area [3]. Our work investigates the possibility of developing a new metric, that can provide an earlier indicator of a publications citation impact based upon co-citation relations.

Taking influences from existing metrics, including Citation Count, HITs [4] and PageRank [5], a family of seven Co-Citation based metrics were developed and evaluated over a large corpus of citation and co-citation data provided by the CiteBase[1] citation registry. Taking this family of seven metrics, PageRank, HITs and Citation Count were added to make ten metrics to be evaluated, with Citation Count representing the target "Gold Standard" metric.

CiteBase, developed as part of the Open Citation project (OpCit) [6], was designed specifically for the purpose of indexing citations over time. These characteristics made CiteBase an ideal candidate for use in this study. Data obtained from CiteBase related to over 300,000 publications, contributing a total of 3.37 million citations. This represents a figure of about 10 citations per publication and logically, 149 Co-Citations per publication, thus a total of 46 million co-citation links are established.

In order to evaluate the effectiveness of each metric a number of highly ranked publications, by Citation Count after 3 years, were selected from CiteBase and tracked from their initial creation to the end of a 3 year period. Once applied, the results from each algorithm were analysed in 3 ways. Firstly, a correlation score was generated from the comparison of rank order between each algorithm and eventual citation count. A high correlation giving a good result implied that the highest cited paper is also highly ranked by each new metric. Secondly, the overall position of the publications within CiteBase were recorded and then compared with the eventual rank position by Citation Count. High rank order correlation from the first test, but low overall rank would suggest that these publications would not be noticed by the new metric. Thirdly, the average age of the top 5% of publications by each metric was calculated in order to evaluate which metric is revealing more recently published articles. These 3 factors then combine to make an "ideal" new metric which should rank the publications as highly, sooner and in a highly Citation Count correlated manner.

The most basic of the co-citation family of metrics (CoRank-LinkCount) provides an effective early indicator of subsequent impact. This algorithm performs well against Citation Count in the first 12 months after publication as an article accrues its first citations. It also exhibits a good correlation with eventual Citation Count, while ranking the set of target publications highly in the overall dataset. Due to this positive performance in the first 12 months, CoRank-LinkCount also reveals a higher number of recent publications in its top 5% of results than Citation Count.

In order to evaluate the results of the remaining metrics, including PageRank which did not show anywhere near as positive an outcome as CoRank-LinkCount, Principal Component Analysis (PCA) was applied to the overall correlation matrix between all 10 algorithms in the study (7 Co-Citation based algorithms, Citation Count, PageRank and HITs). Using the correlation matrix, PCA outputs a series of principal vectors which can be used to re-plot the results in order to graphically show the differences between each algorithm or family of algorithms. This technique was first used on citation metrics by Bollen [7], who indicated significant differences between families of metrics including download metrics, citation metrics and traditional Impact Factor [8].

As a result of Principal Component Analysis, it was found that algorithms based on Citation Count group together tightly. This group includes Citation Count, HITs and CoRank-LinkCount. The other Co-Citation based metrics also grouped together based upon their primary property, e.g. being based upon PageRank or involving publication age factors. Perhaps more interesting is that PageRank, the algorithm used by Google, does not relate strongly to any of the algorithms trialled. This is explainable by looking at the results of the individual tests; while many of the Co-Citation based algorithms performed well in the rank order and publication age tests and not in the correlation test, the exact opposite is true of PageRank.

As the pace of scholarly communications quickens, the demand to evaluate research sooner becomes apparent. Metrics such as CoRank-LinkCount have clear value and potential to provide accurate early indication measures that mimic existing metrics while drawing on sources of new data, with the added possibility of becoming the new "gold standard" metric.

[1] Citebase - http://www.citebase.org

References

1. Brody, T.: Evaluating research impact through open access to scholarly communication (2006)
2. Thelwall, M.: Can google's pagerank be used to find the most important academic web pages? Journal of Documentation **59**(2) (2003) 205–217
3. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American society for information science **24**(4) (1973) 265–269
4. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46**(5) (1999) 604–632
5. Brin, S., Page, L., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
6. Hitchcock, S., Brody, T., Gutteridge, C., Carr, L., Hall, W., Harnad, S., Bergmark, D., Lagoze, C.: Open citation linking: The way forward. D-Lib Magazine **8**(10) (2002)
7. Bollen, J., Van de Sompel, H., Rodriguez, M.: Towards usage-based impact metrics: first results from the mesur project. Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (2008) 231–240
8. Garfield, E.: The history and meaning of the journal impact factor. JAMA: the journal of the American Medical Association **295**(1) (2006) 90