Characterizing the Scholar H-index Via Full-text Citation Analysis

Star X. Zhao
Department of Information Resource Management, Zhejiang University, Hangzhou, China
starzhao@zju.edu.cn

Xiaozhong Liu*
School of Library and Information Science, Indiana University, Bloomington, IN.
liu237@indiana.edu

Fred Y. Ye
School of Information Management, Nanjing University, Nanjing, China and Department of Information Resource Management, Zhejiang University, Hangzhou, China.
yy@zju.edu.cn

Abstract

This study proposes a method to characterize the scholar h-index by full-text citation analysis. The method combines the citation context analysis, graph mining, and supervised topic modeling to modify the oversimplified process of citation count, and provides more sophisticated assumptions for the scholar h-index in two aspects: the context of citation and the supervised topic-related measure.

Keywords: full-text citation analysis, h-index, academic impact, text mining, bibliometrics

Background and Objective

The bibliometrics is an important means to characterize scientific publication, scholar, or domain. In this field, the h-index (Hirsch, 2005), considered as one of the most renowned and successful bibliometrics indicators in recent years (Egghe, 2010; Norris & Oppenheim, 2010), has been applied in many aggregative levels (e.g. scholar/journal/institution/country/ science funding) and extensive fields (e.g. finance/network) (Byström, 2011; Zhao & Ye, 2012). Among them, Scholar h-index is the earliest and most used application. A scholar with an index of h means that he/she has published h papers each of which has been cited at least h times. This measure balances the number of the scholar's high cited papers and the number of citations. In the sense of citation analysis it reveals both two significant aspects of scholar's published works: productivity and impact.

However, a major limitation of traditional citation analysis is that the classical method is focusing on citation counts, while ignores the context, topic or motivation of citations. There are different reasons of citing a paper, such as identifying origin, introducing methodology, providing background, giving credit, criticizing others' work and addressing the interestingness (Garfield, 1964; Liu & Rousseau, 2012). However, most previous h-index implementations ignore most of these qualitative features. Another disadvantage of traditional citation analysis occurs as simplifying the multi-citing to one citing. For instance, if paper A cites two or three different texts of paper B, the citations between A and B will just be simplified to one linkage between them. However, intuitively, based on citation frequency and citation context (or topic), citing paper's credit should NOT be evenly distributed to the cited publication, and some citations should be more important than others (Voos & Dagaev, 1976; McCain & Turner, 1989; Liu, Zhang & Guo 2012).

For scholar's h-index, these limitations lead to essential problems of its validity and reliability. First, the scholar's h-index is constructed by the ambiguous citing meaning and topic, i.e., the indicator oversimplifies the citation relationship because a cited paper can make essential or trivial contribution to the citing paper. Second, the calculation of scholar's h-index omits the multiple citing between two papers. In most circumstances the multiple citing indicates their close relevance, thus this process seems not fair and might lose some important information.

Zhao, S. X., Liu, X., & Ye, F. Y. (2013). Characterizing the scholar h-index via full-text citation analysis. *iConference 2013 Proceedings* (pp. 746-749). doi:10.9776/13356

In this proposed research, based on Liu, Zhang and Guo (2012), we employed a supervised topic modeling algorithm, Labeled LDA (LLDA) (Ramage et al., 2009), to infer the publication and citation topic distribution, where each topic is a probability distribution of words and the label of the topic is an author contributed publication keyword. The publication and citation topic probability distributions, then, can be converted to the vertex (publication) prior and edge (citation) transitioning probability distributions to enhance citation network PageRank (with prior distributions) for calculating topical h-index. More specifically, we assume that words surrounding a target citation (citation context) can provide semantic evidence to infer the topical relevance or reason for the target citation, and that a citation network with prior (topic) knowledge can enhance classical bibliometric analysis, i.e. based on the citation context, if a cited paper contributes to the core topic(s) of the citing paper, this cited paper should get more credit from the citing paper (higher transitioning probability). Because each vertex or edge on the citation network is associated with a topic probability distribution, the enhance PageRank can generate an authority vector, and each score in the vector tells the publication or author topical importance, which will, then, be used to calculate author topical h-index.

Methods and Designs

The method proposed in this study combines the citation analysis and text mining to replace the oversimplified process of citation count. It applies a supervised topic modeling algorithm (Labeled LDA) to produce the publication and citation topic distribution, where each topic label is an author contributed keyword and each topic consists of a probability distribution of words. Then, a weighted citation network can be constructed by the publications (nodes) and citations (edges) according to their topic probability distributions. This method is based on that assumptions that: 1) for a target citation, surrounding words (context) can reveal the citation topical motivation; 2) a cited paper which contributes to the core topic(s) of the citing paper should obtain more weights (credits) from the citing paper; and 3) the publication (node) importance can be scored by the citation network which is associated with a topic probability distribution.

For the new scholar h-index based on the full-text citation analysis, we attempt to use the following steps to measure the author topical importance:

(1) In a paper set, analyze all the full text of the paper, and extract all the topics along with their topic labels (author provided keywords).

(2) Construct the weighted citation network, $G = (V, E)$, with two kinds of prior knowledge: *publication topic prior* and *citation topic transitioning probability distribution*.

Each vertex, $v \in V$, on the citation graph represents a publication, with the publication topic prior probability vector $\{p_{z_{key_1}}(v), p_{z_{key_2}}(v), \ldots p_{z_{key_n}}(v)\}$, where $\mathrm{p}_{\mathrm{z}_{\mathrm{key}_t}}(\mathrm{v})$ is the prior probability of vertex v for topic $\mathrm{z}_{\mathrm{key}_t}$ and $\sum_{i=1}^{|V|} p_{z_{key_i}}(v) = 1$. Each edge, $\mathrm{e} \in \mathrm{E}$, on the graph represents a citation connecting $\mathrm{v_i}$ and $\mathrm{v_j}$ ($\mathrm{v_i}$ cites $\mathrm{v_j}$). The topic transitioning vector for each edge is $\{ p_{z_{key_1}}(v_i|v_j), p_{z_{key_2}}(v_i|v_j), \ldots p_{z_{key_n}}(v_i|v_j) \}$, where $p_{z_{key_t}}(v_i|v_j)$ is the probability of transitioning from vertex $\mathrm{v_i}$ to $\mathrm{v_j}$ for topic z_{key_t}.

(3) Compute each scholar's h-index by employing publication topic distribution $p_{z_{key_t}}(v)$ and citation transitioning probability $p_{z_{key_t}}(v_i|v_j)$.

By this new method, a topic-related scholar h-matrix can be set up, as shown in Fig.1.

Topic List

Author List	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	.....
Author 1	5	25	15	2	1	.....
Author 2	0	0	0	19	6	.....
Author 3	1	2	3	2	1	.....
Author 4	0	0	2	5	2	.....
Author 5	19	15	12	17	15	.....
.....	.....	.....	.....	.....	.....	.....

Figure1. The topic-related scholar h-matrix based on full-text citation analysis

There are three merits for this proposed research. First, Labeled LDA, used to characterize publication and citation for this research, is a supervised topic model that constrains LDA by defining a one-to-one correspondence between LDA's latent topics and user tags (keyword metadata). Labeled LDA can directly learn word–tag correspondences, which has been demonstrated to improve expressiveness over traditional LDA with visualizations of a corpus of tagged web pages. It is a promising method to model topics for h-index, and which could be used to optimize the ranking algorithm, and important for result evaluation and interpretation.

Second, unlike classical scholar h-index, our method produces topic based on author h-index scores. Namely, for different topics, a specific author could have different h-index scores. Consequently, the scholar h-index can be compared in the same research topic. It provides much fairer results for the scholars who involve multiple topics or fields.

Last but not least, this new method, considering full text publication and citation transitioning probabilities, may favor authors that make significant contributions but which have not yet received many citations. For instance, our method will grant more credits to new papers and unknown authors if they are making essential contribution to important (high cited) publications. This is very important for academic information retrieval and recommendation systems also.

Dataset and Evaluation

We used 41,370 publications from 111 journals and 1,442 conference proceedings or workshops on computer science for the experiment (mainly from the ACM digital library), where the full text and citations were extracted from the PDF files. The selected papers were published between 1951 and 2011. From these we extracted 28,013 publications' text (accounting for 67.7% of all the sampled publications), including titles, abstracts, and full text. For the other publications, we used the title, the abstract, and information from a metadata repository to represent the content of the paper.

In order to evaluate our work, we will sample a list of topics (with keyword labels). Domain expert will sample some main conference proceedings or journals for each candidate topic. By using classical and this innovative h-index method, we will 1) identify the most important authors from this community; and 2) predict the most important authors (not yet important) in a number of years. MAP and NDCG indicators will be used for this evaluation.

Outlook

Our methods attempt to provide more appropriate assumptions for the scholar h-index in two aspects: the context of citation and the topic-related measure. In future works, we will implement the ideas and designs by using ACM data. We believe that the citation measures should consider more details of the context, and the full-text mining would be a potential tool for this purpose. Theoretically, these methods can be applied for h-index at other aggregative levels also, such as journal, institution or

research field. Although there are still many difficulties to understand and interpret the semantic or motivation of citations accurately and completely, the full-text citation analysis provides the primary insight to observe and characterize the context of citations.

References

Byström, H. (2011). An index to evaluate fund and fund manager performance. *Applied Economics Letters*, 18(14), 1311-1314.

Egghe, L. (2010). The hirsch index and related impact measures. *Annual Review of Information Science and Technology*, 44, 65-114.

Garfield, E. (1964). Can citation indexing be automated?. *In Statistical Association Methods for Mechanized Documentation, symposium proceedings. Washington National Bureau of Standards Miscellaneous Publication*, 269, 189-192.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572.

Liu, X., Qin, J., & Chen, M. (2011). ScholarWiki system for knowledge indexing and retrieval. *In Proceedings of the American Society for Information Science and Technology. American Society for Information Science and Technology*, 1-4.

Liu, X., Zhang, J. & Guo, C. (2012) Full-text citation analysis: enhancing bibliometrics and scientific publication ranking. *Journal of the American Society for Information Science and Technology (In press)*.

Liu, Y.X., & Rousseau, R.(2012). Interestingness and the essence of citation. *Journal of Documentation*,(In press).

McCain, K.W., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics*,17(1-2),127-163.

Norris, M., & Oppenheim, C. (2010). The h-index: a broad review of a new bibliometric indicator. *Journal of Documentation*, 66(5), 681-705.

Voos, H., & Dagaev, K.S. (1976). Are All Citations Equal? Or, Did We Op. Cit. Your Idem? *Journal of Academic Librarianship*, 1(6), 19-21,

Zhao, S.X., & Ye, F.Y. (2012). Exploring the directed h-degree in directed weighted networks. *Journal of Informetrics*, 6(4), 619-630.