How does Citing Behavior for a Scientific Article Change over Time? A Preliminary Study

Ch’ao LU
Department of Information
Management, Nanjing University
of Science and Technology
No. 200, Xiaolingwei, Nanjing,
210094, China
luchao_njust@qq.com

Chengzhi ZHANG
Department of Information
Management, Nanjing University
of Science and Technology
No. 200, Xiaolingwei, Nanjing,
210094, China
zhangcz@njust.edu.cn

Shutian MA
Department of Information
Management, Nanjing University
of Science and Technology
No. 200, Xiaolingwei, Nanjing,
210094, China
mashutian0608@hotmail.com

ABSTRACT

This study is to investigate how citing behavior for a scientific article changes over time. A highly cited article is chosen to collect citation content from its 902 citing articles. Natural Language Processing and content analysis are adopted to encode the original data and complete the statistics of citation behavior. Citation content analysis indexes, like citation mention, length and location, are used to describe the changes of citation behavior over time. The experimental results suggest that there is correlation between citing time with most of the indexes and that citation mention and citation length are declining while the number of the citation co-occurrences is increasing.

Keywords

Citation analysis, citation content analysis, citation behavior.

INTRODUCTION

Since the Scientific Citation Index was proposed, citation analysis has become one of the core theories in Library and Information Science (Garfield, 1964). Based on the assumption that all the literature cited in citing paper are of equal importance, it has been applied to academic evaluation in journals, authors, institutions and etc. The citation analysis has gained numerous studies on how (Bornmann & Daniel, 2008; Burrell, 2003) and why (Moravcsik & Murugesan, 1975; Small, 1978; Teufel, Siddharthan, & Tidhar, 2009) scholars cite papers. The assumption is also accepted by “half-life” of periodical articles since 1960 (Bernal, 1960; Burton & Kebler, 1960) to evaluate the obsolescence of periodical literature in one specific field or discipline. Considering all the documents in one discipline as a whole, the “half-life” mainly instructs collection building and weeding of library from macroscopic perspective. However, it is difficult to assess the obsolescence of one single scientific article with this theory (Pasterkamp, Rotmans, de Kleijn, & Borst, 2007). A scientific article gets less important and influential with time passing by. If more details about the citations could be detected, their importance might be calculated and the obsolescence would be assessed. Scholars, especially the novices would better understand the value of articles and properly utilize them in scientific researches or in manuscripts. Since the Natural Language Processing Technologies are widely adopted and full-text scientific article data are open accessible, detecting the citation details becomes possible. The citation content analysis has redrawn scholars’ attention and become a new trend in Library and Information Science (Ding et al., 2014; Zhao & Strotmann, 2014). Existing related studies have adopted several citation content indexes like, citation mention (Ding, Liu, Guo, & Cronin, 2013; Wan & Liu, 2014a, 2014b; Zhao & Strotmann, 2015), citation location (Ding et al., 2013; Gabb, Lucic, & Blake, 2015; Hu, Chen, & Liu, 2013; Wan & Liu, 2014a, 2014b) to represent the citation content. As far as we are concerned, there is no study considering the current influence of a scientific article with indexes mentioned above. Therefore, this study is to represent the influence of a scientific article with indexes and to find out how these indexes change over time. Then we can evaluate the influence or value of an article in different points of time. The changes can also be used to detect the ages of the scientific articles in a micro-perspective way and guide novices to properly using them.

ASIST 2015,November 6-10, 2015, St. Louis, MO, USA.

METHODOLOGY

Data Collection

A highly cited paper in Library and Information Science, “*An index to quantify an individual's scientific research output*” (Hirsch, 2005) first introducing the h-index, is chosen to collect full-text papers which cited it. 1,294 papers are retrieved from Web of Science Core Collection. The latest citing paper came out on November 2014. We have got 1,050 full-text articles, and 249 of them are unavailable. Papers like letters, editorials, communications and comments are removed and the final data set contains 902 full-text articles, including 255 PDFs files, 626 html files and 21 XML files.

We collect the citation indexes to represent the citation content as follows: citation date, citation mention, citation location, citation content, citation length, and citation co-occurrences.

Citation Mention is defined as how many times the paper is cited/mentioned in one article (Ding et al., 2013; Wan & Liu, 2014a). In this paper, we count the citation mention of the highly cited paper in all these 902 articles.

Citation Location is where the citation content located. It has been proved that citation locations are related to the citing behavior (Ding et al., 2013; Hu et al., 2013; Wan & Liu, 2014a). Different locations of the cited paper in the article may weigh differently (Hu et al., 2013). Based on previous works, the citation locations are classified into eight types: Introduction, Literature Review/Related work, Methodology, Results, Discussion, Conclusion, NA (not available), and Others in this study. NA encodes the papers where reference list contains the cited paper but does not mention it in the main part. Others are used to encode the papers that have mentioned the cited paper in the main body while the structure is unrecognizable due to some writing styles, like papers with no headings. According to the encoding methods, all the citation sentences will be given a certain location type.

Citation Length is the length of the citation content. Previous studies (Ding et al., 2013; Hu et al., 2013; Jeong, Song, & Ding, 2014; Liu & Chen, 2013; Wan & Liu, 2014a, 2014b) have extracted only one sentence of the citation content and we also use this method. The length of the citation content can suggest the importance of the cited paper.

Citation Co-occurrence means that the highly cited paper was cited along with other papers in the same sentence. This index may suggest the contribution of the cited paper.

Among these indexes, citation content needs collecting manually to guarantee the accuracy of the data. There are different types of files of the full text (including PDF, Html, Xml) and citation styles (Author-Data, Numbered, and so on). It is hard to precisely extract the citation content with programming like other studies did (Ding et al., 2013; Hu et al., 2013). Three students are responsible for citation content collection, and accordingly then the data are rechecked. Other indexes can be processed and calculated automatically. Finally, this study has collected 1,690 citation sentences in 902 full-text articles.

Data Analysis

All the unrecognizable codes and spam codes (like wasted spaces and html tags) are removed from the sentences. Then these data can be used to calculate the length and find out whether the cited paper is co-cited with other papers in the same sentence or not. The length of citation sentences are computed according to the formula (1):

$$len\_re = len\_s \, / \, len\_f \qquad (1)$$

Where *len_s*, *len_f* represents the length of sentence and the full text where we extracted sentences respectively.

In the data analysis part, the citation frequency, mention and length, distribution of citation location and the number of citation co-occurrences are calculated by years.

RESULT AND DISCUSSION

This section will introduce findings according to the six indexes and discuss about them. The findings will help answer the questions as mentioned in Methods and Data. Findings are unfolded in three aspects: citation frequency and citation mention; distribution of locations; and citation content: length of citation sentence, topics, and number of citation co-occurrences.

Citation Frequency and Citation Mention

Figure 1 shows the citation frequency and the citation mention the cited paper received in each year. Figure 1 suggests since 2006 the citation frequency that paper received has increased year by year until 2014, during which a slight decline appeared between 2010 and 2011. After 2014, the citation frequency sharply declined. The citation mention showed the similar trend. By comparing the two curves, it suggests that citation mention usually is higher than citation frequency (Ding et al., 2013; Wan & Liu, 2014a) in the long term. The average citation mention data of per citing paper in each year (the citation mention divided by the citation frequency of the year) is presented in Figure 2. This curve suggests that the citation mention was obviously on the decline through years except for 2006. As shown in Figure 2, the highest average citation mention was 2.6 on the second year after the cited paper getting published. In 2012 the cited paper received the lowest average citation mention (1.7), followed by 2014 (1.71) except for year 2006.

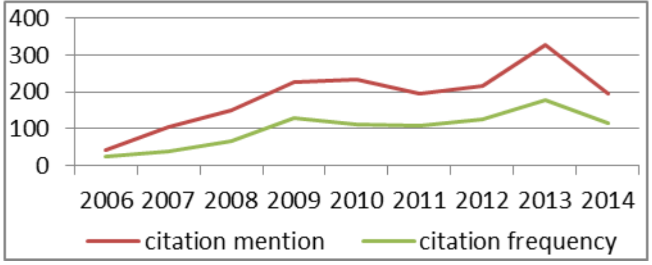

Figure 1. Citation frequency and citation mention between 2006 and 2014.

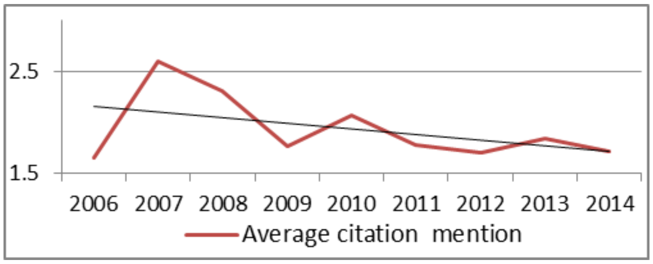

Figure 2. Average citation mention from 2006 to 2014

In 2007, the average citation mention peaked while the citation frequency peaked in 2013. And in 2009 the citation frequency was increasing while average citation mention fell into a bottom. Possible reason may be that in 2009, most papers just mentioned the cited paper as a specific application to statistics or as a background because of its reputation.

Citation Location

Figure 3 presents the locations of the cited paper appeared through years. More than 1/3 of papers cited the paper in Introduction part, which shares the similar results with (Hu et al., 2013).

When considering time, the Introduction and Literature Review took up about a major rate of location distribution and indicated a slow decline. More citations appeared in Method part. In Result part, the trend was on a slow increase after a sharp decrease. Oppositely, Discussion part suggested a growing trend then got a decreasing trend. The citation appearing in Conclusion part was on a declined path from 2006 to 2014.

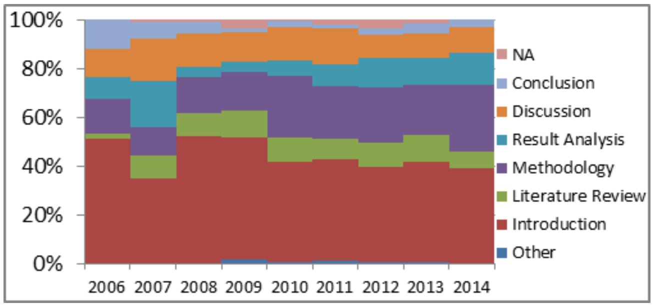

Figure 3. The location change through years

Citation Content

All the length of sentences is divided by the length of their articles. Results are presented in Figure 4. The curve suggests that by years the length of the citation sentences was getting shorter and shorter. In 2009, the curve reached a peak. One possible reason is articles published in 2009 were shorter than those published in other years; another reason is that papers published in 2009 were likely to cite the citation with less mention but longer single sentences considering the average citation mention in Figure 2.

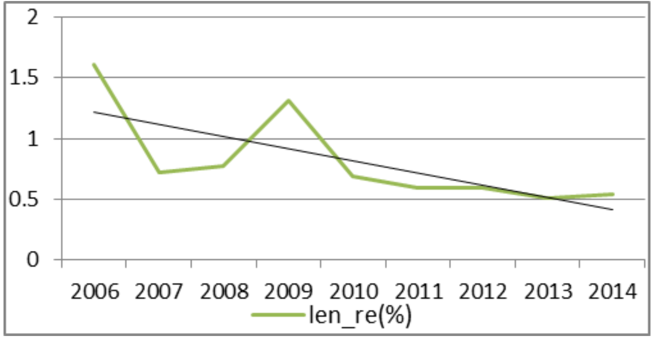

Figure 4. The average length and relative length of the citation content from 2006 to 2014.

The average number of the citation co-occurrences and the standard deviations from 2006 to 2014 are curved in Figure 5. The figures imply that more and more citations shared the single sentences in the articles with years going by. In previous co-author analyses, the authors evenly contribute to the scientific articles. According to Figure 4, the average length of the citation sentences was increasingly shorter. That means the cited papers we used in this article has made less influence on the application articles with time fleeting. The STDEV in 2014 suggests most of the papers tended to cite the article along with more articles. The largest number of the citation co-occurrences is 28 articles in one sentence alone.

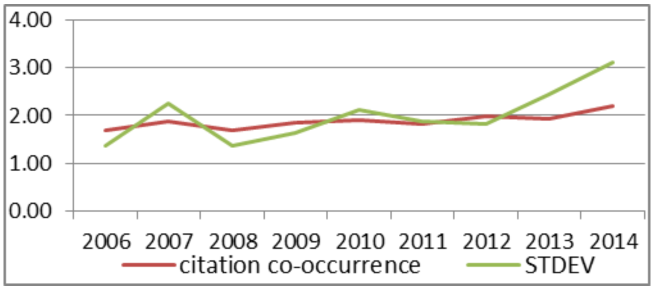

Figure 5. Average citation co-occurrences in the citation sentences and the STDEVs from 2006 to 2014.

Citation mention, Location, Length of citation content, and citation co-occurrences reveal the passed time.

Observations above suggest that the citation content indexes changed over years. With time going by, average citation mention of each year appeared a continuing decline after reaching the peak and citation frequency of the years showed a similar trend as the trend was retarded. Possible explanation may be that the citation frequency changes more slowly than citation mention on the former citing behavior. Citation location is another index that suggests the change of citing behavior. In earlier years, authors tended to cite the paper more in Introduction, Literature review, Discussion and Conclusion; while in later years, authors seemed to cite the article more in Methodology and Result. Possible reason is that the growing knowledge of cited paper makes authors introduce the h-index less in their papers. The authors do not need to talk about the cited paper in Discussion or Conclusion in details. They may choose more important or newer works in Introduction and Literature review. As an effective method, h-index can be mentioned more in Method part or directly appears in Result part when related data are analyzed.

Changes also took place in the content part of the citations. Average length of the citation sentences got shorter and more and more other citations co-occurred with h-index over time. Average length of the citation sentences reflects how specified the article is cited in other papers. The declining tendency suggested that more and more authors had chosen to epitomize h-index in their manuscripts. One reason is that h-index is well-known to scholars in related fields. Another reason explains that more studies adopted the h-index as a mature method to analyze data rather than optimize it with more details. The reason also explains why average citation sentence length did not change much regardless of the length of articles containing them. Furthermore, more and more other citation co-occurred with h-index, such as g-index and other optimizations on h-index. With years passing, authors cited the h-index along with more related optimization studies instead of citing it solely. The co-occurrence well proved that scientific papers are losing their impacts with more related outstanding studies showing up over time. Authors will not neglect these late important studies in their manuscripts.

CONCLUSION

This study is to find the changes of citing behaviors over time. Findings reveal that the citing behaviors did change over time shown by citation indexes: mention, length, location, and the number of the citation co-occurrences. Most of the citation indexes are deeply investigated to reveal the changes. Observations of this study may be used to build single article assessing system to evaluate the age or the real-time influence of the target article with more efforts and help users, especially the novices to read and use the scientific articles.

However, this study still needs improving. For example, this study only chose one scientific article to realize our intention. The limited sampling may lead to some biases though the observations can be supported by other related studies. Another major limitation is that all the original data are collected manually. This method will limit the application of the findings to the real problems. So next efforts will be made to: 1) investigate large-scaled sampling cited document sets to generalize the findings in this preliminary study; 2) develop software or toolkits to collect the citation data like full-text articles and citation content and process them automatically; 3) ultimately develop the assessment system mentioned above for use in real life.

ACKNOWLEDGMENTS

This work was supported in part by National Social Science Fund Project (No.14BTQ033) and Major Projects of National Social Science Fund (13&ZD174).

REFERENCES

Bernal, J. (1960). *The Transmission of Scientific Information: A User's Analysis.* Paper presented at the Proceedings of the international conference on scientific information, 77-95.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80.

Burrell, Q. L. (2003). Predicting future citation behavior. *Journal Of the American Society for Information Science And Technology, 54*(5), 372-378.

Burton, R. E., & Kebler, R. (1960). The “half - life” of some scientific and technical literatures. *American documentation, 11*(1), 18-22.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7,* 583–592.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology, 65*, 1820–1833.

Gabb, H. A., Lucic, A., & Blake, C. (2015). A Method to Automatically Identify the Results from Journal Articles. *iConference 2015 Proceedings*.

Garfield, E. (1964). Science Citation Index-A new dimension in indexing. *Science, 144*(3619), 649-654.

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics, 7*, 887–896.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics, 8*, 197–211.

Liu, S., & Chen, C. (2013). The differences between latent topics in abstracts and citation contexts of citing papers. *Journal Of the American Society for Information Science And Technology, 64*, 627–639.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social studies of science, 5*(1), 86-92.

Pasterkamp, G., Rotmans, J. I., de Kleijn, D. V., & Borst, C. (2007). Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles. *Scientometrics, 70*(1), 153-165.

Small, H. G. (1978). Cited documents as concept symbols. *Social studies of science, 8*(3), 327-340.

Teufel, S., Siddharthan, A., & Tidhar, D. (2009). *An annotation scheme for citation function.* Paper presented at the Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, 80-87.

Wan, X., & Liu, F. (2014a). Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology, 65*, 1929-1938.

Wan, X., & Liu, F. (2014b). WL-index: Leveraging citation mention number to quantify an individual's scientific impact. *Journal of the Association for Information Science and Technology,* 65(12), 2509-2517.

Zhao, D., & Strotmann, A. (2014). In - text author citation analysis: Feasibility, benefits, and limitations. *Journal of the Association for Information Science and Technology, 65*(11), 2348-2358.

Zhao, D., & Strotmann, A. (2015). Dimensions and uncertainties of author citation rankings: Lessons learned from frequency - weighted in - text citation counting. *Journal of the Association for Information Science and Technology*.