

Understanding the Impact Change of a Highly Cited Article: A Content-based Citation Analysis

Chao Lu^{1,2}, Ying Ding^{2,3,4}, Matthew Schnaars², Chengzhi Zhang^{1,5,6*}

1. *Department of Information Management, Nanjing University of Science and Technology, Nanjing, 210094, China*

2. *School of Informatics and Computing, Indiana University Bloomington, Bloomington, 47401, United States*

3. *of information management, Wuhan University, Wuhan, 430072, China*

4. *Library, Tongji University, Shanghai, 200092, China*

5. *Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing University, Nanjing, 210093, China*

6. *Jiangsu Collaborative Innovation Center of Social Safety Science and Technology, Nanjing, 210094, China*

Abstract

Researchers tend to cite highly cited articles, but how these highly cited articles influence the citing articles has been underexplored. This paper investigates how one highly cited essay, Hirsch's "h-index" article (H-article) published in 2005, has been cited by other articles. Content-based citation analysis is applied to trace the dynamics of the article's impact changes from 2006 to 2014. The findings confirm that citation context captures the changing impacts of the H-article, notably impact decay over time in several ways. In the first two years, average citation mention of the H-article increased, yet continued to decline with fluctuation until 2014. In contrast with citation mention, average citation count stayed the same. The distribution of citation location over time also indicates three phases of the H-article: "Discussion," "Reputation," and "Adoption". Based on their locations in the citing articles and their role in different phases, topics of citation context shifted gradually when an increasing number of other articles were co-mentioned with the H-article in the same sentences. In contrast to traditional means of citation analysis, which only uses bibliographic data and does not weigh citation impact over time, these observations show that within the context of the citing articles the H-article's impact manifests in various ways that continued to shift over the nine years.

Keywords:

Citing behavior; content-based citation analysis; citation context; impact decay

MSC:

68T30

JEL:

D830

* Corresponding author. E-mail: zhangcz@njust.edu.cn

1. Introduction

Citation count has been used as the de facto standard to measure the impact of an article, a researcher, or an institution. But how a highly cited article impacts a field and how these influences change over time has not been well explored. It is important to understand how a paper’s impact grows, diffuses, and fades so as to: (1) facilitate scholarly communication and understanding of research obsolescence trends; (2) detect impact changes in different domains and factors of influence; and (3) differentiate the impact of papers even when they have roughly the same number of citations. Using citation counts alone to measure the impact of a paper is a limited approach, in that it ignores impact changes, especially relevant for highly cited articles (MacRoberts & MacRoberts, 1989; Voos & Dagaev, 1976; Aksnes, 2003). Researches on citation contribution (Cano, 1989; Case & Higgins, 2000; Garfield, 1964; Lipetz, 1965; Moravcsik & Murugesan, 1975; Voos & Dagaev, 1976) have found that perceived contributions of an article vary within the text of citing articles. Lipetz (1965) presented 29 categories of citation motivations in physics literature. As different perceived “contributions” in this sense may imply varied impacts of the cited article, this finding only confirms that it is problematic to assume that all citations in an article are interpreted by the citing article in the same manner. Impact decay of articles over time has also been investigated for decades. As scientific knowledge and contributions are dynamic and quickly changing in light of new discoveries, it is important to acknowledge nuanced factors of an article’s influence, including its changing impact over time. Furthermore, numerous studies (Cano, 1989; Moravcsik & Murugesan, 1975; Small, 1978; Voos & Dagaev, 1976) confirm that analyzing citation context can help differentiate various motivations and functions of citations. So how, where, and how many times an article is positioned in relation to other works is a relevant factor to consider when exploring its potential impact, including its impact decay over time. We choose J. E. Hirsch’s (2005) highly cited article, “An index to quantify an individual’s scientific research output” to illustrate this issue of citation change over a nine-year period (where the article is referred as “H-article” and the index as “h-index” hereafter). Figure 1 shows the citation patterns of two articles published in 2008 (Article A) and 2014 (Article B) that cite the H-article to support their arguments (Case & Higgins, 2000). In the citing sentence of Article A where the H-article is mentioned, no other article is co-mentioned; while in the citing sentence of Article B, the H-article is co-mentioned together with 15 other articles. We can therefore assume that the H-article should make a greater contribution to Article A (e.g., 1/1) than Article B (e.g., 1/16).



Figure 1. Different citation contexts of the H-article.

Citation context, which is the contextual information surrounding a citation in the citing articles, can be categorized at the syntactic and semantic levels (Angrosh, Cranefield, & Stanger, 2012; Kaplan, Tokunaga, & Teufel, 2016; Wan & Liu, 2014a; Zhang, Ding, & Milojević, 2013). Syntactic citation context includes citation mention (how many times an article has been mentioned in a citing article) (Ding, Liu, Guo, & Cronin, 2013), as well as citation location (where these references are mentioned in the citing article) (Hu, Chen, & Liu, 2013). Semantic citation context includes citation topic, which captures the topic distribution of citation contexts. While these contextual features work well in detecting citation sentiment (Voos & Dagaev, 1976), citing behavior (Small, 1978), and citing motivation (Moravcsik & Murugesan, 1975), they have not been explored in depth to detect impact change of articles over time.

This paper addresses this gap by applying content-based citation analysis to examine the dynamics of the H-article's impact changes as reflected in citation context shifts from 2006 to 2014, where we divide the period into three phases: "Discussion," "Reputation," and "Adoption." Section 2 contains a brief literature review, section 3 discusses data and methodology, section 4 describes and discusses results, and section 5 draws conclusions and points out future research.

2. Literature Review

Macro-level impact decay

The impact decay of articles has been investigated for many decades at the macro level. Burton and Kebler (1960) first used the concept of "half-life" from physics to describe scientific articles' obsolescence function or impact delay, which they defined as "the time during which one-half of all the currently active literature was published" (p. 19). Half-life has been widely adopted by libraries to weed out literature and construct collections (Line & Sandison, 1974; Schlachter, 1988), or to enhance library services and technical support (Burton & Kebler, 1960; Brown, 1980; Tsay, 1998). For example, Charles (1988) pointed out that the citation count in the Science Citation Index should be normalized to achieve a more interpretable half-life for astronomical papers. Even though the impact decay or obsolescence of scientific publications has been studied, researchers have made a limited effort to analyzing the ways in which the impact of an article actually changes over time., e.g., how, where, and how many times one cited article is mentioned in the body of citing articles, which provides information on how the impact of the cited article changes over years.

Citing behavior

There are many reasons for authors' varying citation practices and trends. Lipetz (1965) identified 29 categories of citation practice motivations, organizing them into four clusters: (1) original scientific contribution or intent of the citing paper, (2) contributions of the citing paper other than its original scientific contribution, (3) identification of relationships between the citing paper and the cited paper, and (4) scientific contribution of the cited paper to the citing paper. Similarly, Moravcsik and Murugesan (1975) divided the citations of physics articles into four categories: conceptual/ operational, evolutionary/ juxtapositional, organic/ perfunctory, and confirmative/ negational, where they found that one third of the references were redundant, one seventh were negational, and two fifths were perfunctory. Case and Higgins (2000) used a questionnaire to identify why authors cited highly cited articles, and found that authors do so to promote their own authority, or to claim that the highly cited article deserves attention or criticism.

Citing behavior thus varies considerably in different articles, where some are heavily cited but are only mentioned sparingly in the citing articles, and others that receive only a moderate number of citations are frequently mentioned by the citing papers (Zhao & Strotmann, 2015). Generally, more than one third of citations occur in the beginning of the citing articles, most of which are perfunctory. Some citations are located in the Method section for operational use and some in the Result and Discussion sections for confirmative use. These diverse locations indicate a range of citation function, and to some extent imply citation impact in the citing articles (Cano, 1989). Different citation contexts that contain the same cited article may also discuss different topics. For example, Ruane and Tol (2008) cited the H-article to point out the function of the h-index. Hack, Crooks, Plohman, and Kepron (2014) initially mentioned the H-article to discuss the function of the h-index, then to define the h-index, and finally to compare the h-index with other indicators. Some articles only mentioned the H-article (Pathak & Bharati, 2014) while others

referenced it along with many other articles (Venable et al., 2014).

Content-based citation analysis

Content-based citation analysis (CCA) focuses on the features of citation context (e.g. mention and location) to differentiate scholarly impact (Ding et al., 2014; Small, 1978; Teufel, 2000). Small (1978) posited that citations are the carriers of specific concepts or topics from the cited articles, which help the concepts interact and influence each other (Liu, Zhang, & Guo, 2013), pointing that we could misunderstand the contribution of the cited articles to the citing articles without taking citation context into consideration. Voos and Dagaev (1976) found that citation mention and citation location analysis help identify different types of citation contributions, where citation count alone fails to detect such nuances. Similar studies were done by Moravcsik and Murugesan (1975) and Cano (1989), who found that one third of citations are located at the beginning of the citing articles. Although citation locations suggest different contributions of the cited articles, all these studies rest on small samples or manual data collection, which are hard to generalize.

Advances in natural language processing (NLP) technologies make it possible to semi-automatically investigate the features of citation context in large-scale, full-text articles (Ding et al., 2014; Ding & Stirling, 2016). Content-based citation analysis has been further applied in solving various problems related to author co-citation analysis (Jeong, Song, & Ding, 2014; Kim, Jeong, & Song, 2016), author ranking (Zhao & Strotmann, 2015) and impact evaluation (Ding et al., 2013; McKeown et al., 2016; Wan & Liu, 2014b). New researches show that combination of features of citation context has a better potential than citation count to accurately evaluate citation contribution (Ding et al., 2013; Hu et al; Wan & Liu, 2014a) which, however, cannot be easily generalized. Some scholars have suggested that the citation topics may not play a major role in an article’s impact over time. For example, Small, Tseng, and Patek (2017) recently argued that when the cited articles are highly cited, their citations become standard symbols and the concepts they carry remain unchanged, but they did not provide temporal evidence to support this argument. We agree with scholars who believe that the meaning of citation context in a highly cited article can indeed change over time, and should thus be investigated to identify impact shifts in the citing articles. Even though many studies report a diversity of citation motivations and patterns, little attention has been given to how the impacts of highly cited articles change over time. To fill this gap, we use the citation context of the H-article (Hirsch, 2005) to quantitatively analyze specific features of the data collected (Figure 2) and reveal its impact change in the nine-year period from 2006 to 2014.

3. Methodology

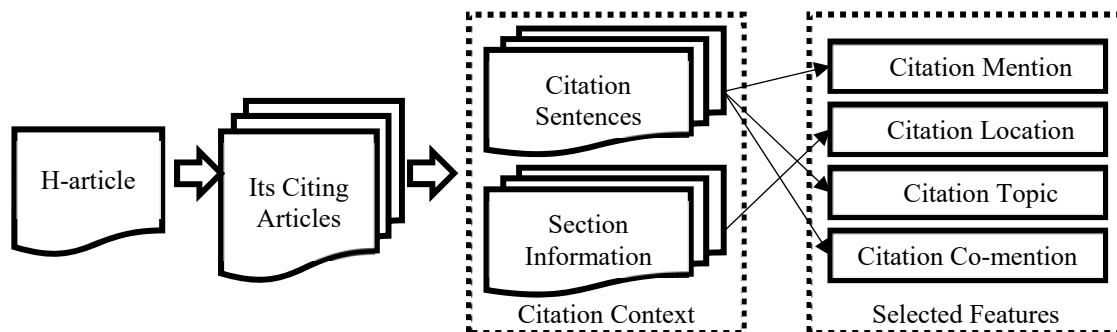


Figure 2. Overview of data collection and selected features.

Data

Cited article data. In this study, we use Hirsch’s 2005 paper entitled “An index to quantify an individual’s scientific research output” (H-article) as the example of a highly cited and influential article. The h-index has been confirmed in the last decade to be of great importance for evaluating individuals’ productivity and impact. Having received a large number of citations in Web of Science (WoS), Hirsch’s seminal paper has continued to attract and influence many scholars from diverse domains.

Full text of citing articles. We collect the full-text articles published between 2006 and 2014 that cite the H-article from the WoS core collection. Articles without full text are omitted. In total, we use 763 full-text citing articles.

Citation context. In examining citation context, studies usually identify the one sentence that contains the citation as the citation content, and the section where it is located as the citation location (Ding et al., 2013; Hu et al., 2013; Jeong et al., 2014; Tang, Wan, & Zhang, 2014; Wan & Liu, 2014a, 2014b). We thus present the citation contexts of a sample article (Table 1) that mention the H-article nine times: once in the Introduction, once in Literature Review, and the rest in the Results and Discussion sections. In total, 1,476 citation contexts from 763 full-text articles are collected (Table 2). As Table 2 shows, the ratio of citation sentences to number of citing articles in each year is in a fluctuating decline after a short increase in 2006 and 2007.

Table 1. Citation context sample.

| ID | Citing Article ID | Year | Title | Citation Context | |
|----|-------------------|------|------------------------------|---|----------|
| | | | | Citation sentence | Location |
| 1 | 1 | 2007 | Soil science and the h index | HIRSCH (2005) suggested the h (Hirsch) index as a measure of scientific 'output'. | Intro.* |
| 2 | | | | The typical h index depends on the discipline or field of science. The h index of an individual scientist is influenced by: the size or number of scientists in the field, the number of papers produced by the scientists in the field, the average number of citations in the field, and the age of the scientist (HIRSCH, 2005). | Lit. |
| 3 | | | | On the other hand, HIRSCH (2005) argues that the larger the field, the larger the number of scientists to share a larger number of citations, so typical h values should not necessarily be larger | R&D |
| 4 | | | | HIRSCH (2005) suggested the relation between h and the number of total citations c, tot N is given by: | R&D |
| 5 | | | | Evidently h is related to the age of the researchers, a relation with age is proposed by HIRSCH (2005): | R&D |
| 6 | | | | HIRSCH (2005) found m 1 characterising a successful scientist, and m 2 for outstanding scientists. | R&D |
| 7 | | | | HIRSCH (2005) also defined c as the average number of citations per paper per year with the following relationship: | R&D |
| 8 | | | | According to HIRSCH (2005) realistically $c > p$, where most contributions to $N_{c,tot}$ is from the highly cited papers (the h papers that have the number of citations $> h$). | R&D |
| 9 | | | | The maximum h index we found was 51 whereas in biology and physics it is over 100 (HIRSCH, 2005). | R&D |

Note: Intro. represents Introduction, Lit. is Literature Review, and R&D is Results and Discussion.

Table 2. Numbers of citing papers collected each year.

| Year | # of citing articles | # of citation sentences | Ratio of citation sentences |
|------|----------------------|-------------------------|-----------------------------|
| 2006 | 17 | 32 | 1.88 |
| 2007 | 23 | 70 | 3.04 |
| 2008 | 60 | 141 | 2.35 |
| 2009 | 88 | 148 | 1.68 |
| 2010 | 97 | 211 | 2.18 |
| 2011 | 109 | 208 | 1.91 |
| 2012 | 125 | 211 | 1.69 |
| 2013 | 133 | 261 | 1.96 |
| 2014 | 111 | 194 | 1.75 |

Features

A set of features is selected herein to describe the impact change in the citation context, which contains two categories: syntactic features and semantic features.

Syntactic features. The syntactic features include an article’s citing article count, citation mention, and citation location. For each citing article, the citing article count always equals one, because the H-article can only be noted once in the reference list of a citing article. *Citation mention* is the number of times that the H-article is mentioned (e.g., “HIRSCH (2005) suggested ...” (Minasny, Hartemink, & McBratney, 2007, p.258).) within the full text of the citing article (Ding et al., 2013; Wan & Liu, 2014a). *Citation location* is the place where a cited paper in the citing article is noted (Hu et al., 2013), such as the Introduction (Intro.), Literature Review (Lit.), Methodology (Meth.), Results, Discussion (Disc.), or Conclusions (Con.). We use the section information to calculate the citation location distribution and the number of distinct locations in citing articles published in each year between 2006 and 2014. For example, in Table 1, the distribution of citation location in the sample article is shown as 11 percent (Intro.), 11 percent (Lit.), and 78 percent (R&D), making three distinct citation locations.

Semantic features. The semantic features include citation co-mention and citation topic. *Citation co-mention* means the number of other citations besides the H-article that are co-mentioned in the same citation sentence of the H-article (Wan & Liu, 2014a). For example, in the first sentence of the sample article in Table 1, the citation co-mention equals one because only the H-article is mentioned. The average citation co-mention of the sample article is also one ($9/9=1$). *Citation topic* is the topic distribution of the citation sentences extracted from the citing articles (Liu et al., 2013). This study uses citation content to extract topics (Figure 2). We adopt these two co-mention and topic citation categories to analyze the H-article’s impact change over time.

Data Analysis

For data analysis, we select several indicators to measure article count, citation mention, citation location, citation co-mention and citation topic. The indicators and their features are shown in Table 3. Table 4 shows one example and explains how we calculate these indicators.

Table 3. Indicators.

| Indicator | Features |
|---|----------------------|
| Average Citing Article Count | Citing Article Count |
| Average Citation Mention | Citation Mention |
| Citation Location Distribution | Citation Location |
| Average Number of Distinct Citation Location | Citation Location |
| Top 30 Keywords of Citation Sentences | Citation Topic |
| Topic Similarity Based on Topics in Fixed Year | Citation Topic |
| Topic Similarity between Every Two Continuous Years | Citation Topic |
| Average Citation Co-mention | Citation Co-mention |

Table 4. Data sample in 2006.

| Article ID | Citing Article Count | Citation Mention | Number of Distinct Locations | Average Citation co-mention |
|------------|----------------------|------------------|------------------------------|-----------------------------|
| 1 | 1 | 2 | 2 | 1.5 |
| 2 | 1 | 3 | 2 | 2 |
| 3 | 1 | 4 | 3 | 2.5 |
| 4 | 1 | 2 | 1 | 1 |
| 5 | 1 | 1 | 1 | 3 |

(1) Average Citing Article Count:

$$\text{Average Citing Article Count} = \frac{\sum_{i=1}^n \text{Citing Article Count}_i}{N} \quad (1)$$

In Table 4, the Average Citing Article Count in 2006 = $\frac{1+1+1+1+1}{5} = 1$

(2) Average Citation Mention:

$$\text{Average Citation Mention} = \frac{\sum_{i=1}^n \text{Citation Mention}_i}{N} \quad (2)$$

In Table 4, the Average Citation Mention in 2006 = $\frac{2+3+4+2+1}{5} = 1.2$

(3) Average Number of Distinct Citation Location (ADCL):

$$\text{ADCL} = \frac{\sum_{i=1}^n \text{Number of Distinct Location}_i}{N} \quad (3)$$

In Table 4, the ADCL in 2006 = $\frac{2+2+3+1+1}{5} = 1.8$

(4) Average Citation Co-mention:

$$\text{Average Citation Co-mention} = \frac{\sum_{i=1}^N \text{Average Citation Co-mention}_i}{N} \quad (4)$$

In Table 4, the Average Citation Co-mention in 2006 = $\frac{1.5+2+2.5+1+3}{5} = 2$

(5) Citation Location Distribution:

Table 5 shows the section information of all the citation mentions of the H-article, so that we can examine the location distributions of these references over time.

Table 5. Location distribution of citation mentions from 2006 to 2014.

| Year | Intro. | Lit. | Meth. | R&D | Con. |
|------|--------|------|-------|-----|------|
| 2006 | 15 | 1 | 4 | 3 | 5 |
| 2007 | 25 | 3 | 8 | 13 | 7 |
| 2008 | 77 | 16 | 21 | 4 | 7 |
| 2009 | 77 | 22 | 18 | 6 | 2 |
| 2010 | 84 | 19 | 59 | 12 | 6 |
| 2011 | 92 | 17 | 41 | 19 | 3 |
| 2012 | 81 | 21 | 49 | 26 | 5 |
| 2013 | 114 | 26 | 49 | 29 | 12 |
| 2014 | 73 | 13 | 59 | 24 | 6 |

Note: The number 15 in 2006 means this year 15 sentences in total mention the H-article in the Introduction.

(6) Top 30 keywords of Citation Sentences:

Topic extraction. Three main methods and algorithms are available to extract document topics: TFIDF (Salton & Buckley, 1988), LDA (Blei, Ng, & Jordan, 2003), and LSA (Dumais, 2004). The first two are most frequently used (Alsaad & Abbod, 2015; Hu et al., 2015; Lee, Lo, Chen, Lin, & Wang, 2015). Both LDA and TFIDF are applied to extract topics in this research, but we finally use the results from TFIDF because it generates better interpretable results than LDA. The top 30 words ranked by TFIDF values are therefore used to represent the topics in each year. General, special, and high-frequency words are removed (e.g. the h-index). The formula for computing the TFIDF values of words is as follows:

$$\text{TFIDF}_w = T f_w \times \log_2 \left(\frac{Doc}{D f_w} \right) \quad (5)$$

$T f_w$ is the frequency that the word w appears in a set of keywords; Doc represents the number of documents in the whole document set; and $D f_w$ denotes the number of documents that contain the word w .

Topic Similarity. Jaccard coefficient is used to track changes of topic similarity at the topic level between two years (Formula (6)):

$$\text{Jaccard}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \times 100\%, (X \cup Y \neq \emptyset) \quad (6)$$

Here, X and Y represent a topic set extracted by TFIDF in different years, respectively. $|X \cap Y|$ represents the number of the keywords the two sets share; $|X \cup Y|$ shows the amount of all the distinct elements the two sets contain.

(1) Topic Similarity Based on Topics in the Fixed Year:

The year 2006 is selected as the fixed year and Formula (6) is used to compute the topic similarity from 2007 to 2014 with 2006 to detect topic differences, such as similarity between 2006 and 2007, and similarity between 2006 and 2008. This approach helps us locate the topic shift from the initial citation context in 2006, one year after the H-article was published.

(2) Topic Similarity between Every Two Continuous Years:

Topic similarity between every two continuous years is calculated using Formula (6) as well, such as topic similarity between 2006 and 2007, and topic similarity between 2007 and 2008. This similarity

component compares the topics of the citation contexts between two continuous years to identify topic shifts from the previous year.

4. Results Analysis and Discussion

Citing Article Count and Citation Mention

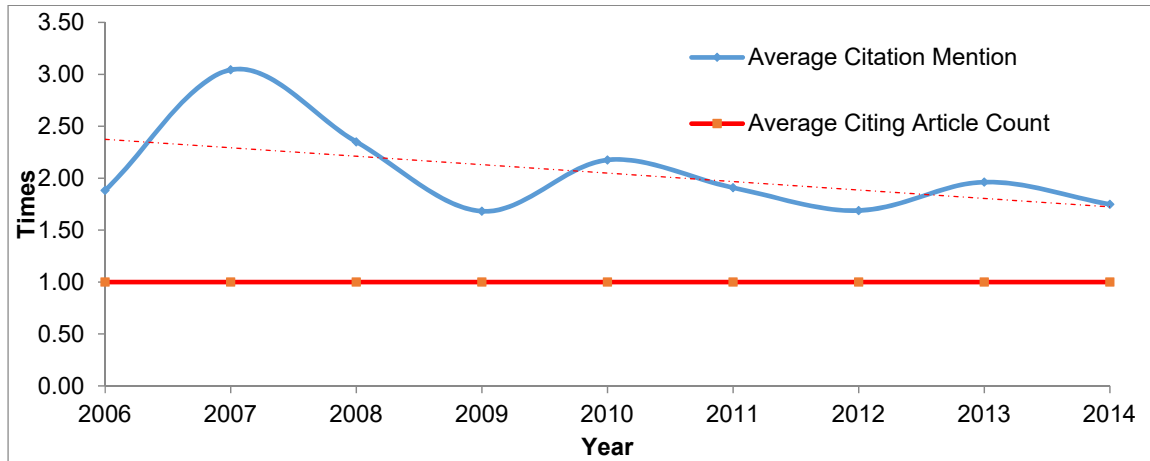


Figure 3. Article counts and citation mentions 2006 to 2014.

We plot the average citing article counts and average citation mentions in Figure 3. The average citing article counts equal one from 2006 to 2014. By contrast, the average citation mentions change in count every year and peak at three in 2007 before fluctuating to below 1.7 in 2009 and 2012. That indicates that the citing papers all mention the H-article more than once (similar to the findings in recent researches (e.g., Ding et al., 2013; Hu et al., 2013)) and less frequently after 2007.

Citation Location

Figure 4 presents the location distribution of citation mentions. Generally, more than 40 percent of citation mentions appear in the Introduction of the article every year, with small fluctuations. We take citation mentions in 2008 as an example to show usage of the H-article in citation context. We find that 43 out of 60 citing articles mention the H-article in the Introduction 77 times, 1.8 times per article on average.

Out of the 77 mentions, 22 simply note that the H-article is popular, e.g. “Since Hirsch’s first publication of the h-index in 2005 [9], this new measurement of academic impact has generated widespread interest” (Baneyx, 2008, p.364). A total of 45 of the 77 mentions introduce the definition or function of h-index, e.g. “A simple and popular one among the possibilities is the h-index, the Hirsch index (Hirsch 2005), which is an indicator for lifetime achievement of a scholar” (Järvelin & Persson, 2008, p.1433); “I have recently [16] shown that self-citations significantly reduce the h index in contrast to Hirsch’s expectations [1]” (Schreiber, 2008, p.188). Ten note specific applications of the h-index, e.g. “In this paper we tried to provide a partial answer by considering the h-indexes [HIRSCH, 2005A, B] of a group of highly cited researchers based on each of the three citation databases” (Bar-Ilan, 2008, p. 258).

From these instances we can see that mentions of the H-article in the Introduction are at times perfunctory as the major reason is to provide the definition of the h-index and its function.

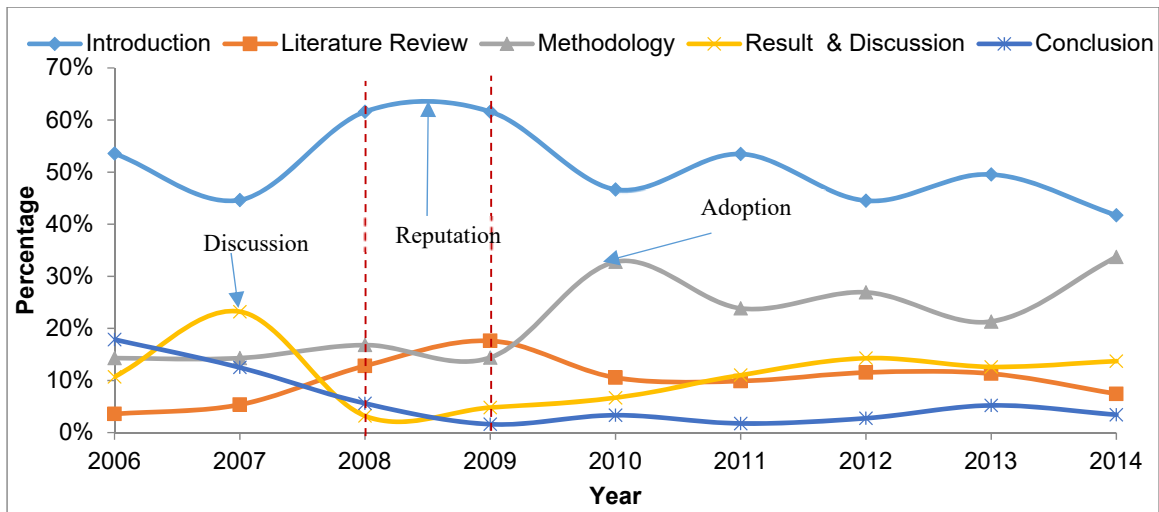


Figure 4. Location distribution of H-article citations 2006-2014.

Citation mentions located in other parts of the article change more frequently than those in the Introduction. For instance, in the R&D, citation mention peaks in 2007. Nine out of 23 citing papers mention the H-article 28 times, which is three times per article on average. Among these mentions, 17 discuss the pros and cons of the h-index, such as Pulina and Ana Helena Dias (2007) addressed: “Hirsch (2005) states that even though C_i properly measures the total impact of a scientist’s activity, it has the following disadvantages: 1) it is hard to find, 2) it may be inflated by few big hits, which may not be representative of the individual if he/she is coauthor with many others on those papers and will correspond to a very atypical value of the a parameter ($a = C_i/h^2$), larger than 5, and 3) it gives undue weight to highly cited review articles versus original research contributions” (p.97). Six mention the H-article to present results, e.g. “The effect of the citing population size was exemplified by HIRSCH (2005) by comparing Physics and Biology, the latter reaching much higher h values” (Imperial & Rodríguez-Navarro, 2007, p.274). After 2007, citation mentions appear less in the R&D.

In the Methodology, citation mention reaches its largest portion in 2010 over the years, when 38 out of 97 citing articles mention the H-article 59 times (1.6 times on average). In these 59 mentions, 18 introduce the h-index by defining it, e.g. “The h index is defined as follows (Hirsch 2005): A scientist has index h if h of his or her N_p papers have at least h citations each and the other ($N_p - h$) papers have h citations each” (Lazaridis, 2010, p.212); 17 describe the function or features of the h-index, e.g. “The h-index has recently got attention and is assumed to be a robust measure for scientific performance and impact (Hirsch 2005)” (Mikki, 2010, p.322); “Due to its simplicity and meaningfulness, Hirsch’s h index (Hirsch 2005) has created quite a stir in the scientific community” (Lazaridis, 2010, p.212). Other mentions include comparing the h-index with its variants (e.g. g-index (Egghe, 2006) and h(2) index (Kosmulski, 2006)). After 2010, citation mention makes up a stable and relative high portion in the Methodology. On the contrary, citation mention shows quite low frequencies in the Literature Review and the Conclusion.

In-depth analysis of citation mentions and the corresponding locations allows us to divide the period (2006-2014) into three phases: “Discussion,” “Reputation,” and “Adoption” (Table 6). In the Discussion phase, many citation mentions of the H-article are distributed in the Introduction and R&D. The H-article is widely and heavily discussed in the R&D (nine of 29 articles and three times per article). Moreover, a range of variants like the g-index (Egghe, 2006) and the h(2) index (Kosmulski, 2006) have been proposed in this phase. In the Reputation phase, citation mentions are largely distributed in the Introduction for reasons related to the fame of the H-article or the definition of the h-index. In the Adoption phase, most of the mentions appear in the Introduction and Methodology sections, where citing articles generally compare the h-index with other indicators or adopt it in their studies.

Table 6. Descriptions of the three phases of citation data collected.

| Phase | Definition | Features | Regarding H-article | Period |
|-------|------------|----------|---------------------|--------|
|-------|------------|----------|---------------------|--------|

| | | | | |
|------------|--|--|---|-----------|
| Discussion | In this phase, many citing articles discuss the function and features of h-index and propose its variants by citing the H-article. | Many citing articles cite and mention the H-article in R&D several times. | Most of the mentions discuss the function and pros/ cons of the h-index. | 2006-2008 |
| Reputation | In this phase, most mentions of the H-article are in the Introductions with only a few in other parts. | Most mentions are perfunctory | Many citing articles cite the definition of the h-index and mention its fame using shorter sentences. | 2008-2009 |
| Adoption | Mentions appear in the Introduction and Methodology part | The h-index is either introduced as a method or compared with other methods. | Articles cite the definition, state the features of the h-index, and compare them with similar indexes. | 2009-2014 |

These three phases indicate that the H-article has been mentioned for different purposes over time, from its optimization to comparison, then to other methods to its application. Figure 5 plots the distinct citation locations of the H-article over the nine-year period of data collection, showing that the diversity of citation location peaks in 2007 and declines thereafter with fluctuations. After combining these data with results in Figure 4, we find that mentions of the H-article are located in various sections (e.g., R&D and Methodology). After 2007, mentions mainly appear in Methodology and Introduction sections only.

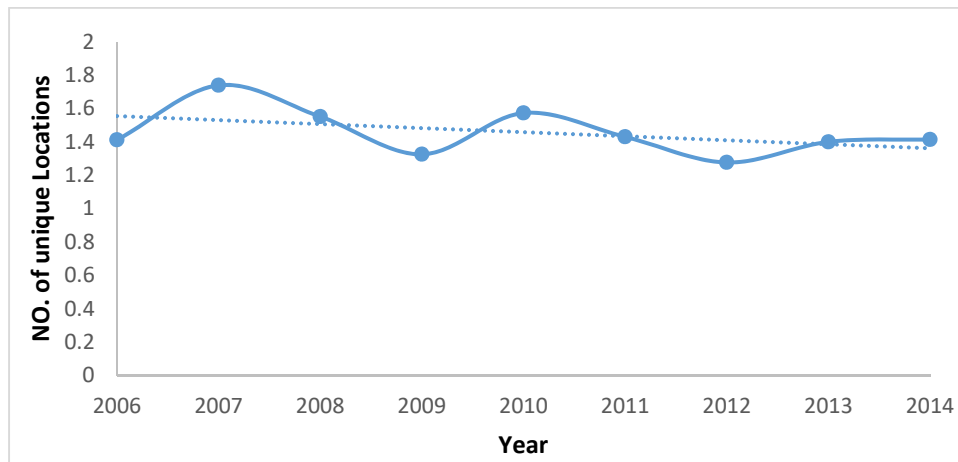


Figure 5. Average Citation Location Diversity of H-article mentions in each year.

Citation Co-mention

Another way to observe the impact change of the article is to analyze how it is co-mentioned with other articles within the same citation contents. Figure 6 illustrates the average citation co-mention and its standard deviation over time. The blue curve indicates a marginal increase in mentions from 1.7 in 2006 to 2.4 in 2014, meaning that more articles are co-mentioned with the H-article during this period.

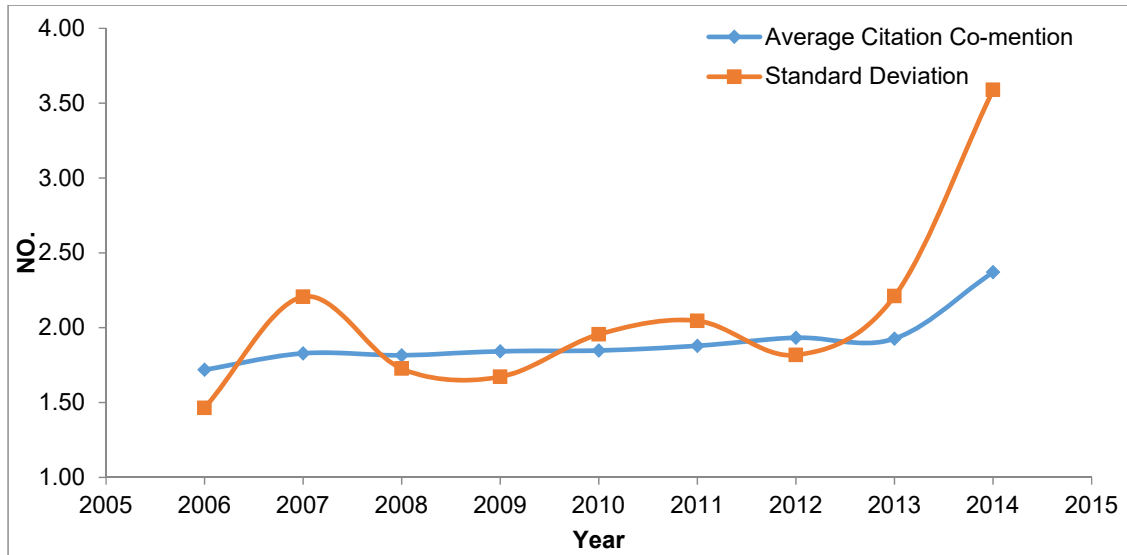


Figure 6. Average number of citation co-mentions in citation sentences and standard deviations 2006-2014.

Citation Topic Similarity

TFIDF value for each word in citation content per year is calculated. The top 30 words with the highest TFIDF values are selected as the topical words in each year (Table 7). Words put in bold indicate that they are either independent from the h-index definition or are newly extracted in that year. The majority of topical words refer to the definition of the h-index, e.g. “measure,” “individual,” and “output.” They are usually found in sentences such as, “The h index for a scientist is the number of papers that the scientist has authored that have received $\geq h$ citations (8)” (Ioannidis, 2010, p.4636). This is reasonable since all the citation sentences mention the h-index to some degree, although the rest of these words also reveal some changes.

In the Discussion phase, most of studies still focus on examining and optimizing the h-index, yet also invent new indicators, using words such as, “advantage,” “parameter,” “braun,” and “ball” (marked in bold). After Egghe proposed the g-index (Egghe, 2006) and Kosmulski proposed the h(2) index (Kosmulski, 2006) in 2006, their articles were frequently co-mentioned with the H-article thereafter. Many publications also discuss the function and potential future of the h-index in 2006 and 2007 (e.g. Ball, 2005; Bornmann & Daniel, 2007; Braun, Glänzel, & Schubert, 2006; Hirsch, 2007; Oppenheim, 2007).

In the Reputation phase, the H-article gains fame in Bibliometrics as well as its fellow studies, since many citing articles mention keywords such as “popular,” and “novel” that attract attention. These words have been frequently co-mentioned with other articles in the Introduction of citing articles, for example, “2007,” “2006,” “egghe,” and “glänzel.”

In the Adoption phase, more and more articles are co-mentioned with the H-article since more words related to years pop up (e.g. “2009” and “2010”). Meanwhile, some other keywords indicate that the H-index has been applied in evaluating the scientific performance of groups or organizations, not just in the evaluation of individual performance, e.g. “group,” “community,” and “field.” The H-article is thus mentioned in the Methodology along with other analytical methods (e.g. social network analysis). As seen in Table 8, the h-index has been combined with other indicators or methods as indicated by citation content.

Table 7. Yearly distribution of top 30 words with high TEIDF values.

| Year | Top 30 words |
|------|--|
| 2006 | output; measure; individual; work; quantify; ranking; physicist; given; simple; performance; high; article; shown; cumulative; use; braun ; result; case; physic; particular; papers; ball ; significant; researcher; order; science; model; assessment; age; arbitrary; |

| | |
|------|--|
| 2007 | individual; researcher; measure; output; value; field; year; science; age; physic; physicist; article; time; average; quantify; ball ; database ; parameter; higher; low; mentioned; productivity; relation; single; metric; identify; advantage ; larger; career; 2006 ; |
| 2008 | researcher; measure; follows; author; output; year; fewer; physicist; individual; 2007 ; 2006 ; bibliometric; article; time; simple; value; quality; popular ; single; equal; original; increase; quantify; definition; count; novel ; ranking; contribution; academic; field; |
| 2009 | 2006 ; individual; output; researcher; 2007 ; quantify; measure; author; article; performance; productivity; time; original; follows; rank; egghe ; year; glänzel ; highest; factor; simple; study; model; quality; braun ; developed; level; way; scholar; raan ; |
| 2010 | individual; measure; researcher; output; author; 2007 ; 2006 ; definition; quantify; article; value; group; 2008 ; factor; science; study; original; time; simple; performance; year; egghe ; metric; popular ; general; quantity; originally; career; single; quotient ; |
| 2011 | researcher; measure; individual; author; article; quality; productivity; output; 2006 ; field ; single; work; factor; academic; bibliometric; time; 2007 ; cumulative; widely; rank; community ; production; egghe ; metric; year; ass; physicist; 2010 ; follows; tool; |
| 2012 | measure; individual; output; author; researcher; 2010 ; productivity; article; time; performance; 2006 ; quantify; scholar; metric; egghe ; 2009 ; bibliometric; year; quality; alonso ; originally; example; 2008 ; evaluating; work; average; evaluate; developed; designed; original; |
| 2013 | measure; researcher; individual; author; productivity; time; output; 2006 ; article; egghe ; academic; year; quantify; quality; single; 2007 ; use; factor; quantity; bibliometric; value; metric; work; contribution; widely ; attention; burrell ; 2009 ; count; significance; |
| 2014 | measure; author; productivity; researcher; individual; article; 2010 ; performance; time; 2006 ; count; account; metric; bibliometric; year; developed; example; factor; work; achievement; quotient; academic; useful; popular ; output; known; information; quality; cumulative; egghe ; |

Table 8. Citation content examples.

| Author/s | Year | Citation Content/ Title |
|-------------------------------|------|---|
| Schuetz, P., & Cafilisch, A., | 2008 | To cover linguistic applications we benchmark the word association network [23] and the graph of the coappearing words in publication titles (co)authored by Martin Karplus [24] who has the third highest h factor [25] among chemists [26]. |
| Cobo, Manolo J., et al. | 2012 | As described in Cobo et al. (2011a), the performance analysis uses bibliometric measures and indicators (based on citations), such as the h-index (Alonso et al., 2009; Hirsch, 2005), g-index (Egghe, 2006), hg-index (Alonso et al., 2010), or q2-index (Cabrerizo et al., 2010) to quantify the importance, impact, and quality of the different elements of the maps (e.g., clusters), and also of the network. |

The Jaccard coefficient is applied to calculate the similarity between the yearly topics in two ways (Figure 7): to compare keywords in each year with the keywords used in 2006 (Similarity A), and to compare keywords between every two continuous years (Similarity B).

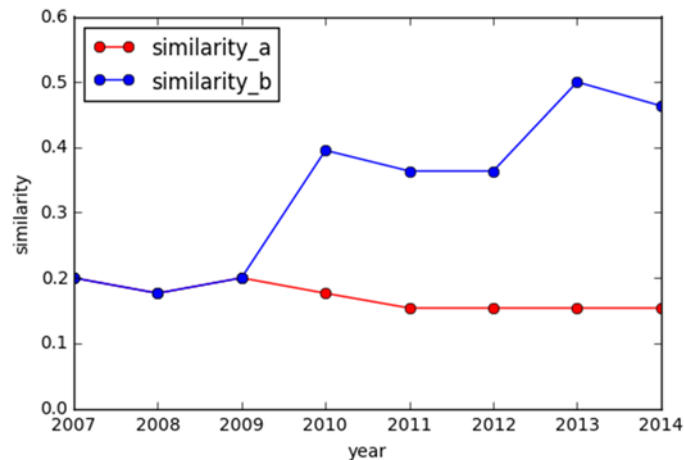


Figure 7. Topic similarity in two different ways based on years.

Similarity A shows a gradual decrease with slight fluctuations after 2009. The decline shows that the citation topics change only slightly over the years. In the first two phases (2007-2008) where the H-article

is usually the sole article mentioned in citation content, the topics change marginally. From 2009 on, when the H-article enters the Adoption phase, the similarity of keywords keeps decreasing, since the newly emerged topics, such as “organizational evaluation” in applied studies (which are usually conducted in differentiated fields) can easily diverge from initial topic’s focus over time. Similarity B after 2009 shows a sharp increase every two or three years (from 2007 to 2010, and from 2010 to 2013 where it peaks), and is usually larger than Similarity A. This indicates that the citation topics between continuous years are somewhat similar (Similarity B), yet vary more broadly in terms of keyword usage, which shows stronger shifts (Similarity A). In the first two phases of Discussion and Reputation (2006-2009), the two types of similarity show similar trends before splitting. These data show evidence of a larger degree of topic drift in the Adoption phase when the H-article is mentioned for specific applications.

5. Conclusion

In our use of citation context and its features of citation mention, citation location, and citation topic to track how the impact of Hirsch’s highly cited article (2005) changes over time, this study contrasts standard citation research that posits the average citing article count as always one, thus discounting the impact of citation mentions over time. The use of average citation mention shows different trends over time, however, as shown in study results collected over the nine-year period from 2006 to 2014. In the first two years, mentions of the H-article increase to peak in 2007 and continue to fall with fluctuation until 2014. The distribution of citation location also indicates different phases, where the citing behavior of the H-article changes from general examination (“Discussion”), its status in the field (“Reputation”), to application by citers (“Adoption”). The average number of the H-article’s co-mentioned articles keeps growing, indicating more and more other cited articles in the citing articles are noted in the same sentences and share contributions with the H-article. The top 30 keywords of citation contents in each year reveal an impact change of the H-article, from mainly citing the definition and function of the h-index to gradually adopting or applying it to other domains. This research therefore demonstrates the dynamic changes of patterns in article citation mentions, and argues that only using citation counts to measure the H-article’s impact changes does not offer a broad measure of its influence over time.

The limitation of this study is that it only selects one highly cited article to highlight impact changes over a period of nine years. More large-scale investigations should be conducted in the future to better understand how and why these impacts change, using other articles of impact. These investigations can help us facilitate the evaluation of articles and their influence in a more nuanced manner, and promote better scholarly communication and understanding of scholarly obsolescence over time. This is an important topic in light of ongoing advances in scientific knowledge and technology, where scholars’ statuses, their funding, and the readers of science are all deeply affected by new knowledge that replaces their own.

Acknowledgments

This work is supported in part by Major Projects of National Social Science Fund of China (No. 16ZAD224). Thanks Star Xing Zhao and Yi Bu for providing suggestions on the manuscript, and to our colleagues in IR&TM Group@NJUST for their technical support and suggestions.

References

- Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, 12(3), 159-170.
- Alsaad, A., & Abbod, M. (2015). Enhanced topic identification algorithm for Arabic Corpora. In Proceedings of the 17th UKSIM-AMSS International Conference on Modelling and Simulation.
- Angrosh, M., Cranefield, S., & Stanger, N. (2012). A citation centric annotation scheme for scientific Articles. Paper presented at the Proceedings of Australasian Language Technology Association Workshop.
- Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, 436, 900.

- Baneyx, A. (2008). "Publish or Perish" as citation metrics used to analyze scientific output in the humanities: International case studies in economics, geography, social sciences, philosophy, and history. *Archivum immunologiae et therapeuticae experimentalis*, 56(6), 363-371.
- Bar-Ilan, Judit (2008). Which h-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics* 74(2): 257-271.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning Research*, 3, 993-1022.
- Bornmann, L., & Daniel, H. D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology*, 58(9), 1381-1385.
- Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69(1), 169-173.
- Brown, P. (1980). The half-life of the chemical literature. *Journal of the American Society for Information Science*, 31(1), 61-63.
- Burton, R. E., & Kebler, R. (1960). The "half-life" of some scientific and technical literatures. *American Documentation*, 11(1), 18-22.
- Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, 40(4), 284-290.
- Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51(7), 635-645.
- Charles, J. P. (1988). Citation Analysis of Astronomical Literature: Comments on Citation Half-lives. *Publications of the Astronomical Society of the Pacific*, 100(623), 106.
- Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, 63(8), 1609-1630.
- 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., & Stirling, K. (2016). Data-Driven Discovery: A new era of exploiting the literature and data. *Journal of Data and Information Science*, 1(4), 1-9.
- 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(9), 1820-1833.
- Dumais, S. T. (2004). Latent semantic analysis. *Annual Review Of Information Science and Technology*, 38(1), 188-230.
- Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131-152.
- Garfield, E. (1964). Science Citation Index: A new dimension in indexing. *Science*, 144(3619), 649-654.
- Hack, T. F., Crooks, D., Plohman, J., & Kepron, E. (2014). Citation analysis of Canadian psycho-oncology and supportive care researchers. *Supportive Care in Cancer*, 22(2), 315-324.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572.
- Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, 104(49), 19193-19198.
- Hu, B., Dong, X., Zhang, C., Bowman, T. D., Ding, Y., Milojević, S., Larivière, V. (2015). A lead-lag analysis of the topic evolution patterns for preprints and publications. *Journal of the Association for Information Science and Technology*, 66(12), 2643-2656.
- 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.
- Ioannidis, J. P. (2010). Is there a glass ceiling for highly cited scientists at the top of research universities?. *The FASEB Journal*, 24(12), 4635-4638.
- Imperial, J., & Rodríguez-Navarro, A. (2007). Usefulness of Hirsch's h-index to evaluate scientific research in Spain. *Scientometrics*, 71(2), 271-282.
- Järvelin, K., & Persson, O. (2008). The DCI index: Discounted cumulated impact-based research evaluation. *Journal of the American Society for Information Science and Technology*, 59(9), 1433-1440.
- Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics*, 8, 197-211.

- Kaplan, D., Tokunaga, T., & Teufel, S. (2016). Citation block determination using textual coherence. *Journal of Information Processing*, 24(3), 540-553.
- Kim, H. J., Jeong, Y. K., & Song, M. (2016). Content- and proximity-based author co-citation analysis using citation sentences. *Journal of Informetrics*, 10(4), 954-966.
- Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI newsletter*, 2(3), 4-6.
- Lazaridis, T. (2010). Ranking university departments using the mean h-index. *Scientometrics*, 82(2), 211-216.
- Lee, Y.-S., Lo, R., Chen, C.-Y., Lin, P.-C., & Wang, J.-C. (2015). News topics categorization using latent Dirichlet allocation and sparse representation classifier. In *Proceedings of the IEEE International Conference on Consumer Electronics*.
- Line, M. B., & Sandison, A. (1974). Progress in documentation: "Obsolescence" and changes in the use of literature with time. *Journal of documentation*, 30(3), 283-350.
- Lipetz, B. A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation*, 16(2), 81-90.
- Liu, X., Zhang, J., & Guo, C. (2013). Fulltext citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, 64(9), 1852-1863.
- MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5), 342.
- McKeown, K., Daume, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., & Fleischmann, K. R. (2016). Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, 67(1), 2684-2696.
- Mikki, S. (2010). Comparing Google Scholar and ISI Web of Science for earth sciences. *Scientometrics*, 82(2), 321-331.
- Minasny, B., Hartemink, A. E., & McBratney, A. (2007). Soil science and the h index. *Scientometrics*, 73(3), 257-264.
- Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5(1), 86-92.
- Oppenheim, C. (2007). Using the h-index to rank influential British researchers in information science and librarianship. *Journal of the American Society for Information Science and Technology*, 58(2), 297-301.
- Pathak, M., & Bharati, K. A. (2014). Botanical survey of India (1971-2010): A scientometric analysis. *Current Science*, 106(7), 964.
- Pulina, G., & Ana Helena Dias, F. (2007). Some bibliometric indexes for members of the Scientific Association of Animal Production (ASPA). *Italian Journal of Animal Science*, 6(1), 83-103.
- Ruane, F., & Tol, R. (2008). Rational (successive) h-indices: An application to economics in the Republic of Ireland. *Scientometrics*, 75(2), 395-405.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.
- Schlachter, G. (1988). Obsolescence, weeding, and bibliographic love canals. *RQ*, 28(1), 7-9.
- Schreiber, M. (2008). The influence of self-citation corrections on Egghe's g index. *Scientometrics*, 76(1), 187-200.
- Schuetz, P., & Caflisch, A. (2008). Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Physical Review E*, 77(4), 046112.
- Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, 11(1), 46-62.
- Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8(3), 327-340.
- Tang, X., Wan, X., & Zhang, X. (2014). Cross-language context-aware citation recommendation in scientific articles. In *Proceedings of the 37th international ACM SIGIR Conference on Research & Development in Information Retrieval*.
- Teufel, S. (2000). Argumentative zoning: Information extraction from scientific text. CiteseerX. Retrieved at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.485>
- Tsay, M.-Y. (1998). Library journal use and citation half-life in medical science. *Journal of the American Society for Information Science*, 49(14), 1283-1292.

- Venable, G. T., Shepherd, B. A., Roberts, M. L., Taylor, D. R., Khan, N. R., & Klimo Jr, P. (2014). An application of Bradford's law: Identification of the core journals of pediatric neurosurgery and a regional comparison of citation density. *Child's Nervous System*, 30(10), 1717-1727.
- 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.
- 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.
- Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, 64(7), 1490-1503.
- Zhao, D., & Strotmann, A. (2015). Dimensions and uncertainties of author citation rankings: Lessons learned from frequencyweighted in-text citation counting. *Journal of the Association for Information Science and Technology*, 67(3), 671-682.