



Where are citations located in the body of scientific articles? A study of the distributions of citation locations



Zhigang Hu^{a,b,c,*}, Chaomei Chen^{a,b,d}, Zeyuan Liu^{a,b,c}

^a Joint-Institute for the Study of Knowledge Visualization and Science Discovery, Dalian University of Technology, China

^b Joint-Institute for the Study of Knowledge Visualization and Science Discovery, Drexel University, USA

^c WISE Lab, Dalian University of Technology, Dalian, China

^d College of Information Science and Technology, Drexel University, Philadelphia, USA

ARTICLE INFO

Article history:

Received 5 June 2013

Received in revised form 7 August 2013

Accepted 14 August 2013

Keywords:

Distribution of citations

Citation location

Structure of scientific articles

Full text analysis

ABSTRACT

We address issues concerning what one may learn from how citation instances are distributed in scientific articles. We visualize and analyze patterns of citation distributions in the full text of 350 articles published in the *Journal of Informetrics*. In particular, we visualize and analyze the distributions of citations in articles that are organized in a commonly seen four-section structure, namely, introduction, method, results, and conclusions (IMRC). We examine the locations of citations to the groundbreaking h-index paper by Hirsch in 2005 and how patterns associated with citation locations evolve over time. The results show that citations are highly concentrated in the first section of an article. The density of citations in the first section is about three times higher than that in subsequent sections. The distributions of citations to highly cited papers are even more uneven.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

In scientific literature, citations are instances of references to previously published scientific articles. Due to various constraints in terms of accessibility to full texts of scientific articles, traditional citation analysis is largely limited to the analysis of references listed at the end of an article (Cano, 1989). Recently, as full texts of scientific articles have become increasingly accessible, analyzing citation patterns at the level of full text has attracted a considerable amount of attention (Ding, Liu, Guo, & Cronin, 2013; Liu & Chen, 2013; Small, 2011).

Many have studied the citation context of a citation instance in terms of functions of citations (Cronin, 1982; Moravcsik & Murugesan, 1975; Teufel, Siddharthan, & Tidhar, 2006), motivations of citations (Bonzi & Snyder, 1991; Brooks, 1985, 1986) and classifications of citations (Agarwal & Yu, 2009; Aljaber, Martinez, Stokes, & Bailey, 2011; Bertram, 1970; Cano, 1989; Di Marco & Mercer, 2004; Finney, 1979; Garzone & Mercer, 2000; Hanney, Grant, & Jones, 2005; McCain & Turner, 1989; Pham & Hoffmann, 2003; Vinkler, 1987). Among these studies, citation location has been identified as a simple yet valid factor in citation classification scheme (Bertram, 1970; Cano, 1989; Finney, 1979; Maričić, Spaventi, Pavičić, & Pifat-Mrzljak, 1998; McCain & Turner, 1989). However, locations, or distributions of citations, have not been systematically studied using scientometrics methods. The purpose of this paper is to bridge this gap and address the following questions:

- How are citations distributed in the full text of scientific articles?
- How are they distributed in each section of a scientific article?

* Corresponding author at: WISE Lab, Dalian University of Technology, Dalian, China. Tel.: +86 13591791279.
E-mail addresses: huzhigang.cn@gmail.com (Z. Hu), cc345@drexel.edu (C. Chen), liuzy@vip.163.com (Z. Liu).

Another interesting issue we will address is how citations to high-impact papers are distributed inside the body of citing articles. For example, where are the most common locations where high-impact references are cited? A citation classic is expected to be cited in the front of an article because of its prominent position in a relevant field. We hypothesize that citations to highly cited papers would be concentrated in the earlier sections of articles.

The aims of this paper are fulfilled by a study of two carefully selected samples. One is the distribution of citations in articles with a 4-section structure, namely Introduction-Methods-Results-Conclusions (or IMRC). The IMRC structure is common in many scientific fields, including the scientometrics field. This kind of articles is chosen as a representative case. We visualize the locations of citations in IMRC-structured articles. In the second analysis, we examine the paper that is cited the most in the *Journal of Informetrics* – the h-index paper by Hirsch, and investigate where it is cited inside the body of citing articles. We found different citation patterns. In other words, the popularity of references may influence the distributions of how they are cited.

2. Data collection and processing

In this paper, we collected 350 articles published in *Journal of Informetrics* from its first issue in 2007 to the first issue in 2013. The full texts of these scientific articles were retrieved from ConSyn (<http://consyn.elsevier.com>), a content syndication system operated by Elsevier®. ConSyn is a full-text database of all the articles in journals published by Elsevier. Unlike another full-text database ScienceDirect®, also from Elsevier, ConSyn provides downloadable articles in XML format instead of PDF. Compared with the latter, full texts in XML are more machine-friendly and easier to process. It allows us to identify the specific citation locations precisely and reliably. Therefore, it becomes feasible to conduct comprehensive case studies of a large amount of full text of scientific articles.

Identifying citations from XML-formatted full texts is straightforward. In XML full text, citation instances are clearly marked with tags, for example, in the form of `<ce:cross-ref refid="bib0015">Egghe, 2009a</ce:cross-ref>`, where *bib0015* represents the reference id in the reference list at the end of a citing paper. *Egghe, 2009a* represents the citation anchor in the body of the citing paper. Fig. 1 illustrates where and how citations and references are extracted. Inside the body of a scientific article, each citation instance is detected and inserted into a database with attributes such as Citation ID, Citation Location, Citing Paper ID, and Cited Reference ID. At the same time, information of a reference is extracted and imported into a reference table, with information such as its Title, Author, Source, and Citation ID. The tables of Citations and References are connected with each other based on Citation IDs.

In the Citation table, the location of a citation is measured in terms of the number of words from the beginning of the citing article. For example, if one citation's location is 398, it means the citation is located 398 words away from the beginning of the article. The relative position of a citation is measured in terms of the length of the article, i.e. the total number of words.

Body of the full text of article

The real career h-index sequence of a particular author is determined from all his publications and the total number of citations received up to a given year (Egghe, 2009a). If a scientist publishes a fixed number of publications every year, the h-index is expected to increase in a linear way (Hirsch, 2005), as has been confirmed by some empirical studies (Egghe, 2009b; Ronald & Bihui, 2008). This linear growth could actually also be obtained without a constant yearly publication, as Burrell showed (Burrell, 2007a). In reality, however, some h-index sequences grow in a non-linear way. Egghe (2009a) showed different cases of linear, concave or convex h-index sequences and conducted a mathematical study of them. These

Corresponding Text in XML Format

```
<ce:para id="par0010" view="all">The real career h-index sequence of a particular author is determined from all his publications and the total number of citations received up to a given year (
<ce:cross-ref refid="bib0015">Egghe, 2009a</ce:cross-ref>). If a scientist publishes a fixed number of publications every year, the h-index is expected to increase in a linear way (
<ce:cross-ref refid="bib0035">Hirsch, 2005</ce:cross-ref>), as has been confirmed by some empirical studies (
<ce:cross-ref refid="bib0020">Egghe, 2009b; Ronald & Bihui, 2008</ce:cross-ref>). This linear growth could actually also be obtained without a constant yearly publication, as Burrell showed (
<ce:cross-ref refid="bib0005">Burrell, 2007a</ce:cross-ref>). In reality, however, some h-index sequences grow in a non-linear way.
<ce:cross-ref refid="bib0015">Egghe (2009a)</ce:cross-ref> showed different cases of linear, concave or convex h-index sequences and conducted a mathematical study of them. These deviations from the expected linear behaviour are interesting, as they contain additional information regarding the career of the corresponding author (
<ce:cross-ref refid="bib0010">Burrell, 2007b</ce:cross-ref>).
</ce:para>
```

Extract Citations information

Citing_Paper	Cit_id	Ref_id	Cit_Label	Cit_Position
10.1016/j.joi.2011.02.003	10	bib0015	Egghe, 2009a	398
10.1016/j.joi.2011.02.003	11	bib0035	Hirsch, 2005	403
10.1016/j.joi.2011.02.003	12	bib0020	Egghe, 2009b	421
10.1016/j.joi.2011.02.003	12	bib0050	Ronald & Bihui, 2008	423
10.1016/j.joi.2011.02.003	13	bib0005	Burrell, 2007a	442
10.1016/j.joi.2011.02.003	14	bib0015	Egghe (2009a)	456

Citation Table

Tail of the full text of article

References

Burrell, Q. L. (2007a). Hirsch's h-index: A stochastic model. *Journal of Informetrics*, 1, 16–25.
 Burrell, Q. L. (2007b). Hirsch index or Hirsch rate? Some thoughts arising from Lian's data. *Scientometrics*, 73, 19–28.
 Egghe, L. (2009a). Comparative study of h-index sequences. *Scientometrics*, 81, 311–320.
 Egghe, L. (2009b). Mathematical study of h-index sequences. *Information Processing and Management*, 45, 288–297.
 Egghe, L. (2010). The Hirsch-index and related impact measures. *Annual Review of Information Science and Technology*, 44, 65–114.
 Guns, R., & Rousseau, K. (2009). Simulating growth of the h-index. *Journal of the American Society for Information Science and Technology*, 60, 410–417.
 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, 16569–16572.

Corresponding Text in XML Format

```
<ce:bib-reference id="bib0015">
<ce:label>Egghe, 2009a</ce:label>
<ce:reference url="http://www.elsevier.com/locate/0000-0000">
<ce:author>
<ce:given-name>L.</ce:given-name>
<ce:surname>Egghe</ce:surname>
</ce:author>
<ce:author>
<ce:given-name>L.</ce:given-name>
<ce:surname>Egghe</ce:surname>
</ce:author>
<ce:title>Comparative study of h-index sequences</ce:title>
</ce:contribution>
<ce:host>
<ce:series>
<ce:volume>81</ce:volume>
<ce:issue>3</ce:issue>
<ce:pages>311–320</ce:pages>
</ce:host>
</ce:bib-reference>
```

Extract reference information

Ref_id	Author	Ref_title	Ref_journal	Volume	Year	Page
bib0015	L. Egghe	Comparative study of h-index sequences	Scientometrics	81	2009	311–320

Reference Table

Fig. 1. Extracting citation and reference information from the full text of a scientific article. On the left, six citations were extracted along with their locations. On the right, information of the reference (Egghe, L, 2009a) was extracted and imported into the reference table.

Download English Version:

<https://daneshyari.com/en/article/10358568>

Download Persian Version:

<https://daneshyari.com/article/10358568>

[Daneshyari.com](https://daneshyari.com)