The Recurrence of Citations within a Scientific Article

Zhigang Hu[1], Chaomei Chen[2] and Zeyuan Liu[1]

[1] *huzhigang@dlut.edu.cn, liuzy@dlut.edu.cn*
WISE Lab, Dalian University of Technology, 116024 Dalian (China)

[2] *cc345@drexel.edu*
College of Computing and Informatics, Drexel University, 19104 Philadelphia (USA)

Abstract

Although listed at the tail of a scientific article only once, a reference is usually cited repeatedly inside the full text of the article. In this research, we investigated the universality of recurring citations in Journal of Informetrics. About 1/4 references are repeatedly cited. For these repeatedly cited references, their citation location and citation context for the first and subsequent times are examined separately. Normally, recurring citations of a same reference tend to be located in the same section instead of different ones. It proves that, even if a reference is cited for multiple times in a single citing paper, it is still focus on the same topic in the same section most of the time. We also explored the reason why recurring citations are happening. By comparing the contexts of two kinds of citations, the first-time citations and the succeeding citations, we found that, for a specific reference, its first-time citation is usually not as intentional as the succeeding citations. Just because of the relative importance of the succeeding citations compared to the first-time citation, recurring citations are reasonable and necessary.

Conference Topic

Citation and co-citation analysis

Introduction

Citations are essential components for scientific articles. Traditional citation analysis is more like reference analysis, since only references listed at the tail of the article are researchers' concern. Citations, which indicate the locations and context where references are cited, are almost ignored in previous research. The reference analysis is much easier and effective most of the time, but in the meanwhile, some important information might be neglected. For example, where are these references are cited inside the citing papers? How are the citations distributed among different sections? By investigating the citation location and the citation context, however, we can understand not only the pattern how references are cited, but also the reason why authors cite it like that.

Nowadays, full-text citation analysis, which is about how references are cited in the body of citing papers, is just beginning (Ding, Liu, Guo, & Cronin, 2013; Hu, Chen, & Liu, 2013; Liu, Zhang, & Guo, 2013; Zhang, Ding, & Milojević, 2013). During to the increasingly availability of structured full texts such as XML-formatted articles, researchers began to turned their attention from references to citations in the body of articles. For example, Ding et al. have examined the distribution of references across text and find that most highly cited works appear in the Introduction and Literature Review sections of citing papers (Ding et al., 2013). Hu et al. visualize the location distribution of citation instances, especially those to highly-cited references. The results show that citations are usually distributed very uneven inside the full texts of scientific articles (Hu et al., 2013).

In full-text citation analysis, recurring-citation is an interesting issue. Recurring-citation refers to the phenomenon that a reference is cited more than once in a citing paper. Take this paper for example, we cite the reference of (Hu et al., 2013) in the first sentence of last paragraph for the first time, and then cite it again in the last sentence of the same paragraph for the second time. In this paper, we call this reference a repeatedly cited reference, or a reference with recurring citations. Recurring-citation is a common fact in citation behaviour. In our previous research, we find that, sometimes, a reference might be repeatedly cited as many as nine times in a single paper (Hu et al., 2013).

In this research, we will investigate the phenomenon of recurring citations. Our concern is the universality and the pattern of recurring-citations, including: (1) how common recurring citations are in scientific articles? (2) where the recurring citations of a single reference are usually located inside the paper? (3) what the difference is between its first-time citation and the succeeding ones? In the end, the reason why recurring-citation is necessary will also be discussed.

Data and Methods

To detect recurring citations of a reference inside a citing paper, the full text of the citing paper need to be observed. There are two types of full texts: one is in unstructured format such as PDFs, which is human-friendly; the other is in structured format such as XMLs/HTMLs, which is machine-friendly. Compared with PDFs, structured full texts, e.g., XMLs, are much easier to process by computer. For example, XML-formatted full texts can be parsed directly using an existing function: *xml_parse()* in PHP. Thus, it is very straightforward to identify citations inside a citing paper. Nowadays, structured XML-formatted full texts are available and downloadable in almost every bibliographic database, such as Elsevier, Springer, John & Wiley, and especially, the open access online journals like PLoS ONE. In this study, the data of full texts was sourced from Elsevier ConSyn (http://consyn.elsevier.com), a content syndication system developed by Elsevier. Since 2011, Elsevier ConSyn provided downloadable articles in XML format.

In Elsevier ConSyn, we retrieved and downloaded all the full texts of 350 articles published in Journal of Informetrics (JOI) from 2007 to 2013. Journal of Informetrics is chosen as the case in this study because it is published by Elsevier and belongs to the field of library and information science. By our own developed program, we parsed these XML-format full texts and extract all the citations inside them. Since each citation instances is clearly marked with a XML tag, i.e. <*ce:cross-ref*>, they can be recognized and extracted easily. All the attributions of each citation, including its location and its citees, were recorded and import into database tables.

By looking into citations' citees, we achieved the cited times of each reference inside each citing paper. If the cited times is equal to one, it means the reference is one-time cited inside this citing paper. While if cited more than once, the reference is considered as repeatedly cited or recurrently cited. In this research, we will count the frequency of each type of reference, e.g., once-cited, twice-cited, triple-cited, etc. In this way, the universality and intensity of recurring-citation can be estimated accordingly.

For repeatedly cited references, their citation locations will be studied. The location of citation can be measured by, from macro to micro scales: character, word, sentence, paragraph and section. In this study, we chose the measurement at the largest scale: section. We will calculate the count of citations in each section and see how citations are distributed in different sections. Generally, a scientific article is made up of four sections, namely *Introduction*, *Data and Methods*, *Results,* and *Discussion and Conclusions*. It is called IMRaD structure usually (see e.g. Agarwal & Yu, 2009; Swales, 1990) To some extent, citation location can reveal the citation motivation. If we are aware of the section where a citation is located, the role of the citation can be figured out to some extent. For instance, if a reference is cited in the section of *Data and Methods*, usually section II, it is probably a helpful citation relevant in the aspect of methodology; while if it located in the section III or the section of *Results*, the citation is more likely about comparable results.

Besides the location distribution of recurring citations, we also examined the difference between a reference's first-time citation and the succeeding ones. We extracted the context when a reference is cited for the first time and when it is cited again in the following parts. The first-time and the succeeding citation contexts will be compared in terms of the count of their citees inside. The more citees/references a citation contains, the less important each citee/reference is. The citation with many citees/references, such as the one in the first sentence of the second paragraph, is called perfunctory citation (Cano, 1989; Oppenheim & Renn, 2004; Pham & Hoffmann, 2003; Voos & Dagaev, 1976), which means authors decide not to cite the citees/references seriously in an excluded way. In this research, we are interesting in which one, the first-time citation or the succeeding citation, is more likely to be perfunctory citation for a multiple-cited reference.

Results and Discussion

The universality of recurring citations

Firstly, we examined how common recurring citations are in the Journal of Informetrics. Among all the 11,327 references inside the 350 articles, 8,417 (74.3%) of them were cited once in a single citing article. The other 2,910 references (account for 25.7%) were cited twice or more, including 1,726 (15.2%) twice-cited references, 613 (5.4%) triple-cited references, and 571 (5.0%) references cited for four times or more. Although one-time citation is the main citation pattern undoubtedly, the phenomenon of recurring-citation cannot be ignored in both frequency and intensity.

Figure 1 shows the frequency distribution of references of each kind, companied with a distribution graph in double logarithmic coordinates. As it shown in the best fitting line, the frequency distribution of multiple citations follows a power law ($y= 21557\ x^{-3.479}$, $R^2=0.9679$), which is a very common law in the field of bibliometrics, such as the distribution of scientific productivity (Lotka, 1926) or keywords (Zipf, 1949). Obviously, it is not accidental that the frequency distribution of recurring citations is in this pattern.

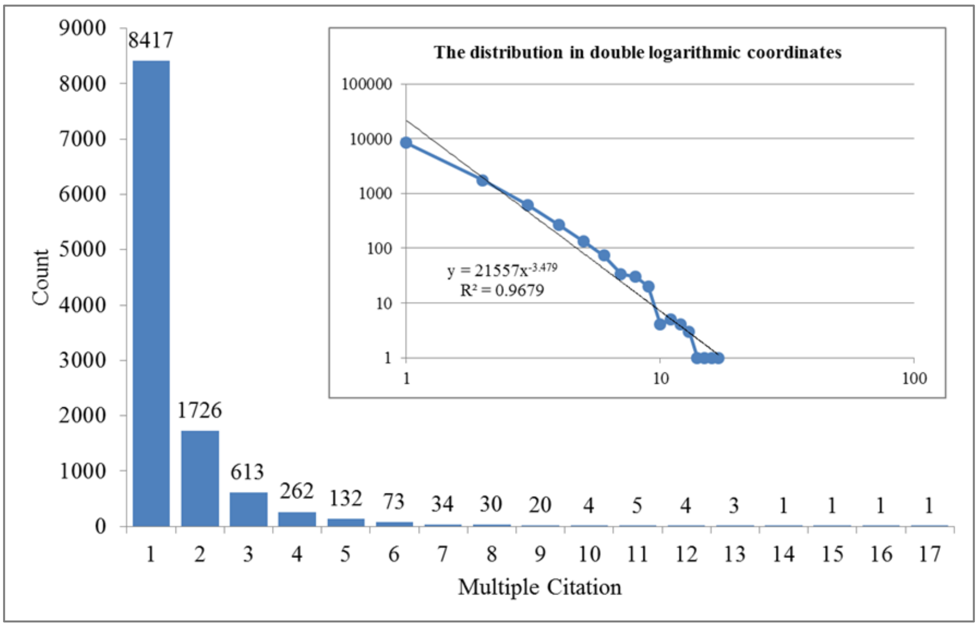

Figure 1. The count of references by their multi-citation times.

The locations of recurring citations

The location pattern of recurring citations is the focus of this research. In this part, we will investigate the location distribution of multi-citation by section. In Journal of Informetrics, 92 articles (26.3% of all) adopt IMRaD structure, which is most used form to organize articles in our research. Thus, we selected all these 92 four-section articles in IMRaD structure as cases, and explored how citations are distributed in the four different sections.

As shown in Figure 2, among all the 3035 citations in these 92 articles, 1238 (40.8%) citations are located in Section I, or the section of *Introduction*; 760 (25.0%) of them are located in Section II (or *Methods*); 769 (25.3%) citations is in the sections of *Results*; and 268 (8.8%) in *Discussion and Conclusions*. This mode of section distribution of citations meets our expectation on citation locations, since it is the widely accepted fact that authors are likely to cite most in the section of *Introduction*.

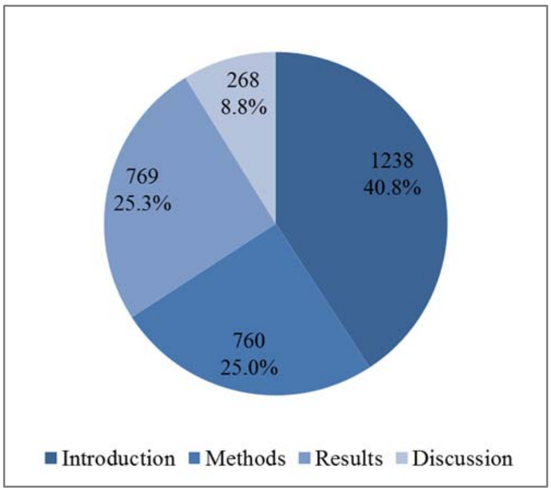

Figure 2. The count of citations in each section.

Based on the section distribution of citations, we are then able to investigate the section combination distribution of a reference's recurring citations. For each repeatedly cited reference, its recurring citations could be located in any sections, either the same section or the different sections. Since twice cited references are the simplest and most common (59.3%) types of repeatedly-cited references, they were chosen for calculating combined-section distribution.

For each twice-cited reference, we recorded the located sections of both citations. The counts of the 10 types of section combinations are shown in Table 1. Among all the 796 twice-cited references, the most common ones are those cited in Section I for both the first and the second time. 224 (28.2% of all) references belong to this type. 124 (15.6%) references are cited in Section I for the first time and Section II for the second time. References that are cited in section IV twice are least common (18 or 0.8%). Totally, 444 (55.8%) references are cited in the same section twice, while 350 (44.2%) ones are cited in the difference sections.

Table 1. The combined section distribution of the twice citations of references

Located Section of the second citation Located section of the first citation	*Sec I*	*Sec II*	*Sec III*	*Sec IV*
Sec I	**224** 28.2%			
Sec II	**124** 15.6%	**90** 11.3%		
Sec III	**68** 8.6%	**42** 5.3%	**112** 14.1%	
Sec IV	**64** 8.1%	**24** 3.0%	**28** 3.5%	**18** 2.3%

Although more twice cited references are cited in the same section, we cannot say that a reference's multiple citations tend to be located in the same sections except that the expected proportion of the multiple citations located in the same section is calculated and compared. Thus, we assume that a reference's twice citations are located independently and randomly, just like two arbitrary citations in the article. Under this hypothesis, the expected distribution of section combinations of twice citations can be calculated as follow:

$$
\begin{aligned}
&(\mathit{Sec\ I}, \mathit{Sec\ I}) : (\mathit{Sec\ I}, \mathit{Sec\ II}) : (\mathit{Sec\ I}, \mathit{Sec\ III}) : (\mathit{Sec\ I}, \mathit{Sec\ IV})\\
&: (\mathit{Sec\ II}, \mathit{Sec\ II}) : (\mathit{Sec\ II}, \mathit{Sec\ III}) : (\mathit{Sec\ II}, \mathit{Sec\ IV})\\
&: (\mathit{Sec\ III}, \mathit{Sec\ III}) : (\mathit{Sec\ III}, \mathit{Sec\ IV})\\
&: (\mathit{Sec\ IV}, \mathit{Sec\ IV})\\
&= 40.8\%\times 40.8\% : 40.8\%\times 25.0\%\times 2 : 40.8\%\times 25.3\%\times 2 : 40.8\%\times 8.8\%\times 2\\
&\quad : 25.0\%\times 25.0\% : 25.0\%\times 25.3\%\times 2 : 25.0\%\times 8.8\%\times 2\\
&\quad : 25.3\%\times 25.3\% : 25.3\%\times 8.8\%\times 2\\
&\quad : 8.8\%\times 8.8\%
\end{aligned}
$$

=16.6% : 20.4%: 20.7% : 7.2% : 6.3% : 12.7% : 4.4% : 6.4% : 4.5% : 0.8%

Figure 4 shows the expected and observed proportions of the section combinations of each kind. If the expected values match the observed well, it means the twice citation are located independently and randomly indeed; otherwise, it means that there is a certain tendency in how to cite a reference twice. In Figure 4, we have not seen the match between the expected and observed values. For example, based on our initial hypothesis, the proportion of (Sec I, Sec I) should be 16.6%, not even closed to 28.2% as observed; the proportion of (Sec I, Sec III) should be 20.7%, while the observed value is 8.6%, which is much lower. Neither of them presents the match as assumed.

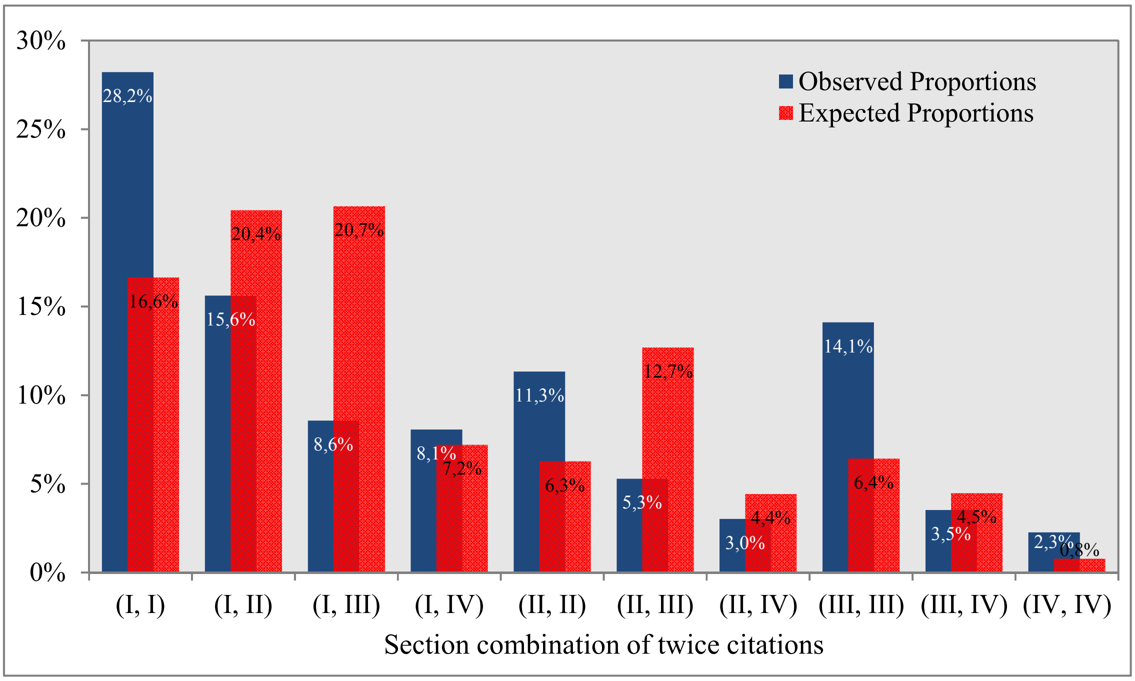

Figure 3. The distribution of the expected and observed proportions of located section combinations of twice citations.

According to the comparison between the expected and observed values, these 10 combinations can be divided into two classes: the above-expectation combinations and the below-expectation ones. The combinations of (Sec I, Sec I), (Sec II, Sec II), (Sec III, Sec III), (Sec IV, Sec IV) and (Sec I, Sec IV) belong to the former. Their observed proportions are higher than expected significantly. The other 5 combinations, i.e., (Sec I, Sec II), (Sec I, Sec III), (Sec II, Sec III), (Sec II, Sec IV) and (Sec III, Sec IV), belongs to the latter. From this division, we can see that, the references with twice citations located in the same section are preferable to those with twice citations located in different sections. The only exceptions are the references cited inside (Sec I, Sec IV), which have an above-expectation proportion (2.3% v.s. 0.8%), though its twice citations located in the different sections.

Why do authors tend to cite a reference multiple times inside the same section? The explanation could be simple. Normally, a reference is only helpful for a single topic, usually existing in a concentrated part of an article, such as a section. Few references are necessary for several different topics, or in different sections. That is why references are preferred to be cited in a single section. This explanation also interprets why the combination of (Sec I, Sec IV) is an exception. The first and the fourth section, although farthest away with each other, are actually discussing about the same topic at the same level, i.e., the hindsight and foresight of research questions.

The context of recurring citations

We have revealed how common recurring citations are and where these recurring citations are usually located, and now we will examine their contexts. Firstly, the citation contexts of repeatedly cited references for the first and the succeeding times were extracted separately. There are totally 11,448 first-time citation contexts and 5,469 succeeding ones extracted. We will explore the difference between these two groups of citation contexts in terms of citation intensity, which can be estimated by how many citees they contained.

The count of citees contained in each citation context is calculated one by one. Averagely, a citation contains 1.94 citees, or put it another way, authors cite 1.94 references once at a time. As it shown in Figure 6, although most citations (64.2% of all) cite only one single citee/reference, there are still more than 1/3 of citations contain two or more citees/reference. 1457 (8.7%) citations cite even five or more references once.

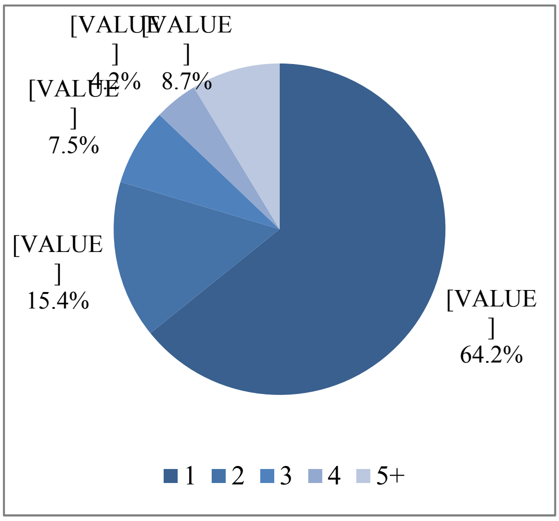

Figure 4. The distribution of citations by the count of contained citees.

Separately, the counts of citees contained by the first-time and succeeding citations are investigated. The first-time citations contain 2.13 citees on the average, while the succeeding citations contain 1.94 citees. Figure 5 shows the specific distribute of both of them by their count of contained citees. For the first-time citations, totally 38.5% of citations cited two citees or more; while for succeeding citations, only 30.1% did. It means the first-time citations are more likely to be perfunctory citations than the succeeding citations. In other words, authors normally cite a reference more casually and perfunctorily for the first time; and then cite it again in the following paragraphs more formally and solemnly. In other words, usually, authors just mention a reference in the beginning, and then seriously use it when citing it later again.

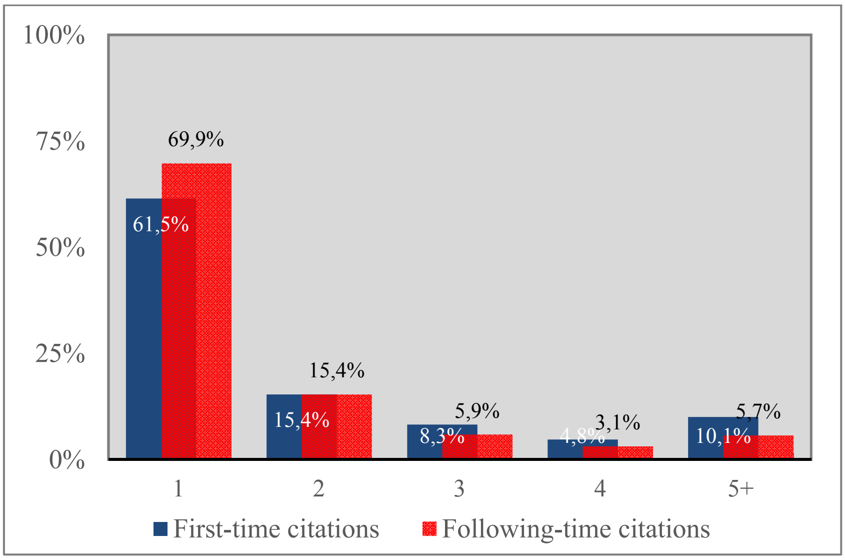

Figure 5. The distribution of the first-time and succeeding citations by their count of contained citees.

Conclusions

Recurring citations are common in scientific publications. In Journal of Informetrics, about 1/4 references are repeatedly cited in citing papers. Although not the mainstream of citation pattern, recurring-citation is undoubtedly a phenomenon that cannot be ignored in full-text citation analysis, an increasing hot research field in recent year.

In this study, we investigate the recurring-citation phenomenon in two perspectives: the citation location and the citation context. In citation location analysis, we find that a reference's recurring citations tend to be located in the same section or closely with each other. It shows that a reference is only cited in a single topic normally. When the topic switches, the reference has little chance to be cited again.

The context of recurring citations contexts are also examined in terms of their citation intensity. As it shown in the result, for a repeatedly cited reference, its first-time citation is usually kind of perfunctory. The reference is always cited accompanied with other references together. When it is cited another time in the following part of the citing paper, the citations are more exclusively and solemnly. Precisely because the succeeding citations are usually more importantly, recurring citations are reasonable and necessary inside scientific articles.

Acknowledgments

The research was supported by the Natural Science Foundation of China (NSFC) under Grant 71103022, 61301227 and 71473028.

References

Agarwal, S., & Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics*, *25*(23), 3174–80. doi:10.1093/bioinformatics/btp548

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, *40*(4), 284–290.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, *7*(3), 583–592. doi:10.1016/j.joi.2013.03.003

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*(4), 887–896.

Liu, X., Zhang, J., & Guo, C. (2013). Full‐text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, *64*(7), 1852–1863.

Lotka, A. J. (1926). The frequency distribution of scientific production. *Journal of the Washington Academy of Science*, *16*, 317–323.

Oppenheim, C., & Renn, S. P. (2004). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, *51*(5), 225–231.

Pham, S. B., & Hoffmann, A. (2003). A new approach for scientific citation classification using cue phrases. *AI 2003 Advances in Artificial Intelligence*, 759–771.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. *Journal of Advanced Composition* (Vol. 11, p. 272). Cambridge: Cambridge University Press.

Voos, H., & Dagaev, K. S. (1976). Are all citations equal? Or did we op. cit. your idem? *Journal of Academic Librarianship*, (1), 20–21.

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*, *64*(7), 1490–1503.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley Press.