The Effects of Co-citation Proximity on Co-citation Analysis

Shengbo Liu[1] and Chaomei Chen[2]

[1] *liushengbo1121@gmail.com*
Dalian University of Technology, WISElab, China

[2] *chaomei.chen@drexel.edu*
College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104-2875, USA

Abstract

In this paper we investigate the effects of co-citation proximity on the quality of co-citation analysis through four experiments of co-citation instances found in full-text scientific publications. First, we compared the distributions of co-citation instances at four levels of proximity in journal articles with the traditionally used article-level co-citation counts. Second, we analyzed how co-citation instances at different proximity levels are distributed across organizational sections in articles. Third, the distribution of co-citation proximity over different co-citation frequency groups is investigated. Fourth, we identified the occurrences of co-citations at different proximity levels with reference to the corresponding traditional co-citation network. The results show that sentence-level co-citations not only preserve the essential structure of the corresponding traditional co-citation network but also form a much smaller subset of the entire co-citation instances typically considered by traditional co-citation analysis. Implications for improving our understanding of underlying factors concerning co-citations and developing more efficient co-citation analysis methods are discussed.

Introduction

In traditional co-citation analysis, two references are considered as co-cited by an article regardless the proximity of the positions of corresponding citations within the article. A major pragmatic reason is due to the lack of access to full-text versions of articles. More recently, repositories such as Pubmed Central make it possible to analyze full-text articles algorithmically. The general question is whether the proximity of co-cited references is expected to produce any insights that traditional article-level co-citation analysis cannot offer.

Studies that make use of such repositories began to emerge. For instance, Elkiss el al. (2008) found that papers co-cited at a finer granularity (within the same sections, paragraphs, or sentences are more similar to each other than papers co-cited at the article level. Gipp and Beel (2009) and Callahan et al (2010) have shown that contextual analysis could augment the validity of co-citation analysis.

This paper presents four experiments that are designed to reveal the effects of co-citation proximity on the quality of co-citation analysis. First of all, the distribution of co-citation proximity in different journals is studied. Co-citations in a paper are considered at four levels of proximity: the sentence level, the paragraph level, the section level and the article level. Higher-level co-citations do not include co-citations found at lower levels. Second, the distribution of co-citations at different proximity levels across sections is analyzed. Third, the distribution discipline of co-citation proximity in different levels under different co-citation frequencies circumstances are analyzed, the relationship between co-citation proximity and co-citation frequency is investigated. Finally, the differences between networks based on different co-citation proximity and traditional co-citation network are compared.

The co-citation proximity analysis requires not only bibliographic information, but also the full text of an article. In this research, we utilize the PubMed Central database. In particular, references and full text information from BMC Bioinformatics, BMC Systems Biology, and BMC Biology are extracted and analyzed.

Future work is discussed, including incorporating the notion of co-citation proximity in author co-citation analysis and journal co-citation analysis, and the application of co-citation contextual analysis in traditional co-citation analysis.

Related Work

Co-citation Analysis

Co-citation analysis was proposed by Small (1973) and Marshakova (1973) independently. Co-citation analysis quantifies the relationship between two co-cited papers, with the assumption that more frequently co-cited documents indicate a stronger relationship. Document co-citation analysis was extended to author co-citation analysis by White and Griffith (1981). Author co-citation analysis is used to find the research similarity of the co-cited authors. Both document co-citation analysis and author co-citation analysis are applied to many scientific areas. This research focuses on document co-citation analysis.

Co-citation contextual analysis

Citation context can be defined as the sentences that contain the citation of a particular reference. Contextual information can be used to reveal the nature of a citation. It can be used to generate a summary of an article (Qazvinian, 2008). Nanba and Okumura (1999, 2005) collected citation context information from multiple documents cited by the same article and generated a summary of the article based on such citation contextual information. They extracted citing sentences from citation context and generated a review. Mei (2008) and Mohammad (2009) found that the summarization is very different from the abstract of the article. Nakov et al. (2004) introduced the term citances. A citance is defined as a set of sentences that surround a particular citation. For example, the sentence "This comparison is made using BLASTX [18]" is a citance of the citation to [18]. The citances can be used in abstract summarization and other Natural Language Processing (NLP) tasks such as corpora comparison, entity recognition, and relation extraction. Bradshaw (2002) used citation contextual information in scientific literature retrieval, and augmented the retrieval efficiency.
Although many studies focused on citation contextual, few studies have addressed co-citation context. Small (1973) proposed the co-citation analysis method, but did not make use information in citing sentences. In 1979, he did the co-citation analysis based on the co-citation contextual and analyzed the content in which the co-citation paper mentioned (Small, 1979).
Recently, researchers start to consider the position of co-citation in co-citation analysis, and have made some insightful observations. Elkiss el al. (2008) studied co-citations in an article at four levels: the sentence level, the paragraph level, the section level, and the paper level. They found that papers co-cited at a finer granularity are more similar to each other than papers co-cited at a coarser granularity. For example, papers co-cited at the sentence level have a stronger relationship than papers co-cited at the section level. Gipp and Beel (2009) focused their research on co-citation similarity based on co-citation position. In their research, co-citations could occur in five categories: within the same sentence, the same paragraph, the same chapter, the same journal and the same journal but different edition. In each category, a co-citation is given a different value of 1, 1/2, 1/4, 1/8 or 1/16. The result shows that the weighted co-citation analysis has much better similarity than traditional co-citation analysis. Callahan et al. (2010) used a similar method to calculate the co-citation strength; a co-citation can occur at different levels of a paper. A co-citation at the paper level is assigned a weight of 1, and for each level deeper an additional weight of 1 is added. However, the weighing scheme in their approach is rather subjective and the sample size they considered was too

small to draw more general conclusions. In this study, our goal is to address some of the remaining issues.

Method

Four sub-studies are described as follows.

Co-citation proximity

Co-citations in a citing paper are considered at four levels of proximity, namely, the article level, the section level, the paragraph level and the sentence level (See Fig.1). If two references are cited within the same sentence, the co-citation instance is called a sentence-level co-citation. If two references are cited in different sentences but within the same paragraph, it is called a paragraph-level co-citation. Similarly, two references cited in different paragraphs but within the same section define a section-level co-citation. Finally, if two references are cited in different sections but within the same paper, we have an article-level co-citation. We expect that sentence-level co-citations represent the strongest bonds between references, whereas paragraph-, section-, and article-level co-citations represent weaker and weaker bonds, respectively.

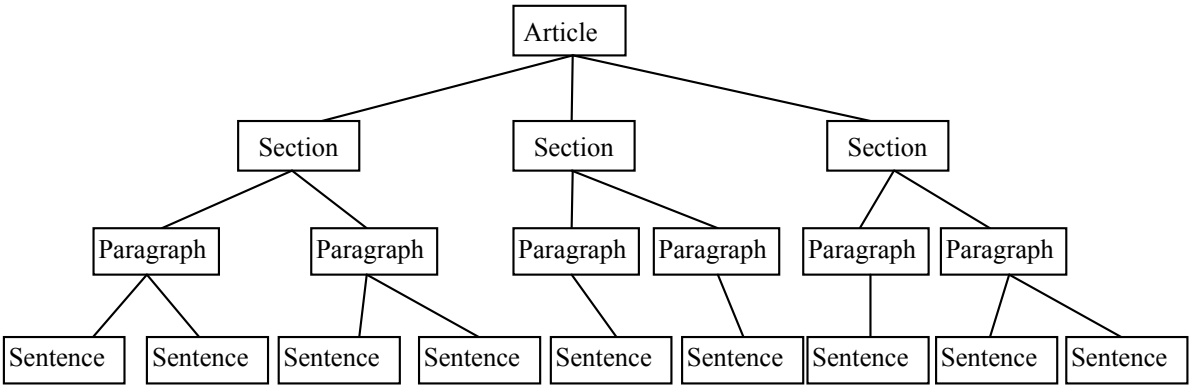


Figure 1. A four-level co-citation proximity scheme.

Distribution of co-citation proximity

Co-citations across different proximity levels are characterized by the distribution of co-citation proximity. Articles from three journals are used in this experiment. These journals are selected from the PubMed Central (PMC). PMC provides the full text of articles in XML, which makes it a valuable source of citation proximity information. The three journals are BMC Bioinformatics, BMC Systems, and BMC Biology. The numbers of articles in the three journals are 3720, 413 and 419, respectively, from periods of 2001-2010, 2007-2010 and 2003-2010, respectively.

References may be co-cited at different levels within one paper, but in this study we measure the strengths of co-citations in terms of the occurrences of the nearest proximity. For example, one reference is cited twice in a paper, and another reference co-cited with it in sentence level and paragraph level, then their co-citation is set to 1 at the sentence level. In future work, we will address co-citations across all levels of proximity.

Co-citation proximity in different sections

The distribution of co-citation proximity across different sections is computed. Sections are identified based on the XML mark-ups. Typical section headings include *introduction*,

background, *method, datasets, result*, *implementation, discussion*, and *conclusion*. However, there are exceptions. For examples, some sections are labelled as *construction and content, testing, evaluation, experiment* and *application*. These sections are lumped in a catch-all category called *others* in our study. We expect more co-citations in *introduction*, *background*, *method*, and *discussion* sections than sections such as *result* and *conclusion* sections in which authors are expected to focus on reporting details about their own work.

The relationship between co-citation frequency and co-citation position

The distribution of co-citation proximity by co-citation frequency is studied. We expect that highly co-cited references should be considerably co-cited within near proximity due to some underlying connections between the co-cited references. We choose *BMC Bioinformatics* for this experiment because it has more papers (3720 papers) and a longer period time (2001-2010) than the other two journals. There are three steps in this analysis. First, the distribution of co-citation frequency is computed. Second, the distribution of co-citation proximity by co-citation frequency is computed and presented.

In this experiment, we constructed a total of 22 data sets, including 20 subsets for papers co-cited 1 to 20 times, one for papers co-cited 21~30 times, and one for papers co-cited 30 or more times. These subsets were further divided into four groups, namely 1~6, 7~12, 13~18, and over 18 co-citations. Finally, the h-index (Hirsch, 2005) is used to identify high and low co-citation references. Although there are many methods to identify the high and low co-citation references, such as mean or median, h-index is relatively well-known in the field of scientometrics and can easily divide a ranked list into two parts (Chen, 2007). The h-index is originally designed to measure the productivity and impact of the published work of a scientist or a group of scientists. The index is based on the set of the scientist's most cited papers and the number of citations that they have received. The h-index is used as an index to measure the high co-cited data sets in this study. The number of co-citation pairs that are co-cited at least *h* times is taken here as the co-citation h-index for the entire data set. The data set is divided into two groups. Group 1 contains co-citation pairs that have less than *h* times of co-citation and group 2 contains that that have greater than or equal to *h* co-citations. We expect that highly co-cited references are more likely to have sentence-level proximity.

Network overlay of co-citation proximity

In addition to social network analysis and visualization, many researches focus on co-citation networks. Through the analysis of co-citation networks, the evolution of the subject structure can be revealed, and hotspots in research frontiers can be detected (Chen, 2006). Software systems such as Pajek, Ucinet, and CiteSpace have been used in co-citation network analysis.

This experiment will identify the differences between network structures corresponding to different co-citation proximity levels based on articles published in the *BMC Bioinformatics* journal. Citespace (Chen, 2006) is used to visualize these co-citation networks. First, a traditional co-citation network is visualized as a base network. Then, a finer-grained co-citation network at a particular proximity level is superimposed on the traditional network. The traditional co-citation network is generated with a threshold of 4 or more co-citations. Finer-grained proximity-level networks use a threshold of 3 or more co-citations. Because of a narrower scope, the lower threshold at a finer granularity remains to be a sub-network of the overall base network. Although the article-level co-citations should be consistent with co-citations in the base network, we expect co-citations at lower levels of proximity would highlight the most important topics in the traditional network.

Results

Results are presented in the same order of the corresponding methods introduced in the earlier section.

Distributions of Co-citation proximity

The distributions of the co-citation proximity in three journals are shown in Table1.

Table 1 : The distribution of the co-citation proximity in three journals

	Proximity				
Journals	*Sentence*	*Paragraph*	*Section*	*Article*	*Total*
BMC Bioinformatics	94,755	163,527	407,235	1,694,083	2,359,600
BMC Sys Biology	15,280	42,000	86,619	442,258	586,157
BMC_Biology	16,668	40,135	102,205	458,707	617,715

As shown in Figure 2, co-citations at various proximity levels have very similar distributions for these three different journals. 2~4% of co-citations were made within the same sentences. 6~7% were within the same paragraphs. About 15~17% of co-citations appeared at the section levels. Over 70% of co-citations occurred at the article level. This suggests that traditional co-citation analysis would be biased towards co-citations that are loosely coupled at the article level and the tighter co-citations at sentence and paragraph levels are likely to be overshadowed by loosely connected references. Although the results are based on three journals, the pattern seems to be consistent enough to conjecture that this may be the case for a broader range of journals. The next question is to what extent tightly and loosely coupled references differ in terms of the patterns they form.

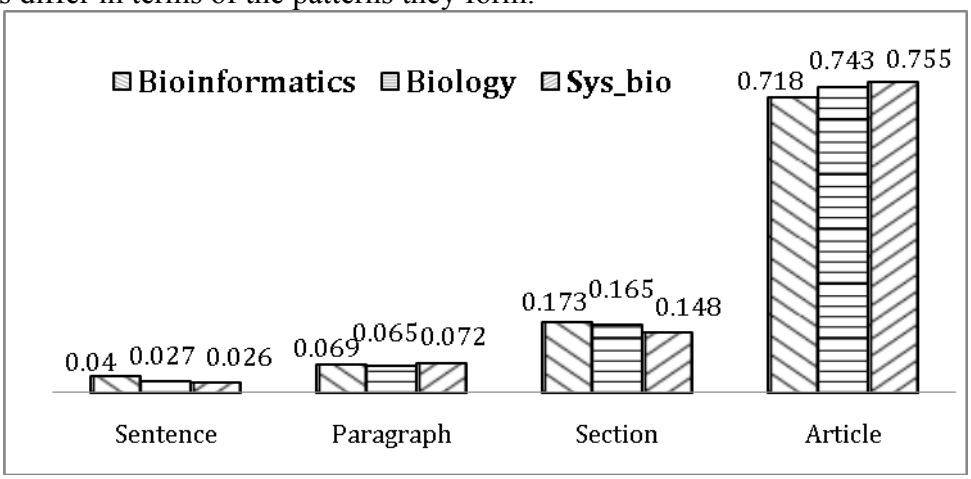


Figure 2. The distribution of co-citation position in three journals

Distributions of co-citations across organizational sections

Table 2 shows the distributions of different proximity level co-citations across organizational sections in articles. Most of the co-citations appear in the background section, and the least in introduction section. Figure 3 shows the percentage of co-citations at each proximity level in different sections. The percentage of co-cited references at the sentence level is lower than co-cited references at the paragraph level and section level in background section, but higher than them in method, dataset, result, implement, discussion and conclusion sections.

Table 2. The co-citation distribution in sections of *BMC Bioinformatics* articles.

	Organizational Section								
Proximity	*Introduction*	*Background*	*Method*	*Dataset*	*Result*	*Implement*	*Discussion*	*Conclusion*	*Others*
Sentence level	905	52,983	4,554	3,464	2,529	1,312	3,279	2,416	23,313
Paragraph level	1,322	100,593	6,416	5,493	3,799	1,751	5,572	3,004	35,577
Section level	4,339	288,028	12,213	9,268	7,637	3,058	11,813	3,727	67,152

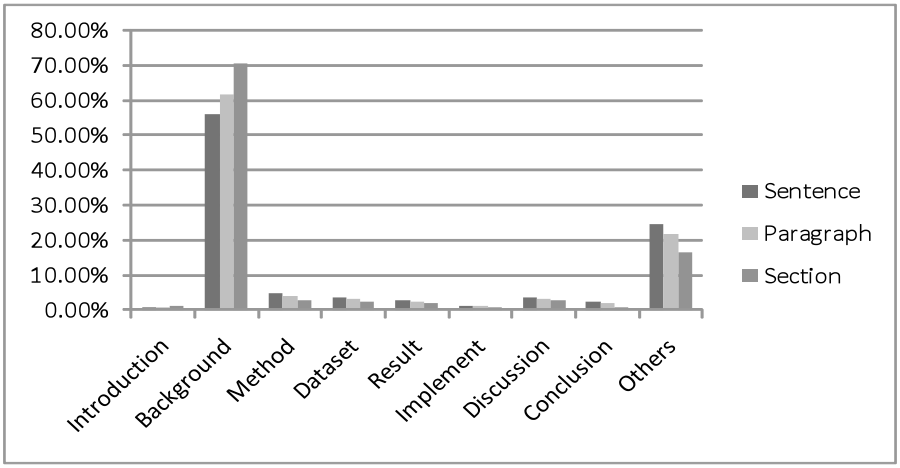


Figure 3. Distribution of co-citation position in different sections.

Co-citation Frequency and Proximity

Table 3 shows the relationship between co-citation frequency and proximity. As the co-citation frequency goes up by the number of co-citied papers appears to drop down.

Table 3.Relationship between co-citation frequency and co-citation proximity.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21-30**	**>30**
Sentence	77465	9029	3223	1643	893	513	384	291	229	136	99	59	72	96	52	74	68	28	37	40	198	126
Paragraph	144520	11429	3368	1533	787	528	330	240	149	114	78	51	53	55	38	24	25	16	23	21	85	60
Section	376664	19970	5190	2176	1116	666	427	232	159	153	61	85	34	39	36	29	21	19	18	17	82	41
Article	1597573	63882	15492	7268	3179	1971	1176	757	588	437	257	189	179	146	99	129	107	45	93	62	280	174
Total	2196222	104010	27273	12620	5975	3678	2317	1520	1125	840	495	384	338	336	225	256	221	108	171	140	645	401

Table 3 shows that the number of co-citations at proximity levels varies considerably across the range of co-citation frequency. We use the proportion of co-citations at various proximity levels in 22 data sets to represent the general trends (See Figure 4).

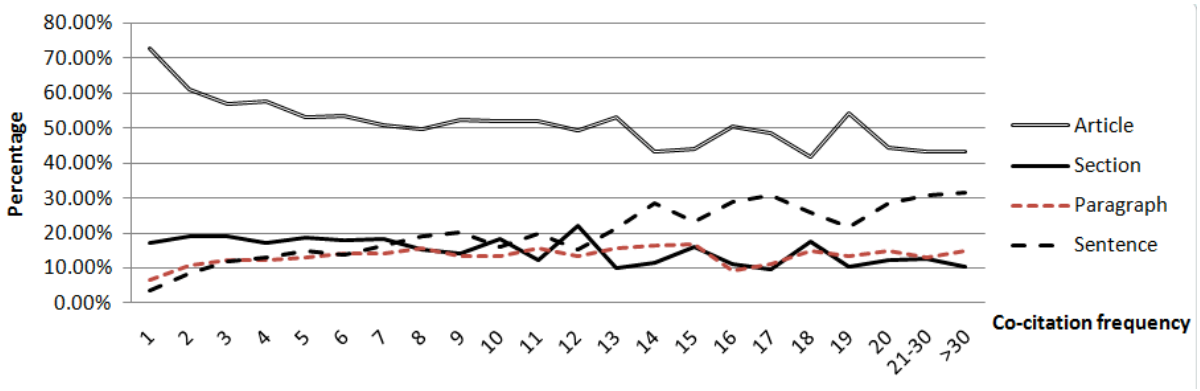


Figure 4. Proportion of co-citations at the four co-citation proximity levels.

In Figure 4 the horizontal axis represents co-citation frequency. The vertical axis represents the proportions of co-citations at various proximity levels. For references co-cited once only, most of them (73%) were co-cited at the article level, section-level co-citations were the second most popular one (17%), followed by paragraph- and sentence-level co-citations for 6.5% and 3.5% respectively.

One prominent trend is that the share of sentence-level co-citations increases along with co-citation frequency at the expense of the share of article-level co-citations. In contrast, paragraph- and section-level citations essentially remain the same across all frequencies of co-citations. The proportion of co-citations at the sentence level became the second largest for co-citation frequency greater than 13. When the co-citation frequency reached 30 times or more, sentence-level co-citation accounts for more than 30% of all co-citations.

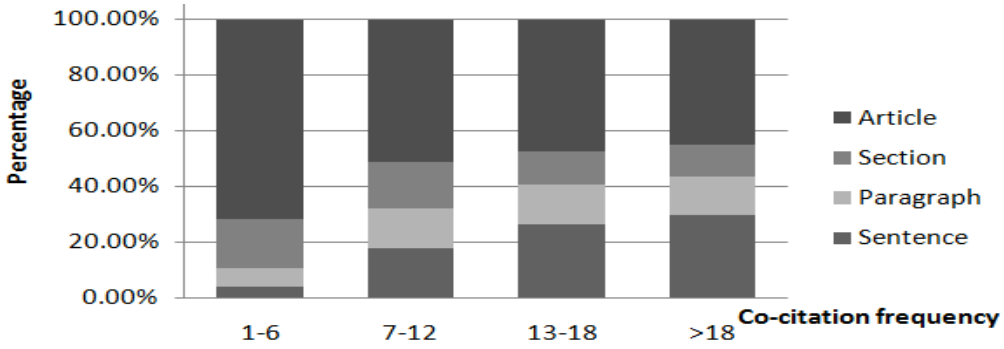


Figure 5. Distribution of 4 co-citation proximity groups over 4 co-citation frequency groups.

Figure 5 provides an alternative depiction of the distribution of co-citations at various proximity levels over co-citation frequencies. It is evident that for 13-18 and >18 groups the size of the sentence blocks is much larger than that in 1-6 and 7-12 groups.

These observations suggest that traditional co-citation analysis using a lower co-citation threshold is more likely to be biased than co-citation analysis using a higher co-citation threshold because sentence-level co-citations become more prominent in high co-citation groups and reduce the prominence of loosely coupled article-level only co-citations.

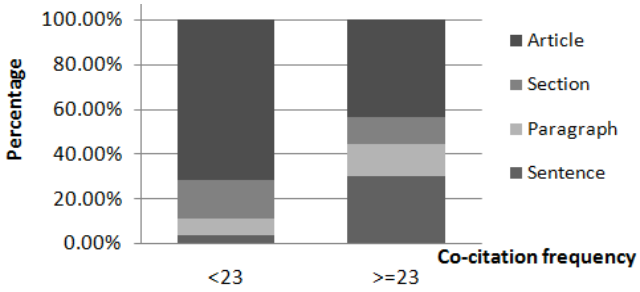
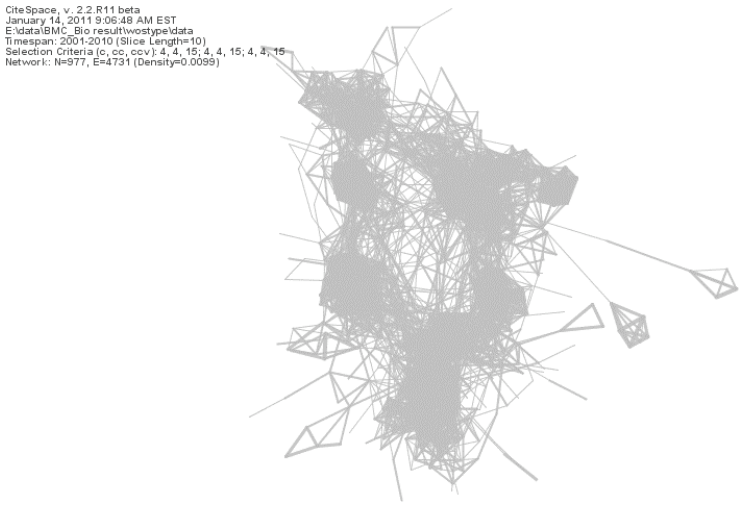


Fig.6 High and low co-citation split by co-citation h-index (h=23) and corresponding proximity.

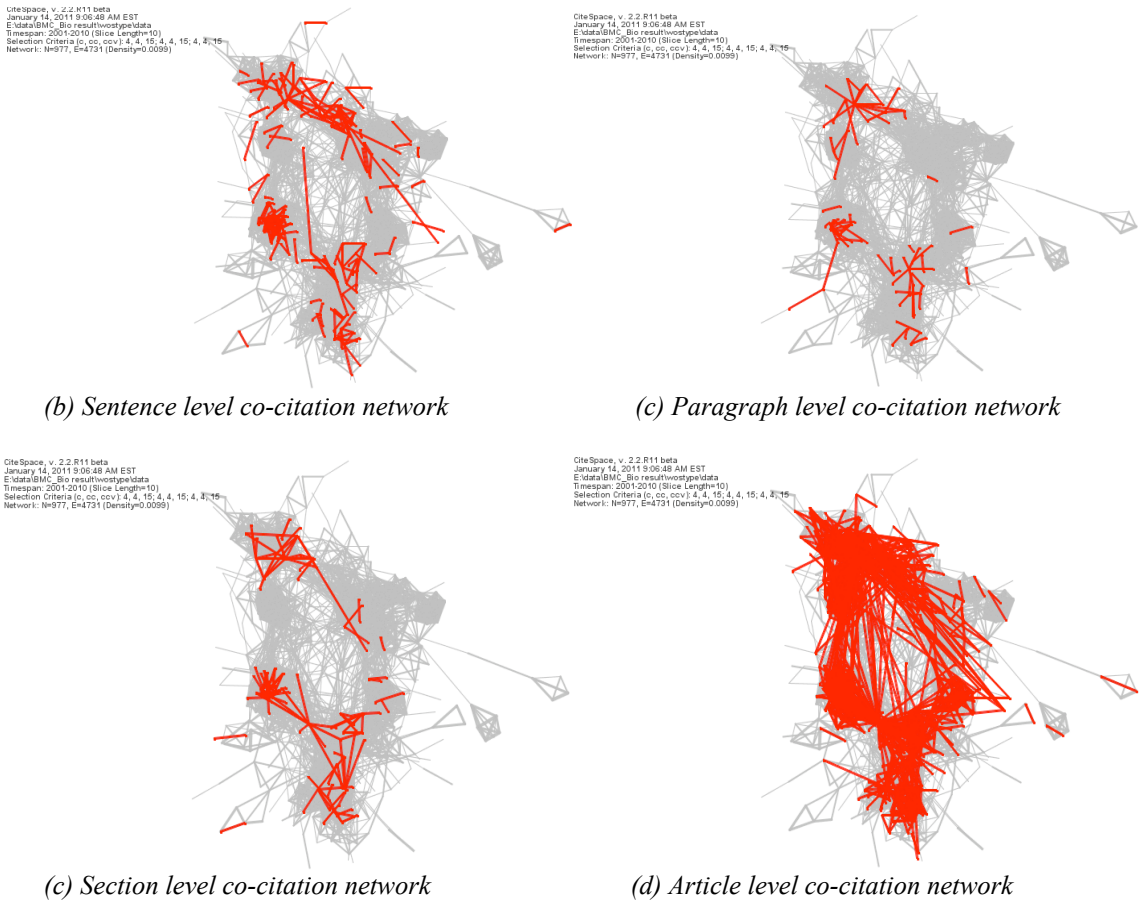
Figure 6 shows the relationship between high and low co-citation frequencies at co-citation proximity levels. High and low co-citations are defined by a co-citation h-index of 23. The high co-citation group has 13 datasets and the low co-citation group has 22 datasets. If mean or median method is used to divide the datasets, the high co-citation group will contains 20 or 18 datasets and the low co-citation group will contains 15 or 17 datasets. The results are similar. In the high co-citation group, article- and sentence-level co-citations are prominent. In contrast, in low co-citation group, article-level only co-citations are overwhelming.

Co-citation proximity in context

A traditional co-citation network and overlays of co-citation networks at four levels of proximity are shown in Figure 7. The proximity-level network overlays are superimposed over the traditional co-citation network in darker colours. The traditional co-citation network contains 977 references and 4,731 co-citation links (Figure 7a). The sentence-level network has 267 edges (Figure 7b). The paragraph-level network has 83 edges (Figure 7c). The section-level network has 126 edges. The article-level network has 1825 edges (Figure 7d).
The article level network has much more information than other three networks, and covers 38.58% of the edges in the traditional co-citation network. This is consistent with the high proportion of co-citations found at this level. Networks associated with the other three proximity levels form sub-networks of the traditional co-citation network. On the other hand, proximity level networks seem to cover the areas of the highest density in the original traditional co-citation network. Although sentence-level co-citations represent about 4% of co-citation instances at all levels, they represent 5.64% of the edges in the traditional co-citation network. In contrast, paragraph and section level co-citations represent 1.75% and 2.66%, respectively. Most of the sentence level co-citations are essential to the traditional co-citation network.



(a)Traditional co-citation network



(b) Sentence level co-citation network

(c) Paragraph level co-citation network

(c) Section level co-citation network

(d) Article level co-citation network

Figure 7. Traditional co-citation network and proximity-level co-citation network overlays.

Discussion

First, we have found that the distribution of co-citation at each proximity level is very similar across three different journals. However the relationship between the distribution and the journal impact factor is not obvious, further study with more journals is needed. If this distribution can be found in other journals and other fields, a new method for co-citation analysis can be developed to take the effect of co-citation proximity into account and improve the quality and accuracy of co-citation analysis.

The distribution of co-citation at proximity levels across organizational structures of articles is not entirely what we expected. Sentence-level co-citations are more popular in method, result, discussion, and conclusion sections than paragraph- and section-level co-citations.

The analysis of the relationship between the frequency and proximity of co-citations has revealed the increasing role of sentence-level co-citations in high-frequency co-citation groups. Traditional co-citation analysis, largely due to the lack of access to adequate data, does not distinguish co-citations made with different granularity.

Our study improves the understanding of the roles played by high and low frequency co-citations in the overall co-citation network. On the one hand, we have shown that the traditional co-citation analysis tends to be overwhelmed by many loosely coupled references that their co-citations can only be found at the highest level of proximity, the article level. On the other hand, our results also indicate that traditional co-citation analysis represents a superset of the essential structure that would be characterized by finer-grained proximity-level co-citations. The biases towards loosely coupled co-citations tend to be reduced and even diminished as the threshold of co-citations traditionally used to sample co-citation instances raises.

The proximity of co-citation appears to have implications on improving the quality and sensitivity of co-citation analysis. For example, the results of our study suggest that sentence-level co-citations are potentially more efficient in identifying the essential structure of the underlying literature than co-citations loosely coupled at the article level because 1) sentence-level co-citations constitute only a fraction of the entire co-citation pool; one may expect a 20-time reduction in terms of the size of dataset, and 2), more importantly, sentence-level co-citations appear to retain the most important structural components in the traditional co-citation network and therefore the fidelity of the traditional co-citation analysis can be expected to be adequately preserved. Furthermore, the four-level proximity framework provides a flexible methodology such that one may decide to take one or more proximity levels into account so as to expand the breadth and depth of the coverage.

Our study has identified several potential routes for future research. For example, the role played by sentence-level co-citations suggests that text analysis of citing sentences would be an important direction to pursue. In this paper we have focused on issues concerning co-citation proximity in document co-citation analysis. Similar studies are needed to investigate patterns in author co-citation analysis and journal co-citation analysis.

Conclusion

We have studied the distributions of co-citations at four levels of proximity and found that sentence-level and article-level only co-citations play a predominant role in forming the overall co-citation network. In conclusion, our results indicate that sentence-level co-citations are potentially more efficient candidates for co-citation analysis because they tend to preserve the essential structural components of the corresponding traditional co-citation network and they tend to appear much infrequent in comparison to loosely coupled article-level only co-citations. These findings are important to improve our understanding of some of the fundamental factors that may influence the outcome of co-citation analysis.

Acknowledgments

Shengbo Liu is currently a visiting doctoral student at Drexel University. This research is supported by National Natural Science Foundation of China (grant number 71003011) and Fundamental Research Funds for the Central Universities of China (Grant 852010). Thanks to the reviewers for the useful suggestions.

References

Aaron Elkiss, Siwei Shen, Anthony Fader, Gunes¸ Erkan,David States & Dragomir R. Radev(2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1), 51–62.

Alison Callahan, Stephen Hockema & Gunther Eysenbach.(2010). Contextual Cocitation: Augmenting Cocitation Analysis and its Applications. *Journal of the American Society for Information Science and Technology*, 61(6), 1130-1143.

Bradshaw.B.(2002). Reference directed indexing: Indexing scientific literature in the context of its use. Northwestern University

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377.

Chen, C., Zhang J., Zhu W. & Vogeley M. (2007) . Delineating the citation impact of scientific discoveries. Proceedings *of the 7th ACM/IEEE-CS joint conference on Digital libraries*(pp.19-28). Vancouver,ACM

Gipp, B. & Beel, J. (2009). Citation Proximity Analysis (CPA)—A new approach for identifying related work based on co-citation analysis. *Proceedings of the 12th International Conference on*

Scientometrics and Informetrics (pp. 571–575). Leuvuen: International Society for Scientometrics and Informetrics.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS*,102 (46), 16569–16572.

Marshakova, I. V. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya*, 2(6), 3–8.

Mei Qiaozhu & Zhai ChengXiang. (2008). Generating impact-based summaries for scientific literature. *Proceedings of ACL '08*(pp.816–824). Columbus:ACL

Nanba, H., & Okumura, M. (1999). Towards multi-paper summarization using reference information. *The 16th International Joint Conference on Artificial Intelligence*,(pp. 926-931). Stockholm:IJCAI

Nanba, H., & Okumura, M. (2005).Automatic detection of survey articles. *The Research and Advanced Technology for Digital Libraries* (pp. 391–491). Berlin: ECDL.

Nakov, P.I., Schwartz, A.S., & Hearst, M.A. (2004). In Citances: Citation sentences for semantic analysis of bioscience text. *SIGIR 2004Workshop on Search and Discovery in Bioinformatics*, Sheffield:SIGIR

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev& David Zajic.(2009). Using citations to generate surveys of scientific paradigms. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 584–592). Boulder: Association for Computational Linguistics.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265–269.

Small, H. (1979). Co-citation context analysis: The relationship between bibliometric structure and knowledge.*Proceedings of the ASIS Annual Meeting.* (pp. 270–275). Medford: Information Today.

Vahed Qazvinian & Dragomir R. Radev. (2008). Scientific paper summarization using citation summary networks. *Proceedings of the 22nd International Conference on Computational Linguistics*, (pp. 689-696). Stroudsburg: Association for Computational Linguistics.

White, H.D.& Griffith, B.C. (1981). Author cocitation: Aliterature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3), 163–171.