Scientometrics (2012) 91:495–511
DOI 10.1007/s11192-011-0575-7

The proximity of co-citation

Shengbo Liu · Chaomei Chen

Received: 29 November 2011 / Published online: 20 December 2011

Abstract Traditional co-citation analysis has not taken the proximity of co-cited references into account. As long as two references are cited by the same article, they are retreated equally regardless the distance between where citations appear in the article. Little is known about what additional insights into citation and co-citation behaviours one might gain from studying distributions of co-citation in terms of such proximity. How are citations distributed in an article? What insights does the proximity of co-citation provide? In this article, the proximity of a pair of co-cited reference is defined as the nearest instance of the co-citation relation in text. We investigate the proximity of co-citation in full text of scientific publications at four levels, namely, the sentence level, the paragraph level, the section level, and the article level. We conducted four studies of co-citation patterns in the full text of articles published in 22 open access journals from BioMed Central. First, we compared the distributions of co-citation instances at four proximity levels in journal articles to the traditional article-level co-citation counts. Second, we studied the distributions of co-citations of various proximities across organizational sections in articles. Third, the distribution of co-citation proximity in 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 (1) the majority of co-citations are loosely coupled at the article level, (2) a higher proportion of sentence-level co-citations is found in high co-citation frequencies than in low co-citation frequencies, (3) tightly coupled 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

S. Liu (✉)
WISE Lab, Dalian University of Technology, Dalian, China
e-mail: liushengbo1121@gmail.com

C. Chen
College of Information Science and Technology, Drexel University, 3141 Chestnut Street,
Philadelphia, PA 19104-2875, USA
e-mail: chaomei.chen@drexel.edu

underlying factors concerning co-citations and developing more efficient co-citation analysis methods are discussed.

Keywords Co-citation proximity · Co-citation analysis · Citation contextual · PubMed Central

Introduction

Traditional co-citation analysis does not take into account the proximity of references co-cited by an article. Some references are cited within the same sentence, whereas other references may be cited in further-apart positions in an article. Intuitively, we expect references cited within the same sentence have tighter connections than references cited in different sections of an article. How are references distributed in terms of their positions in text? Does the proximity of citations reflect any more profound connections at various organisational levels of a scholarly publication?

A major pragmatic reason for the almost absence of studies of the nature of co-citation proximity is due to the lack of access to full text versions of articles. More recently, repositories such as PubMed Central (PMC) 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 et 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.

We present four experiments in order 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 PMC database. In particular, references and full text information from 22 BioMed Central (BMC) journals 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 studies the relationship between two co-cited papers, with the assumption that more frequently co-cited documents indicate a stronger relationship.

Co-citation analysis was proposed by Small (1973) and Marshakova (1973) independently. Co-citation analysis has been applied to analyze the intellectual structure of many scientific areas (Chen, 2004, 2006; Chen et al. 2008; Small and Greenlee, 1986; Small and Sweeney, 1985).

Co-citation context

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. The attributions and functions of the cited article can be indentified through analyzing the semantic of the cited sentences (Siddharthan and Teufel 2007). It can be used to generate a summary of an article Qazvinian and Radev (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 et al. (2009) found that the summarization of the citation context 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 studied the context of co-citation and analyzed the content in which the co-citation paper mentioned (Small 1979). In addition, he analyzed the sentiment of the co-citation context (Small 2010).

Recently, researchers start to consider the position of co-citation in co-citation analysis, and have made some insightful observations. Elkiss et 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 one, and for each level deeper an additional weight of one 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.

We have conducted a preliminary study of co-citation proximity analysis based on three BMC journals (Liu and Chen 2011). Four levels of co-citation proximity were defined in that paper in association with sentences, paragraphs, sections and the article as a whole. The result showed that the distributions of four co-citation levels in three journals are very similar. Over seventy percents of co-citations are occurred in article-level and just two to four percents of co-citations are occurred in sentence-level. Another important finding was the relationship between co-citation frequency and the co-citation proximity in a single journal BMC Bioinformatics. We found a higher proportion of sentence-level co-citations in high co-citation frequencies than low co-citation frequencies.

In this article, we extend the data from three BMC journals to 22. We are particularly interested in verifying if patterns identified in our preliminary results remain valid in the much larger data set.

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.

Distribution of co-citation proximity

Co-citations across different proximity levels are characterized by the distribution of co-citation proximity. Articles from 22 journals are used in this experiment. These journals are selected from the PMC. PMC provides the full text of articles in XML, which makes it a valuable source of citation proximity information. These BMC journals are selected because their impact factors are higher than 2 and each selected journal has 300 or more articles.

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.

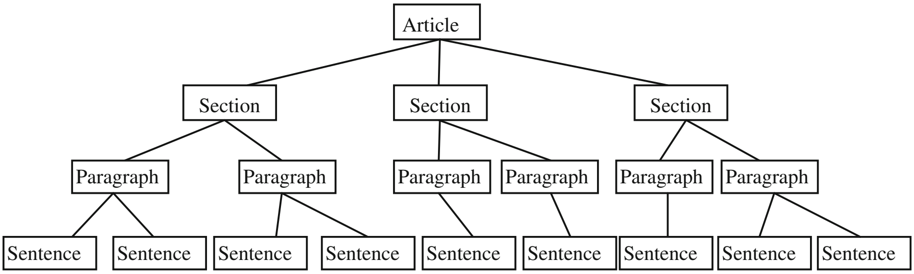


Fig. 1 A four-level co-citation proximity scheme

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 one 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. The *BMC Bioinformatics* journal is chosen. 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. We expect that more sentence-level co-citations occur in *introduction* and *background* sections, for authors always describe similar works in these sections.

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

In terms of the distribution of co-citation proximity by co-citation frequency, 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. *BMC Bioinformatics* was chosen in the previous work (Liu and Chen 2011). We add another two journals *BMC Genomics* and *BMC Cancer* for this experiment to check whether they have the same distribution. Furthermore we also put 22 BMC journals together to do this experiment. 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, first, we constructed each co-citation frequency as a data set. Then these subsets were further divided into four groups based on the amount of co-citation pairs in each co-citation frequency. 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 et al. 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 one contains co-citation pairs that have less than *h* times of co-citation and group two 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, much of research focuses 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 four or more co-citations. Finer-grained proximity-level networks use a threshold of three 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 22 journals are shown in Table 1.

As shown in Fig. 2, the distribution of the co-citation proximity was described by percentage method. The co-citations at sentence and paragraph proximity levels have very similar distributions for these different journals. But in section and article levels, the percentages of the distribution have a high degree of heterogeneity. The highest percentage in article levels is 75.95%, and the lowest percentage just take up 63.32%. The journals in the figure are listed by impact factor from low to high. And the distributions of various proximity levels have less relationship with impact factor. The figure shows that 2–4% of co-citations were made within the same sentences. 6–10% were within the same paragraphs. About 15–23% of co-citations appeared at the section levels. Over 63% of co-citations occurred at the article level. The average distributions of the four co-citation proximities in 22 journals are 3.16, 7.29, 18.16, and 71.39%. The distributions in each journal are very similar to the average distributions. 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 these 22 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.

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

Table 1 The distribution of the co-citation proximity in 22 journals

Journals	Proximity				
	Sentence	Paragraph	Section	Article	Total
BMC_Neurol	10,365	21,554	55,581	194,464	281,964
BMC_Pbulic_Health	69,400	127,391	378,069	1,279,075	1,853,935
BMC_Genet	16,506	38,799	88,690	330,382	474,377
BMC_Med_Res_Methodol	11,050	14,854	34,820	186,302	247,026
BMC_Infect_Dis	28,955	63,885	156,460	430,366	679,666
BMC_Cell_Biol	16,802	46,479	112,462	342,446	518,189
BMC_Biotechnol	12,856	30,271	68,818	230,302	342,247
BMC_Immunol	10,054	29,626	74,387	222,928	336,995
BMC_Cancer	58,720	153,657	420,034	1,104,611	1,737,022
BMC_Neurosci	35,634	96,845	236,153	874,855	1,243,487
BMC_Struct_Biol	11,835	29,525	65,087	309,119	415,566
BMC_Med_Genet	17,859	43,623	98,280	353,936	513,698
BMC_Mol_Biol	19,475	51,068	119,915	391,515	581,973
BMC_Microbiol	35,800	98,085	234,560	854,730	1,223,175
BMC_Dev_Biol	24,060	56,901	146,035	568,091	795,087
BMC_Bioinformatics	94,755	163,527	407,235	1,694,083	2,359,600
BMC_Genomics	105,539	263,979	636,878	3,177,687	4,184,083
BMC_Plant_Biol	21,107	52,736	137,352	562,106	773,301
BMC_Med	16,134	25,038	52,941	265,449	359,562
BMC_Syst_Biol	15,280	42,000	86,619	442,258	586,157
BMC_Evol_Biol	74,371	178,547	442,743	2,065,353	2,761,014
BMC_Biol	16,668	40,135	102,205	458,707	617,715
Average	32,874	75,842	188,878	742,671	1,040,265

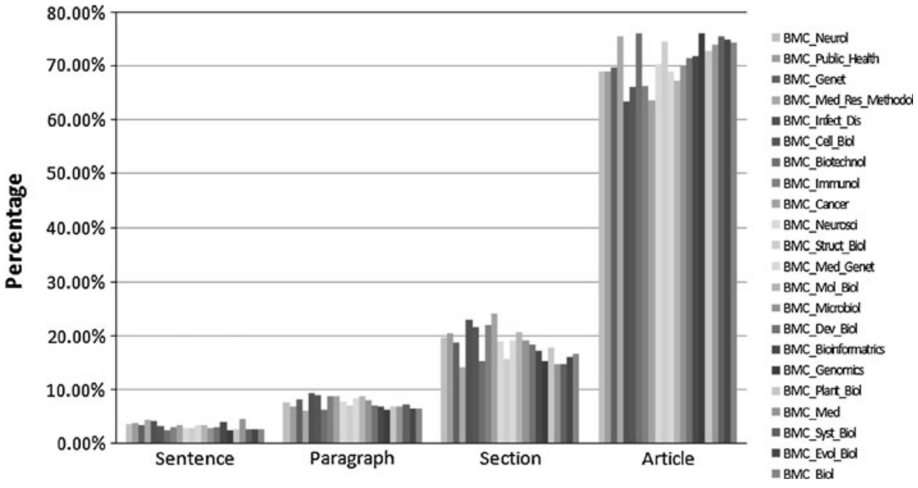


Fig. 2 The distribution of co-citation position in 22 journals

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

Proximity	Organizational Section							
	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

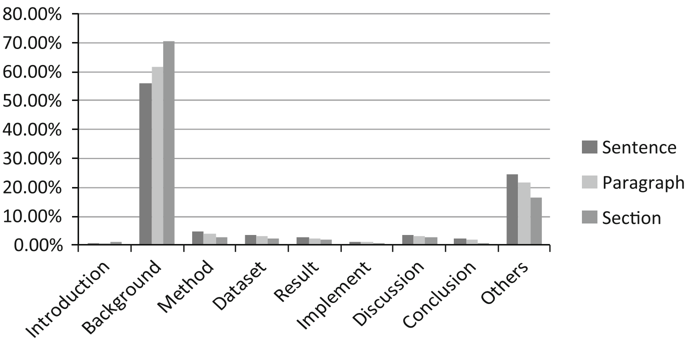


Fig. 3 Distribution of co-citation position in different sections

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.

Co-citation frequency and proximity

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

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 these data sets to represent the general trends (See Fig. 4).

In Fig. 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.

Although the trend is very clear, some of the points did not follow the tread. When the co-citations frequency equals to 25, the sentence-level co-citations take up zero percent and article-level co-citations take up very high percents (64%). There are two co-citation pairs co-cited 25 times:

(a) ALTSCHUL SF, 1997, NUCLEIC ACIDS RES, V25, P3389 and KABSCH W, 1983, BIOPOLYMERS, V22, P2577.
(b) BERMAN HM, 2000, NUCLEIC ACIDS RES, V28, P235 and KABSCH W, 1983, BIOPOLYMERS, V22, P2577.

The article "KABSCH W, 1983, BIOPOLYMERS, V22, P2577" appears in both pairs. The DSSP program mentioned in this paper is widely used and highly cited to compute the protein secondary structure. The article "ALTSCHUL SF, 1997, NUCLEIC ACIDS RES,

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

Frequency	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Sentence	77506	9029	3223	1643	893	513	384	291	229	136	99	59	72	96	52	74	68	28
Paragraph	144552	11429	3368	1533	787	528	330	240	149	114	78	51	53	55	38	24	25	16
Section	376686	19970	5190	2176	1116	666	427	232	159	153	61	85	34	39	36	29	21	19
Article	1597643	63882	15492	7268	3179	1971	1176	757	588	437	257	189	179	146	99	129	107	45
Total	2196387	104310	27273	12620	5975	3678	2317	1520	1125	840	495	384	338	336	225	256	221	108
Frequency	19	20	21	22	23	24	25	26	29	30	31	37	40	43	48	50	54	
Sentence	37	40	20	32	13	24	0	27	41	1	10	4	14	7	28	30	32	
Paragraph	23	21	11	6	15	4	3	7	7	11	9	16	9	6	2	4	3	
Section	18	17	11	6	20	4	15	2	2	5	11	12	4	5	2	2	0	
Article	93	62	21	44	44	16	32	16	37	13	32	42	13	25	16	14	19	
Total	171	140	63	88	92	48	50	52	87	30	62	74	40	43	48	50	54	

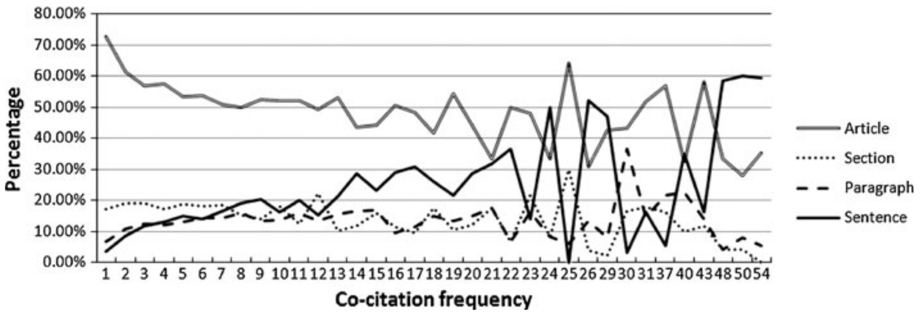


Fig. 4 Proportion of co-citations at the four co-citation proximity levels in *BMC_Bioinformatics*

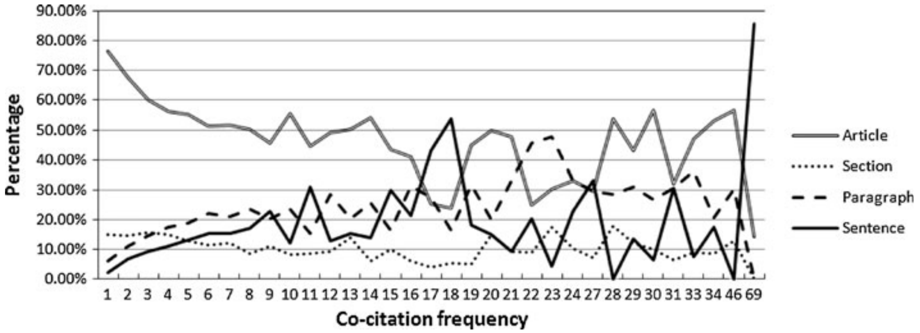


Fig. 5 Proportion of co-citations at the four co-citation proximity levels in *BMC_Genomics*

V25, P3389" is the highest cited paper in journal *BMC_Bioinformatics* which has been cited 274 times. This paper is highly cited for the PSI-BLAST tool which is widely used for searching protein sequence similarities. These two papers are all cited in the research of protein structure, but most of the co-citations appeared in different sections and the average sentence distance of the co-citation is about 55 sentences. Although they have highly co-citation frequency, the relationship between them is not as close as we except in traditional co-citation analysis. So the relationship of the co-cited articles is not only related to the co-cited times, but also related to the co-cited proximity. Figures 5 and 6 show the relationship between co-citation frequency and proximity in *BMC Genomics and BMC Cancer*. The distributions of the four co-citation proximity levels in *BMC_Genomics* with the co-citation frequency lower than 18 are similar to *BMC_Bioinformatics*. When the co-citation frequency grows higher than 18, the distributions of the four co-citation proximity levels became disorder. The reason is that there are very few co-citation pairs in high co-citation frequency, and it is hard to form a trend. The co-citation frequency in *BMC_Cancer* is lower than other two journals. There are just 12 datasets in *BMC_Cancer*, but the distributions of the four co-citation proximity levels are similar to *BMC_Bioinformatics*.

We provide an alternative depiction of the distribution of co-citations at various proximity levels over co-citation frequencies in three journals. The co-citation frequencies are divided into four groups based on the sum of the co-citation pairs in each co-citation frequency. Figure 7a shows the relationship between co-citation pairs and co-citation frequency in *BMC_Bioinformatics*. Most of the co-citation pairs are co-cited just once. The

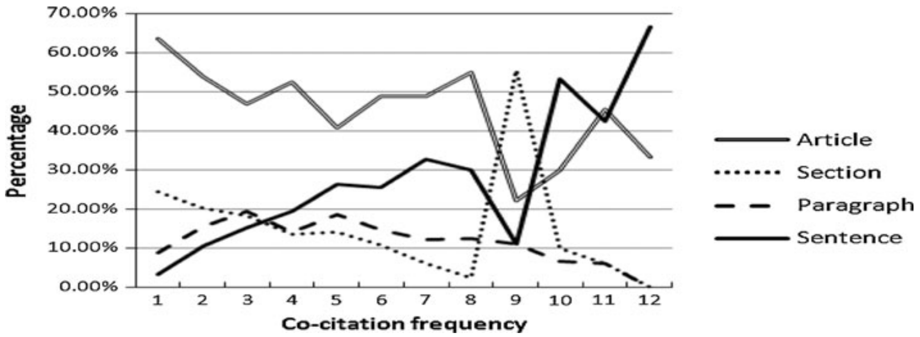


Fig. 6 Proportion of co-citations at the four co-citation proximity levels in *BMC_Cancer*

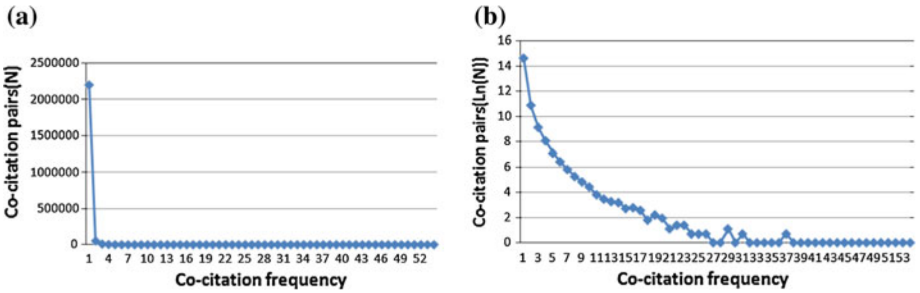


Fig. 7 The relationship between (**a**) co-citation pairs and (**b**) co-citation frequency

logarithm transformation is shown in Fig. 7b in order to reduce the scope the data (Fig. 7b). The datasets can be divided by the logarithmic value of co-citation pairs in each co-citation frequency. For instance, the co-citation frequency datasets in *BMC_Bioinformatics* are divided into the following four groups, 1–2, 3–6, 7–12 and 13–54. Each group takes up about 25% co-citation pairs. As shown in Fig. 8, we got the similar distribution of co-citations at various proximity levels over four co-citation frequency groups in three journals respectively (Fig. 8a, b, c). Followed by the co-citation frequency changing from low to high, the article-level co-citation percentage decreases all the time and the sentence-level co-citation percentage grows higher and higher. We also got the same results about this trend in the dataset of 22 BMC journals together. There are two trend lines in each figure, one is the article trend line and another is sentence trend line. These trend lines fit very well with the distributions and have fitted value higher than 0.9. Although the section-level co-citation has the same distribution trend with article-level co-citation, the section-level co-citation takes up the least percentage in the highest co-citation group and the article-level co-citation always takes up the most percentage in the four 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.

Figure 9 shows the relationship between high and low co-citation frequencies at co-citation proximity levels in *BMC_Bioinformatics* and *BMC_Genomics*. High and low co-citations are defined by co-citation h-index of 23 and 21. In *BMC_Bioinformatics*, the

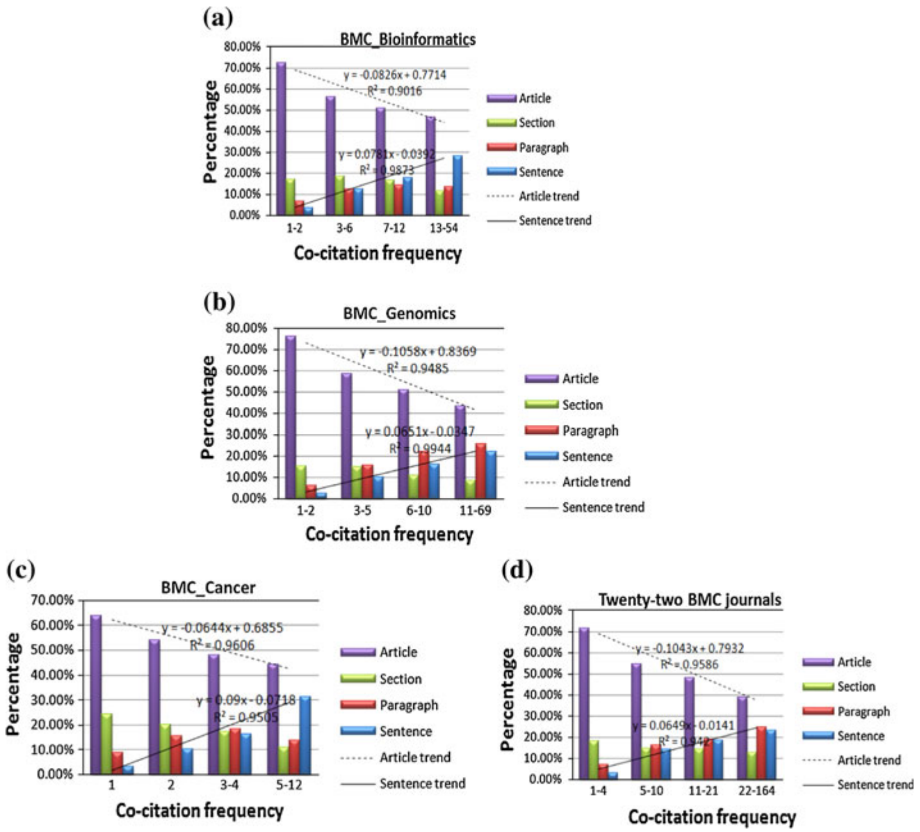


Fig. 8 Distribution of 4 co-citation proximity groups over 4 co-citation frequency groups

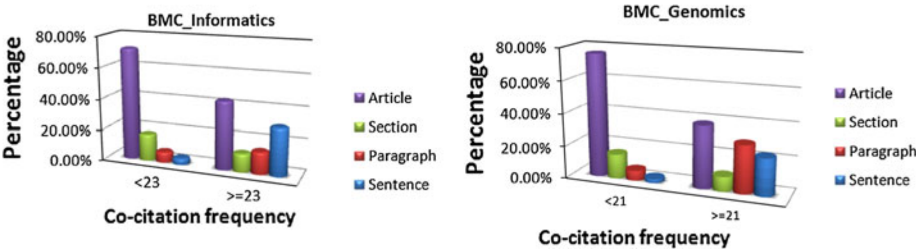


Fig. 9 High and low co-citation split by co-citation h-index and corresponding proximity in BMC_ Bioinformatics(*left*) and *BMC_Genomics*(*right*)

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. In *BMC_Genomics*, the high co-citation group has 13 datasets and the low co-citation group has 20 datasets. The results have some differences in two journals. The distributions in the low co-citation group are similar, but in the high co-citation group, article- and sentence-level co-citations are prominent in *BMC_Bioinformatics* while article- and paragraph-level co-citations are prominent in *BMC_Genomics*. The changing trends of the two groups in both journals are similar, the percentages of article- and section-levels co-citations decrease while the percentage of sentence- and paragraph-levels co-citations grow.

Co-citation proximity in context

A traditional co-citation network and overlays of co-citation networks at four levels of proximity are shown in Fig. 10. 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 (Fig. 10a). The sentence-level network has 267 edges (Fig. 10b). The paragraph-level network has 83 edges (Fig. 10c). The section-level network has 126 edges (Fig. 10d). The article-level network has 1,825 edges (Fig. 10e).

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.

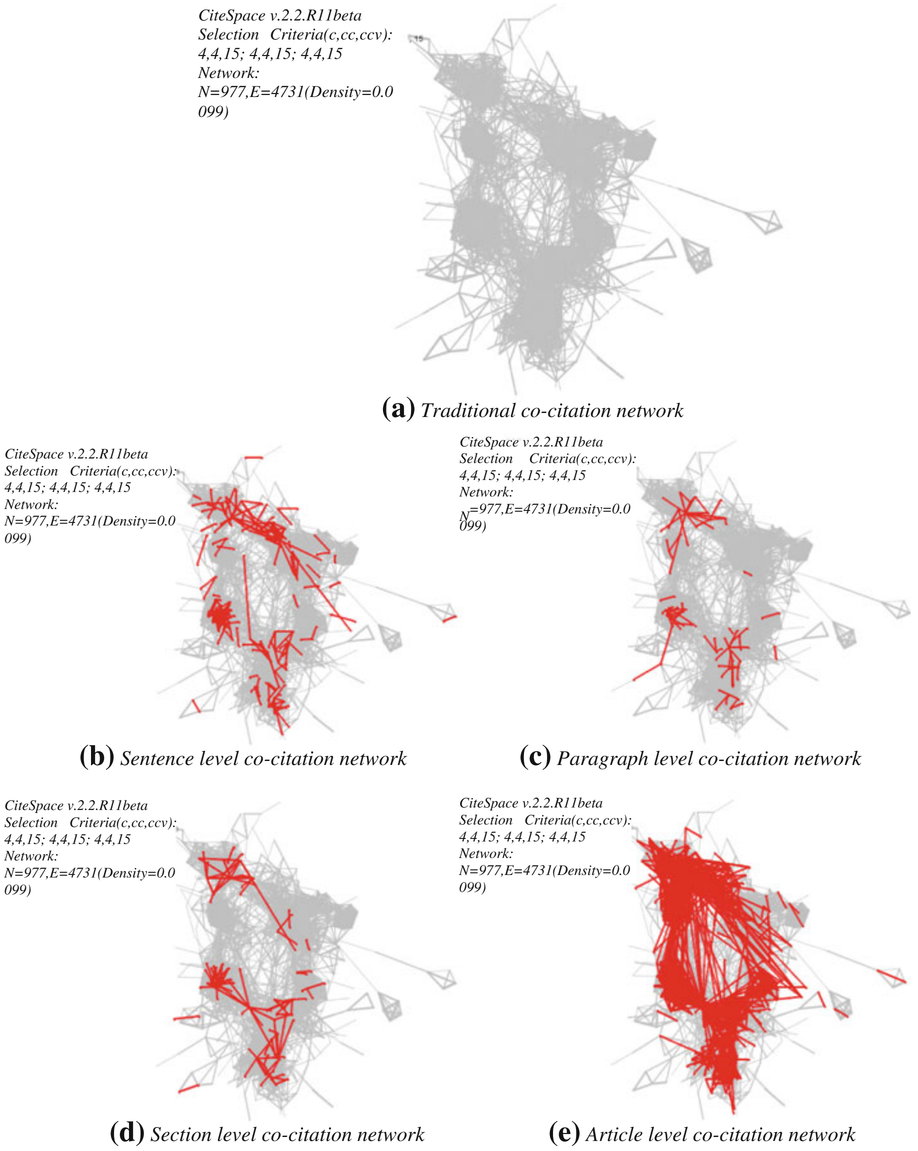
Discussion

First, we have found that the distribution of co-citation at each proximity level has some differences across 22 different journals, but the whole distribution tendencies are similar that the article level take up the most percentage and section level next, sentence level account the lowest percentage. The relationship between the distribution and the journal impact factor is not obvious. 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



(a) *Traditional co-citation network*

(b) *Sentence level co-citation network*

(c) *Paragraph level co-citation network*

(d) *Section level co-citation network*

(e) *Article level co-citation network*

Fig. 10 Traditional co-citation network and proximity-level co-citation network overlays

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.

Conclusions

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. The distributions of the co-citation proximity in 22 journals have some differences, but the main trends are similar. With the co-citation frequency going up, the sentence-level co-citation tends to take up more percents and the article-level co-citation tends to take up less percents. But not all the high co-citation frequencies have the high sentence-level co-citation percentage. 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 was supported by National Social Science Foundation of China (grant number 08BTQ025). An extended version of a paper presented at the 13th International Conference on Scientometrics and Informetrics, Durban (South Africa), 4–7 July 2011 (Liu S. & Chen C., 2011).

References

Siddharthan, A., & Teufel, S. (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. In *Proceedings of NAACL/HLT-07*.

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

Callahan, Alison., Hockema, Stephen., & Eysenbach, Gunther. (2010). Contextual cocitation: Augmenting cocitation analysis and its applications. *Journal of the American Society for Information Science and Technology, 61*(6), 1130–1143.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences of the United States of America, 101*(suppl 1), 5303–5310.

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. In *Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries* (pp. 19–28). Vancouver: ACM.

Chen, C., Song, I. Y., Yuan, X. J., & Zhang, J. (2008). The thematic and citation landscape of data and knowledge engineering (1985–2007). *Data & Knowledge Engineering, 67*(2), 234–259.

Elkiss, Aaron., Shen, Siwei., Fader, Anthony., 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.

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. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569–16572.

Liu, S., & Chen, C. (2011). The effects of co-citation proximity on co-citation analysis. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of ISSI 2011-The 13th International Conference on Scientometrics and Informetrics* (pp. 474–484), Durban, 4–7 July 2011.

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

Qiaozhu, M., & Xiang, Z.C. (2008). Generating impact-based summaries for scientific literature. *Proceedings of ACL'08* (pp. 816–824). Columbus: ACL.

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.

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.

Qazvinian, V., & Radev, D. R. (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.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., & Zajic, D. (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.

Small, H. (2010). Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics*. doi:10.1007/s11192-011-0349-2.

Small, H., & Greenlee, E. (1986). Collagen research in the 1970s. *Scientometrics, 10*(1–2), 95–117.

Small, H., & Sweeney, E. (1985). Clustering the science citation index using co-citations.1. A comparison of methods. *Scientometrics, 7*(3–6), 391–409.