Yang, lijun and Han, Liangxiu and Liu, Naxin (2019) A new approach to journal co-citation matrix construction based on the number of co-cited articles in journals. Scientometrics. ISSN 0138-9130

https://e-space.mmu.ac.uk

A New Approach to Journal Co-citation Matrix Construction based on the Number of Co-cited articles in Journals

Lijun Yang[a], Liangxiu Han[b], Naxin Liu[a]

[a]School of Information Management, Sun Yat-sen University, Guangzhou 510006, China

[b]School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester M1 5GD, United Kingdom

Abstract: Co-citation analysis is one of the most important methods in information science. Journal co-citation analysis has been widely used to analyze the relevance, relationship and structure of underlying articles between journals. Accurate construction of co-citation matrix is a key to accurate journal co-citation analysis. However, the traditional co-citation matrix construction based on co-citation frequency of journals does not accurately reflect the similarity between journals. This paper proposes a new construction method of co-citation matrix based on the number of co-citation articles in journals. The experimental validation has been conducted with real datasets from Chinese Social Science Citation Index (CSSCI) and National Knowledge Infrastructure (CNKI). The results show that the proposed method can accurately capture the similarity between journals and outperform the existing approaches (i.e. co-citation frequency and co-citation ratio approaches). In addition, the proposed method does not need the full-text index of a paper, which provides added value in the field.

Keywords: journal co-citation analysis; co-citation matrix; number of co-cited articles; co-citation frequency; co-citation ratio

1 Background

Co-citation analysis pioneered by American intelligence scientist Henry Small(1973) and former Soviet intelligence scientist Marshakova(1973) has been widely used for identifying topically related publications for search engines and clustering relevant publications to understand the structure of science. Journal co-citation analysis as a branch of co-citation analysis mainly focuses on analyzing the relevance and similarity between journals.

In the context of journal co-citation analysis, the complicated co-citation relationship between journals can be clearly expressed by constructing co-citation matrix between journals. The element of co-citation matrix is called co-citation strength. The greater the

co-citation strength is, the closer the relationship between the two journals (or more similar) is. Most existing researches(e.g., Qiu 2009; Liu and Chen 2012; Jeong et al 2014) construct co-citation matrix based on co-citation frequency (the number of times two journals are co-cited) to express co-citation strength for representing similarity and relevance between journals. For example, Qin's work(2010) based on co-citation frequency analyzed journal correlationship. Wageningen UR Library developed a journal recommendation system to serve researchers according to the co-citation relationship of articles (M.G.P. van Veller and W. Gerritsma 2015), which provided more accurate co-citation analysis. This system was specifically based on their researchers' articles and also used the co-citation frequency approach for citation preferences for each of the five science groups that comprise Wageningen UR. As the author stated in their work, the limitation of their work is that the accuracy of co-citation analysis will be affected if there is a new article outside of those articles. However, the existing work based on the co-citation frequency does not reflect real similarity between journals. This is because the similarity calculation should also consider whether the content of an article between journals are similar, which is the core attribute of journals representing the extent of similarity. To illustrate this, we here provide a concrete example to show that two pairs of journals have same co-citation frequency while the number of co-citation articles between these two pair of journals vary greatly.

    For instance, the co-citation frequencies between two journals: <*Information Science*> and <*Documentation, Information & Knowledge*>, and between <*Documentation, Information & Knowledge*> and <*Journal of Library Science in China*> in the first issue of 2010 are all two. However, ten articles in <*Information Science*> and <*Documentation, Information & Knowledge*> were co-cited whereas only four articles are co-cited in <*Documentation, Information & Knowledge*> and <*Journal of Library Science in China*>. From this example, it is noted that one pair of journals have ten similar articles, whereas the other pair of journals just have four similar articles. So the similarity of these two pairs of journals are not same. This further proves that the distance (or similarity) between two journals based on co-citation frequency only is not sufficient. It is therefore necessary to have

1 a new approach to measure similarity.

2 In this work, we have proposed a new approach for accurate construction of co-citation

3 matrix based on the number of co-cited articles between journals, which has been validated

4 with a case study using real datasets from CSSCI and CNKI. The rest of the papers is

5 organized as follows: Section two describes related work; Section three details the proposed

6 approach; Section four presents an experimental validation with a case study; Section 5

7 concludes the work.

8 **2 Related work**

9 Since McCain(1990) introduced the method in a typical author co-citation analysis, much

10 effort has been devoted to this area (e.g., Leydesdorff & Vaughan 2006, Waltman 2013,

11 Mongeon 2016, Bu et al 2017, Susana et al 2018).

12 The early researches (White 1981; Qiu 2008) for co-citation analysis only concerned

13 about the diagonal elements of a co-citation matrix instead of the all other elements of the

14 whole matrix.

15 For most subjects, the more articles a journal has, the more citation frequencies it is

16 likely to get. A journal has different journal attributes such as the start time/year, the number

17 of issues per year, and the number of articles per issue, etc., which affect the accuracy of

18 co-citation analysis. For example, for those journals that created earlier, published more

19 issues each year with large volumes normally obtain more citation frequencies. This means

20 the co-citation analysis based on co-citation frequency only will not be accurate. To avoid the

21 influence of the number of journal articles on the total citations of journals, the authors(Wang

22 et al 2009) extended the previous work and constructed a co-citation matrix using co-citation

23 ratio. This method is still based co-citation frequency approach which used a co-citation ratio

24 to indicate the co-citation strength. Their specific steps are as follows: suppose A, B are two

25 journals, C(AB) is co-citation frequency of A and B, C(A), C(B) are the total cited

26 frequencies of A and B, respectively, then co-citation strength of A and B can be calculated

27 by

$$Strength(A,B) = \frac{C(A,B)}{C(A)} \times \frac{C(A,B)}{C(B)} = \frac{[C(A,B)]^2}{C(A) \times C(B)} \qquad (1)$$

Based on the authors' claim (Wang et al 2009), their method could reduce errors compared with traditional methods. However, as shown in equation 1, their method was still based on co-citation frequency and therefore not appropriate for accurate co-citation analysis.

Some scholars also argue that the closer the reference position is, the stronger the relationship between references is, and vice versa. This means, citations in the same sentence have a closer relationship than citations in different sentences in the same paragraph. Base on this idea, Eto(2013) also divided co-citation into four levels: enumeration, same sentence, same paragraph, across paragraph, and weights assigned to 4,3,2,1, respectively. And he also investigated the effect of reference position on information retrieval. Gipp and Beel(2009) divided co-citation into five levels: citation in same sentences, citation in same paragraphs, citation in same sections, citation in a journal, citation in different volumes in a journal, and weights assigned to 1, 1/2, 1/4, 1/8, 1/16, respectively. However, the main problem with these methods is the subjectivity of the weight assignment.

To address the problem of subjective issues in the weight assignment, some researchers (Liu et al 2013) have proposed a co-citation matrix construction method based on citation content and level. The main steps of this method is as following: (1)based on position, classifying co-citation into four levels including sentence, paragraph, section and article levels; (2)weighting co-citation relationship of each level by average similarity of context for this level; (3)calculating the co-citation strength of each two articles; (4)constructing co-citation matrix by the following equation:

$$Strength(A,B) = \sum_{i=1}^{n} (Weight(P_i) \times Frequency_{AB}(P_i)) \qquad （2）$$

where, $Strength(A,B)$ is co-citation strength of article A and B, n is the number of levels, P_i is co-citation relationship in level i, $Weight(P_i)$ is the weight of level i, $Frequency_{AB}(P_i)$ is co-citation frequency of article A and B in level i.

Compared with the citation position-based matrix construction approaches, this method

has a "quantitative" weight assignment, which is more objective. However, due to the fact that only a few databases can provide the full-text XML indexing information of the articles, which is difficult to put it into practice. Therefore, the above methods are not suitable for journal co-citation analysis, especially for those who do not provide full-text indexing XML information database.

3 The Proposed Method

3.1 Rational behind

The similarity between journals is normally decided based on the extent of content similarity between journals instead of using co-citation frequency.

Here if A, B, C represent three different journals, A1, A2, A3 represent three articles from journal A; B1, B2, B3 represent three articles from journal B; C1, C2, C3 represent three articles from journal C. Based on the traditional co-citation frequency method, the Group I and Group II shown in the Figure 1 below have co-citation frequency as 3 respectively, which means the similarity between Journal A and B is same as the one between Journal A and C from the perspective of co-citation frequency.
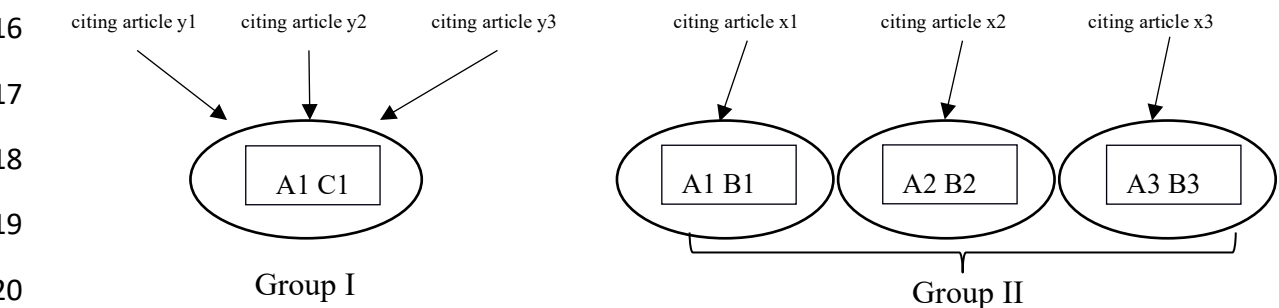
Fig.1. An example scenario of co-citation relationship between journal A, B and A, C

However, this is not accurate in some cases. In fact, if in Group I (journal A and journal C), A1 and C1 are co-cited by paper y1, y2 and y3; and in Group II (journal A and journal B), A1 and B1 are co-cited by paper x1, A2 and B2 are co-cited by paper x2, A3 and B3 are co-cited by paper x3.

This means only two articles in Group I (A and C) are co-cited by other articles while six articles in Group II (A and B) are co-cited by other articles. This means the similarity between Journal A and B is more than the one between Journal A and C.

3.2 Construction of co-citation matrix based on number of co-citation articles

Based on the section 3.1, we propose a new approach for co-citation matrix construction based on the number of co-citation articles, as shown in equation 3

$$Strength(A,B) = \frac{J(A,B)}{J(A)+J(B)} \qquad (3)$$

Where *Strength(A,B)* represents an element of co-citation matrix. J(A) and J(B) represent the number of articles in journal A and B, respectively.

J(A,B) represents the deduplicated number of articles co-cited in journal A and B. This is because the number of co-cited articles between journals can't exceed the original number of articles in journals and duplicated co-cited articles can only be counted as one when calculating the number of co-cited articles. So we can define the number of co-cited articles of journal A and journal B by J(A, B).

In most of cases, a journal that publishes more papers has more chances to obtain more citations, hence there are more chance to gain citations. In addition, different journals publish different numbers of articles. Therefore, we use division for fair calculation of the strength of co-citation.

3.3 Consideration of similarity between journals

In this work, given two journals A and B, the calculation of the distance between two journals is defined as follows:

$< A, B > = \|K_A, K_B\|$, where K_A and K_B represent keywords vectors of Journal A and B, respectively, and $K_A = \{(A_{keyword\text{-}1}, A_{keyword\text{-}1\_freq}), (A_{keyword\text{-}2}, A_{keyword\text{-}2\_freq}), \cdots\}$,

$K_B = \{(B_{keyword\text{-}1}, B_{keyword\text{-}1\_freq}), (B_{keyword\text{-}2}, B_{keyword\text{-}2\_freq}), \cdots\}$; $A_{keyword\text{-}i}$, $B_{keyword\text{-}i}$ are keywords of journal A and B, respectively, i=1,2,$\cdots$; $A_{keyword\text{-}i\_freq}$, $B_{keyword\text{-}i\_freq}$ are the frequencies of $A_{keyword\text{-}i}$, $B_{keyword\text{-}i}$, respectively, i=1,2,$\cdots$; $\|\cdot\|$ could be either Euclidean distance, or Dice coefficient, or Cosine coefficient or Jacobian coefficient.

Similarly, we can calculate the distance between Journal A and Journal Group X as

follows: $$< A, X > = \frac{1}{n} \sum_n < A, B_n > \qquad (4)$$

where B_i is a Journal of Journal Group X, i=1,2,⋯.

4. Experimental evaluation with a case study

In this section, we will use a case study with real datasets from CNKI to evaluate the model accuracy and efficiency. We only focus on comparison study with two traditional co-citation frequency, co-citation ratio approaches. It should be noted that, we didn't include the co-citation content and level based approaches due to limited functions for analyzing Chinese databases (such as CNKI, CSSCI, etc.).

4.1 Dataset description

This work uses a publicly available database CSSCI (Chinese Social Science Citation Index) and CNKI (China National Knowledge Infrastructure) for co-citation analysis, with a particular focus on discipline of "Library, Information and Archival Science". The choice of data is because journals in CSSCI are more standard and academic rigor, and each journal in the CSSCI list has a relatively clear theme, which is suitable for the empirical object of co-citation analysis of journals, while CNKI is the largest academic document database in China with more than 6,100 journals which provides a complete and representative dataset for analysis. At the same time, bibliographic information can be retrieved and downloaded from CNKI, which means co-citation matrix can be constructed at a lower cost. We will mainly choose the journals published in 2010 since the maximum citation number of articles normally occurs within 3~5 years after publication (Huang 2011).

For convenience, we introduce some symbols as follows:

Table 1 Symbols of analysis object

| No.	Journal name
J1	Journal of Academic Libraries
J2	Journal of Library Science in China
J3	Journal of the National Library of China
J4	Library

| J5	Library and Information Service
J6	Documentation, Information & Knowledge
J7	Library Work and Study
J8	Library Development
J9	Library Tribune
J10	Research on Library Science
J11	Library Journal
J12	Library & Information
J13	Data Analysis and Knowledge Discovery
J14	Information Studies: Theory & Application
J15	Journal of the China Society for Scientific and Technical Information
J16	Information Science
J17	Journal of Intelligence
J18	Information and Documentation Services
J19	Archives Science Bulletin
J20	Archives Science Study

4.2 Co-citation analysis

To analyze the data, it is necessary to preprocess the dataset by exclusion of irrelevant articles such as notice of meeting, journal brief introduction etc. After the preprocessing step, we have conducted co-citation analysis on three methods (co-citation frequency, co-citation ratio, and our proposed method) and performed comparison. We have used cluster analysis which is a method that divides a group of heterogeneous populations into subgroups with higher isomorphic properties. If journals are clustered into the same group, then these journals tend to be similar. Figure 2, 3, and 4 show the clustering analyses of journals based on three different approaches. In these three figures, the vertical axis represents the number of journals (also shown in Table 1), and the horizontal axis represents the rescaled distance cluster combination. Table 2 is the summary of experimental results based on these three figures.

1
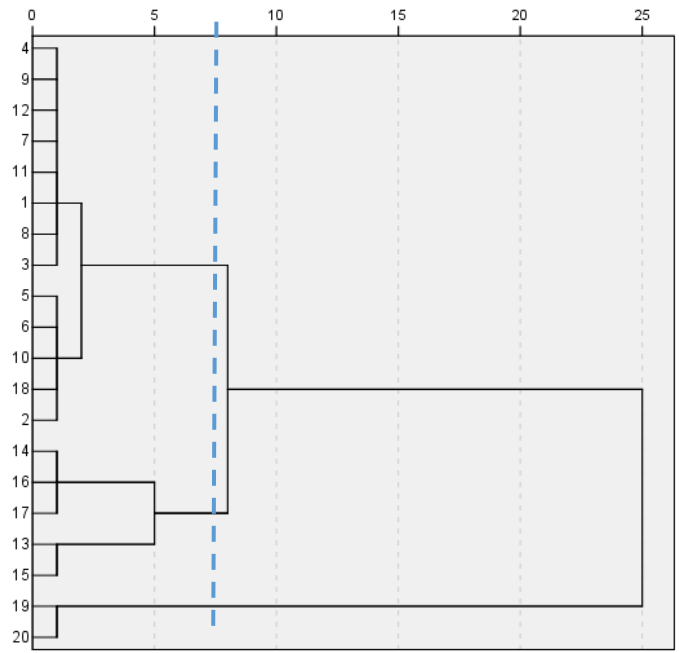
Fig. 2. Dendrogram based on co-citation frequency

2

3

4
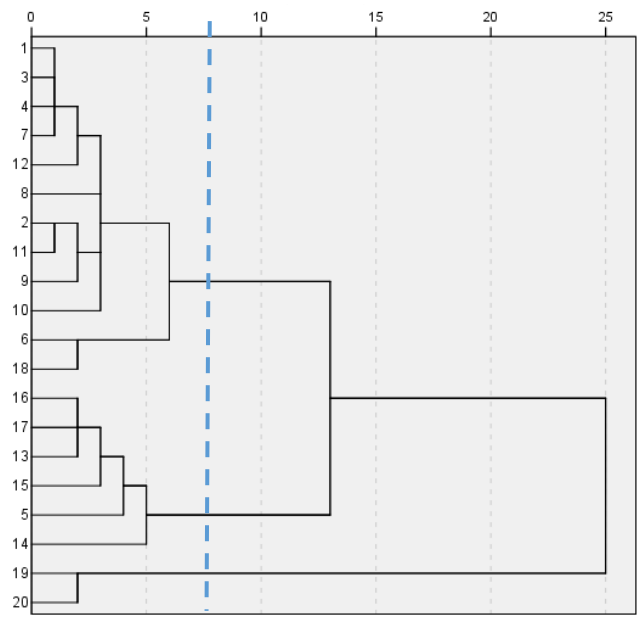
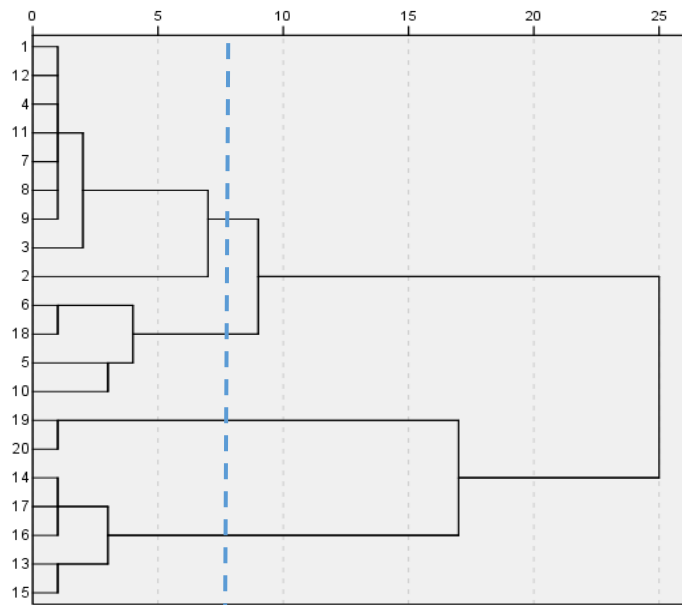Fig. 3. Dendrogram based on co-citation ratio

5

6

Fig.4. Dendrogram based on the number of co-cited articles

Table 2 Experimental results based on three different methods

	Co-citation frequency	Co-citation ratio	The proposed method
Group I	J13, J14, J15, J16, J17,	J5, J13, J14, J15, J16, J17	J13, J14, J15, J16, J17
Group II	J19, J20	J19, J20	J19, J20
Group III	J1, J2, J3, J4, J5, J6, J7, J8, J9, J10, J11, J12, J18	J1, J2, J3, J4, J6, J7, J8, J9, J10, J11, J12, J18	J1, J2, J3, J4, J7, J8, J9, J11, J12
Group IV			J5, J6, J18, J10

4.3 Result analysis and discussion

In this section, we will analyze and discuss the experimental results based on the Figure 2, 3, 4 and Table 2.

(1) Comparison between the co-citation frequency-based approach and the proposed approach

We have compared two methods: the co-citation frequency and the proposed method. The summary of the analysis result is shown in Table 2. In Table 2 (the second column and the fourth column), it shows that the journals have been clustered into four groups (I, II, III and

IV) based on proposed method. For Group I and II, the results based on co-citation frequency and the proposed method are same. For Group I, there are five journals in it, which are clustered into the same group because they all fall within the remit of three disciplines (library science, information science and archival science). For instance, library science mainly focuses on information organization; information science mainly focuses on information development and utilization; and the archival science focuses on information preservation (He 2005).

For Group II, there are two journals (J19, J20) being clustered together. Because J19 and J20 are only two archival journals out of the 20 journals mentioned above. From the discipline point of view, there is rarely an overlap between the archival science journal and library and information science journal. The archival science is relatively independent and at the present there are only two core archival journals in China. Researchers often tend to cite core journals with high quality and high impact, thus significantly reducing the opportunities for co-citation with other journals in library and information sciences.

It is also noted that some journals may publish some articles which are not strongly related to the scope of the journal, in order to further prove the effectiveness of our proposed method, we have conducted another experiment for similarity calculation using equation 4 in Section 3.3. This similarity calculation is mainly based on keywords as they represent the content of articles, reflecting similarity between journals. For example, in Figure 2, journals including J2, J5, J6, J10 and J18 belong the same subgroup. In Figure 4, journals including J2, J1, J12 J4, J11, J7, J8, J9 and J3 belong another subgroup. To determine which subgroup is more appropriate for J2, we have calculated similarity based on keywords vector of journals in these two subgroups, respectively, because the keywords represent the content of an article, which is more accurate for similarity calculation. The experimental result is shown in Table 3.

Table 3 Average distance between J2 and subgroup-1, subgroup-2

	Euclidean distance	Dice coefficient	Cosine coefficient	Jacobian coefficient
subgroup-1	77.6	0.25	0.469	0.15
subgroup-2	53.1	0.32	0.531	0.21

In general, the greater the distance is, the less the similarity is; the greater the correlation coefficient is, the greater the similarity is. For example, the Euclidean distance between J2 and subgroup-1 is 77.6 which is more than that of J2 and subgroup-2. It means that compared with journal in subgroup-1, J2 is more like those journals in subgroup-2. From the perspective of correlation coefficient, we can get the same conclusion.

So we can say that the data in Table 3 shows that the clustering result using our proposed approach outperforms the co-citation frequency-based approach.

(2) Comparison between the co-citation ratio-based approach and the proposed approach

Table 2 (the third column and the fourth column) also shows the different co-citation results obtained from co-citation frequency and the proposed approaches.

Similarly, for the two co-citation matrix construction methods, the distribution of *J5* has changed. We have also calculated the distance between *J5* and subgroup3 and subgroup4, respectively. The experimental result is shown in Table 4.

Table 4 average distance between *J5* and subgroup 3, subgroup4

	Euclidean distance	Dice coefficient	Cosine coefficient	Jacobian coefficient
Subgroup3	123.25	0.45	0.67	0.32
Subgroup4	111.20	0.51	0.85	0.40

Similarly, from Table 4, we can conclude that whatever distance is selected, the results based on proposed method in this article are better than the one based on the co-citation ratio-based approach.

5. Conclusion and Future work

This paper has proposed and validated a new co-citation matrix construction approach based on the number of co-cited articles. We have conducted co-citation analysis experiments with a real dataset and performed comparison study between the proposed approach, co-citation frequency, co-citation ratio based approaches. Based on similarity analysis including clustering analysis and four most commonly used methods (Euclidean distance, Dice coefficient, Cosine coefficient Jacobian coefficient), our proposed method outperforms existing co-citation frequency and co-citation ratio-based approach. Moreover, the proposed method doesn't require a full-text XML indexing and has a wide range of applications.

It is well-known that some other co-citation analysis methods such as author co-citation analysis could be different from journal co-citation analysis. This is because a journal normally has a fixed number of papers per issue while an author could publish a various number of papers. Therefore, the future work will be focusing on investigating correlation between the proposed method and other methods such as author co-citation analysis.

Author contributions

Lijun Yang: Conceived and designed the analysis, collected the data, contributed data or analysis tools, performed the analysis, wrote the paper.

Liangxiu Han: Conceived and designed the analysis, collected the data, contributed data or analysis tools, performed the analysis, wrote the paper.

Naxin Liu: Conceived and designed the analysis, collected the data, contributed data or analysis tools, performed the analysis, wrote the paper.

References

Bu Y., Ni Sh. K. & Huang W. B. (2017). Combining multiple scholarly relationships with author cocitation analysis: A preliminary exploration on improving knowledge domain mappings, Journal of Informetrics, 11(3), 810-822

Eto M. (2013). Evaluations of Context-based Co-citation Searching. Scientometrics. 94(2), 651-673

Gipp B. & Beel J. (2009). Identifying Related Documents for Research Paper Recommender by CPA and COA. Proceedings of International Conference on Education and Information Technology. Berkeley: International Association of Engineers, 636-639

1 He L.R. (2005). Analysis on Correlative Tendency of Frequency of Cited Journals or Impact Factor and
2 Article Counts. Chinese Journal of Scientific and Technical Periodicals. 16(4), 500-503
3
4 Huang L.P. (2011). Literature Aging Research Based on Citation Analysis: An Analysis in Library and
5 Information Science and Management Science. Journal of Intelligence. 30(10), 30-35
6
7 Marshakova, I.V.(1973). (1973). System of Document Connections Based on References. Scientific and
8 Technical Information Serial of VINITI, 6(2), 3-8
9
10 Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. Journal of
11 Informetrics, 8(1), 197–211.
12
13 Leydesdorff, L. & Vaughan, L. (2006), Co-occurrence matrices and their applications in information
14 science: Extending ACA to the Web environment. Journal of the American society for information science
15 and technology, (57)12, 1616–1628
16
17 Liu Sh.B. & Chen Ch.M. (2012). The proximity of co-citation. Scientometrics, (91)2, 495–511.
18
19 Liu Sh.B., Zhang Ch.B., Ding K. & Liu Z.Y. (2013). The Improvement of Co-citation Analysis Based on
20 the Citation Context and Citation Position. Journal of the China Society for Scientific and Technical
21 Information. 32(12), 1248-1255
22
23 McCain K.W. (1990). Mapping authors in intellectual space: a technical overview. Journal of the American
24 society for information science, 41(6):433-443
25
26 Mongeon, P. & Paul-Hus. A. (2016). The journal coverage of Web of Science and Scopus: a comparative
27 analysis, Scientometrics, 106(1), 213-228
28
29 Qin Ch.J. (2010). The Empirical Research of the Knowledge Domains Map Between Subject Relationship
30 Based on Journal Co-citation Analysis Method, Journal of Modern Information, (5):9-11
31
32 Qiu J.P., Ma R.M. & Li Y.J. (2008). Reconsiderations on Co-citation Analysis, Journal of the China Society
33 for Scientific and Technical Information, 27(1), 69-74
34

Qiu J.P. & Li J.P. (2009). The Co-citation Analysis of the Journal of Library and Information Science. Information Studies: Theory & Application. 39(2), 17-19

Small H.(1973). Co-citation in the Scientific Literature：A New Measure of the Relationship Between Two Documents. Journal of American Society for Information Science, 24(4), 265-269

Susana S.G., Juan G., & David M.F. (2018). Reference density trends in the major disciplines, Journal of Informetrics, 12(1), 42-58

Veller, M.G.P. van & Gerritsma, W. (2015). Development of a journal recommendation tool based upon co-citation analysis of journals cited in Wageningen UR research articles. Qualitative and quantitative methods in libraries 4(2): 233-257.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., & Visser, M. S. (2013). Some modifications to the SNIP journal impact indicator? Journal of Informetrics, 7(2),272–285.

Wang X.W. & Liu Z.Y. (2009). Study on journal classification based on co-citation ratio analysis. Science Research Management，30(5), 187-195

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