Author Relationship Mining based on Tripartite Citation Analysis

Feifei Wang[1], Junwan Liu[2], Siluo Yang[3]

[1] *feifeiwang@bjut.edu.cn,* [2] *liujunwan@bjut.edu.cn*
School of Economics and Management, Beijing University of Technology, 100024 Beijing (China)

[3] *58605025@qq.com*
School of Information Management, Wuhan University, 430072 Wuhan (China)

Abstract
This study scrutinizes potential author relationships according to the findings based on the tripartite citation analysis. It focuses on Author co-citation analysis (ACA), author bibliographic-coupling analysis (ABCA) and author direct citation analysis (ADCA). By algorithm design and empirical analysis, the deduction from results of ACA, ABCA and ADCA to potential author relationships mining could be probable, and the empirical process would be practicable.

Conference Topic
Citation and co-citation analysis

Introduction
Citation analysis is a mature quantitative research method in Bibliometrics and Scientometrics. It is widely used in scientific evaluation, scholarly communications, academic behavior analysis, and information retrieval. Author citation analysis mainly includes three types: author co-citation, author coupling, and author direct citation.

Author co-citation analysis (ACA) is the most widely used method for the empirical analysis of disciplinary paradigm, and is frequently studied and improved upon. Many ACA studies have been conducted since Small (1973) introduced document co-citation analysis and White and Griffith (1981) introduced ACA. Bibliographic coupling was proposed as early as 1963 (M. M. Kessler, 1963). However, author bibliographic-coupling analysis (ABCA), i.e. author-coupling relationships, did not get much attention until it is formally put forward and empirically studied by Zhao (Zhao & Strotmann, 2008).

Direct citation is sometimes also called inter-citation or cross citation (Zhang et al., 2009). Compared with co-citation and bibliographic coupling, direct citation is a direct citing relationship without a third party paper. Although researchers are aware of direct citation analysis and employed from time to time (Shibata et al., 2008), it was ignored because of the unavailability of data, difficulty of implementation, and long time windows to obtain a sufficient linking signal for clustering. However, scholars are gradually paying attention to this topic (Boyack & Klavans, 2010). A number of studies have focused on journal direct citation or comparative analysis of methods. The author direct citation analysis was more clearly explored by Wang et al. (2012). Wang used this method to reveal the knowledge communication and disciplinary structure in Scientometrics. This process is named "author direct citation analysis" (ADCA) (Yang & Wang, 2015).

All of these three kinds of citation analysis methods can reveal separately the author relationship in a field. Then, how about the similarities or diversity among the tripartite citation relationships at author level? And, how could the tripartite relationships be synthetically presented to the readers or the result users? We have tried to answer these two questions in previous studies (Wang, 2014), even though the effort is still limited. Persson (2010) and Gómez-Núñez et al. (2011, 2014, 2015) tried to combine these citation measures

in a normalized way to weight existing direct citation relationships between articles or journals.

The following question is worthy of investigation as well: Could we discover potential author relationships according to the findings based on the tripartite citation analysis? To give an example: in a field, author A's paper and author B's paper both are cited by the same paper C, then A and B have co-citation relationship, which can be marked as (A, co-citation, B). Author C and author D, when citing the same paper in their respective articles, have bibliographic-coupling relationship, marked as (C, bibliographic-coupling, D). In addition, if C and A cite each other, then C and A have direct-citation or cross-citation relationship, marked as (C, directly citing, A) or (A, directly citing, C) or (A, cross citation, C). According to these primary relationships, could we deduce an integrated relationship between A and D, or B and C, even B and D? And, what will be the association strength in these potential relationships? These are the key problems that we answer in this study.

Data and methodology

Basic Data

Since the journal *Scientometrics* is one of the most representative communication channels in the field of Scientometrics, it reflects the characteristic trends and patterns of the past decades in scientometric research (Schubert A 2002). This study is based on bibliographic data based on all types of documents published in *Scientometrics* from 1978 to 2011, retrieved from the Web of Science. Author names including the cited authors were normalized because some authors may report their names differently in different papers. We identified each author by his or her surname and first initial only; the same applies to cited authors.

Methodologies

In this study, bibliometrics method is applied to identify the core authors (94 first authors who have published 5 or more papers and simultaneously have a cited frequency over 10) in Scientometrics filed. Author co-citation analysis (ACA), author bibliographic-coupling analysis (ABCA) and author direct citation analysis (ADCA) are respectively used to discover author relationships with co-citation, bibliographic-coupling and direct-citation. Co-occurrence analysis and deductive reasoning methods are used to mine potential author relationships on the findings of the tripartite citation analysis. VBA program processes all kinds of citation analysis data. The final results of author relationship are visualized with Pajek.

Results and discussion

According to the tripartite citation analysis, we obtain three original relation matrixes and their corresponding normalized matrixes (Fig. 1). The normalization method is based on Salton's Cosine similarity measures, which returns similarity values ranging between 0 and 1. In order to describe the directivity of citing behaviour and achieve vectorial deducing, the direct citation matrix is unsymmetrical.

Core author co-citaion matrix

	Garfield E	Glanzel W	Braun T	Egghe L
Garfield E	1	0.7535	0.8359	0.5426
Glanzel W	0.7535	1	0.8916	0.7579
Braun T	0.8359	0.8916	1	0.5736
Egghe L	0.5426	0.7579	0.5736	1

Core author bibliographic-coupling matrix

	Garfield E	Glanzel W	Braun T	Egghe L
Garfield E	1	0.4249	0.3612	0.2881
Glanzel W	0.4249	1	0.9171	0.4069
Braun T	0.3612	0.9171	1	0.26
Egghe L	0.2881	0.4069	0.26	1

Core author direct citation matrix

	Garfield E	Glanzel W	Braun T	Egghe L
Garfield E	1	0.0022	0.0312	0.0027
Glanzel W	0.2844	1	0.414	0.1365
Braun T	0.2511	0.173	1	0.0073
Egghe L	0.0974	0.221	0.1058	1

Figure 1. Normalized matrixes of tripartite citation analysis.

The following five steps could help us realize author relationship mining based on tripartite citation analysis, such as "A→C, B→D, B→C". These steps can also be seen as an algorithm in relation mining.

First step: Obtaining the fundamental citation relationship with strength(>0) among core authors from original matrixes

Tripartite adjacency matrixes are transformed into corresponding adjacency lists. ACA list $\{L_{1i},Q_{1i}\}$ versus matrix $\{O_{1i}, P_{1j}\}$, and relational degree X_i (i stands for the ID of author pair) in list can replace X_{ij} (i/j stand for different authors in the matrix) . ABCA list $\{L_{2i},Q_{2i}\}$ versus matrix $\{O_{2i}, P_{2j}\}$, and relational degree Y_i versus Y_{ij}. ADCA list $\{L_{3i},Q_{3i}\}$ and $\{L_{3j},Q_{3j}\}$ versus matrix $\{O_{3i}, P_{3j}\}$, and relational degree Z_i and Z_j versus Z_{ij} (the order between i and j denotes the citing direction). We used the Adjacency list in calculation process.

Second step: Filtering no-explicit-relationship author pairs

The no-relationship author pairs (X_i=0, Y_i=0, Z_i=0, and no cooperation), are filtered as $\{O_{4i},P_{4j}\}$ in the Adjacency matrix, and $\{L_{4i},Q_{4i}\}$ in the Adjacency list, which forms the basic object in subsequent analysis.

Third step: Mining the relationship of A →C from$\{L_{1i},Q_{1i}\}\{L_{3i},Q_{3i}\}\{L_{4i},Q_{4i}\}$

Remark the $\{L_{4i},Q_{4i}\}$ as $\{A_k,C_k\}$ (k stands for the number of author pairs), the goal is finding the Dk with the relations $\{A_k→D_k, C_k\text{-}D_k\}$. We looked for the synchronous relations with strengh between A_k and D_k, C_k and D_k, from $\{L_{1i},Q_{1i}\}\{L_{3i},Q_{3i}\}$, and matched the author pairs in $\{A_k,C_k\}$. The pseudo code is as follows:

If one author in the pair of $\{A_k,C_k\}$= one author in a pair of $\{L_{1i},Q_{1i}\}$, and another one in the pair of $\{A_k,C_k\}$= one author in a pair of $\{L_{3i},Q_{3i}\}$, and another one in the pair of $\{L_{1i},Q_{1i}\}$= another one in the pair of $\{L_{3i},Q_{3i}\}$

Then mark the "one author in the pair of $\{A_k,C_k\}$" (so as the "one author in a pair of $\{L_{1i},Q_{1i}\}$") as Cα, the "one author in a pair of $\{L_{3i},Q_{3i}\}$" (so as the "another one in the pair of $\{A_k,C_k\}$") as Aα, the "another one in the pair of $\{L_{1i},Q_{1i}\}$" (so as the "another one in the pair of $\{L_{3i},Q_{3i}\}$") as Dα

End with the relation between A_α and C_α according to D_α, and their relation strength equaling to the product of X_α and Y_α. If the order of author pair in $\{L_{4\alpha},Q_{4\alpha}\}$(i.e., $\{A_k,C_k\}$) is in reverse of the order of author pair in $\{L_{3\alpha},Q_{3\alpha}\}$(i.e., $\{A_k,D_k\}$), then the relation strength between Aα and Cα will be the negative value.

Finally, choose the top value (Take the absolute value of the negative value) as the final relation strength of A_α and C_α.

Fourth step: Mining the relationship of B →D from$\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}\{L_{4i},Q_{4i}\}$

Remark the $\{L_{4i},Q_{4i}\}$ as $\{B_k,D_k\}$ (k stands for the number of author pairs), the goal is to find the A_k with the relations $\{A_k→D_k, A_k\text{-}B_k\}$. We looked for the synchronous relations with strengh between A_k and D_k, A_k and B_k, from $\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}$, and matched the author pairs in $\{A_k,C_k\}$.This process is similar with the process of A→C, so the pseudo code is omitted.

Fifth step: Mining the relationship of B →C from$\{L_{1i},Q_{1i}\}\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}\{L_{4i},Q_{4i}\}$

Remark the rest (no relationship like A→C and B→D) of $\{L_{4i},Q_{4i}\}$ as $\{B_k, C_k\}$ (k stands for the number of author pairs), the goal is to find the A_k and D_k with the relations $\{A_k→ D_k, A_k\text{-}B_k, C_k\text{-}D_k\}$. We looked for the synchronous relations with strengh between A_k and D_k, A_k and B_k, Ck and D_k, from $\{L_{1i},Q_{1i}\}\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}$, and matched the author pairs in $\{B_k,C_k\}$. The pseudo code as follows:

If one author in the pair of $\{B_k,C_k\}$= one author in a pair of $\{L_{2i},Q_{2i}\}$, and another one in the pair of $\{B_k,C_k\}$= one author in a pair of $\{L_{1i},Q_{1i}\}$, and another one in the pair of $\{L_{2i},Q_{2i}\}$=one author in the pair of $\{L_{3i},Q_{3i}\}$, and another one in the pair of $\{L_{1i},Q_{1i}\}$= another one in the pair of $\{L_{3i},Q_{3i}\}$

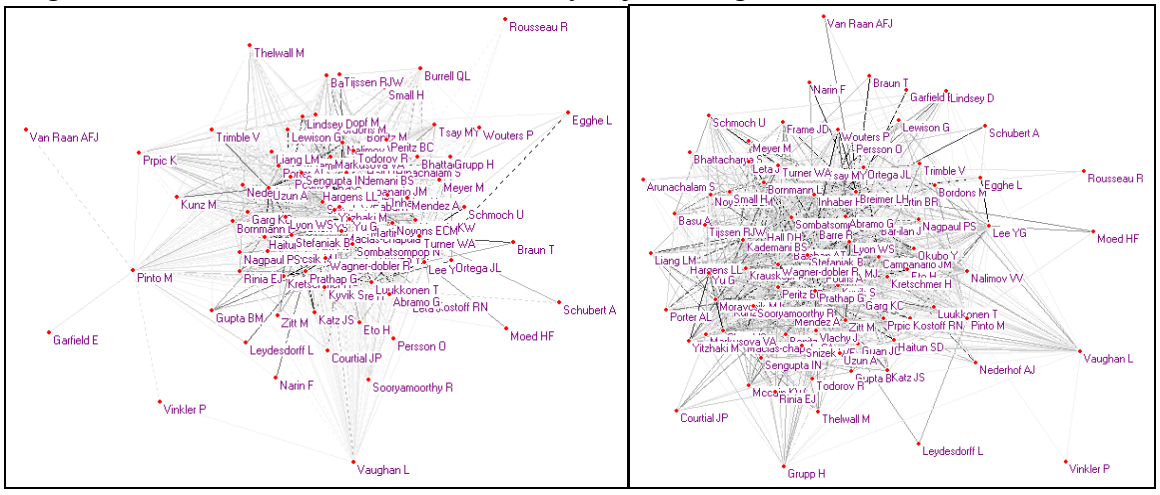
Then mark the "one author in the pair of $\{B_k,C_k\}$" (so as the "one author in a pair of $\{L_{2i},Q_{2i}\}$") as B_χ, "another one in the pair of $\{B_k,C_k\}$" (so as "the one author in a pair of $\{L_{1i},Q_{1i}\}$") as C_χ, one author in the pair of $\{L_{3i},Q_{3i}\}$(so as the "another one in the pair of $\{L_{2i},Q_{2i}\}$") as A_χ, another one in the pair of $\{L_{1i},Q_{1i}\}$(so as the "another one in the pair of $\{L_{3i},Q_{3i}\}$) as D_χ

End with the relation between B_χ and C_χ according to A_χ and D_χ, and their relation strength equaling to the product of X_χ and Y_χ and Z_χ. If the order of author pair in $\{L_{4\chi},Q_{4\chi}\}$(i.e., $\{B_k,C_k\}$) is in reverse of the order of author pair in $\{L_{3\chi},Q_{3\chi}\}$(i.e., $\{A_k,D_k\}$), then the relation strength between B_χ and C_χ will be the negative value.

Finally, choose the top value (take the absolute value of the negative value) as the final relation strength of B_χ and C_χ.

So far, all relationship among author pairs in $\{L_{4i},Q_{4i}\}$ have been built.

According to the above algorithm, potential relationships among not-directly-related core author set could be discovered by VBA programme and Access databases. The final results among A→C, B→D and B→C are visulized by Pajek as Figure 2 and 3.



(a) (b)

Figure 2. (a) Author relationship network of A→C. (b)Author relationship network of B→D.

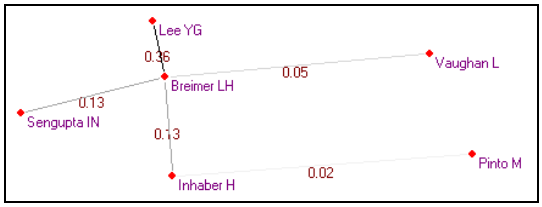


Figure 3. Author relationship network of B→C.

In Figure 3, the labels in the lines denote the value of the relationship similarity for authors in pairs. According to the results, there are different levels of potential relationship between Breimer LH and other authors, such as Inhaber H、Lee YG、Sengupta IN、Vaughan L.

Conclusions

Based on the algorithm design and empirical analysis, the deduction from results of ACA, ABCA and ADCA to potential author relationships mining could be probable, and the

empirical process would be practicable. The findings in Scientometrics field can help scholars discover more research fellows, which can promote scientific research cooperation and broader knowledge communication.

References

Boyack, K. W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389-2404.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers, *Journal of the American Society for Information Science and Technology*, *14*(1), 10-25.

Gómez-Núñez, A. J., Batagelj, V., Vargas-Quesada, B., Moya-Anegón, F., & Chinchilla-Rodríguez, Z. (2014). Optimizing SCImago Journal & Country Rank classification by community detection. *Journal of Informetrics*, *8*(2), 369-383.

Gómez-Núñez, A. J., Vargas-Quesada, B., de Moya-Anegón, F. & Glänzel, W.(2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, *89*(3), 741-758.

Gómez-Núñez, A. J., Vargas-Quesada, B. & Moya-Anegón, F. (2015). Updating the SCImago journal and country rank classification: A new approach using Ward's clustering and alternative combination of citation measures. *Journal of the Association for Information Science and Technology*, published online. http://dx.doi.org/10.1002/asi.23370.

Persson,O.(2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, 4(3), 415-422.

Schubert, A. (2002). The web of Scientometrics: A statistical overview of the first 50 volumes of the journal. Scientometrics, *53*(1), 3-20.

Shibata, N., Kajikawa, Y., Takeda, Y. & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, *28*(11), 758-775.

Small, H. (1973). Cocitation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265-269.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820-1833.

Wang, F. (2014). Influence Analysis of Core Authors in Scientometrics from an Integrated Perspective of Publication and Citation. *Science of Science and Management of S.&T.(China)*, *35*(12): 45-55.

Wang, F., Qiu, J. & Yu, H. (2012). Research on the cross-citation relationship of core authors in scientometrics. *Scientometrics*, *91*(3), 1011-1033.

White, H.D. & Griffith, B. (1981). Author cocitation: A literature mesaure of intellectual structures. *Journal of the American Society for Information Science*, *32*(3), 163-171.

Yang, S. & Wang, F. (2015). Visualizing Information Science: Author Direct Citation Analysis in China and around the World. *Journal of Informetrics*, *9*(1), 208-225.

Zhang, L., Glänzel, W. & Liang, L. (2009). Tracing the role of individual journals in a cross-citation network based on different indicators. *Scientometrics*, *81*(3), 821-838.