Journal of Informetrics 11 (2017) 810–822

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

# Combining multiple scholarly relationships with author cocitation analysis: A preliminary exploration on improving knowledge domain mappings

Yi Bu[a], Shaokang Ni[b], Win-bin Huang[b,*]

[a] School of Informatics and Computing, Indiana University, Bloomington, IN, USA
[b] Department of Information Management, Peking University, Beijing, China

ARTICLE INFO

*Article history:*
Received 14 December 2016
Received in revised form 17 June 2017
Accepted 17 June 2017

*Keywords:*
Author cocitation analysis
Coauthorship analysis
Author bibliographic coupling analysis
Scholarly network
Scientific intellectual structure
Knowledge domain mapping
Exploratory factor analysis (EFA)
Bibliometrics

ABSTRACT

Author cocitation analysis (ACA) is a branch of bibliometrics and knowledge representation that aims to map knowledge domains. However, ACA has been criticized because count-based measurement is too simple, and resulting maps are insufficiently informative. Since different scholarly relationships, e.g., coauthorship and author bibliographic coupling relationships, can extract out different relationships among authors in various perspectives, combining them with ACA for constructing knowledge domain mappings is our major purpose. The proposed method constructs the hybrid matrix from all relationships in four steps: relationship normalization, calculating the similarity between scholarly relationships, calculating adjustment parameters, and constructing hybrid relationships. The important parameters for integrating these matrices are calculated according to the distance in the hyperspace transformed from the similarity among the scholarly relationships by exploratory factor analysis. Compared with ACA, the results of the proposed method show: (1) More sub-fields in the given discipline can be identified when combining other scholarly relationships; (2) The more scholarly relationships added into ACA, the more details in terms of research area the method will find; (3) Good visualization in clustering is depicted when we combine other scholarly relationships. As a result, the proposed method offers a good choice to understand researchers and to map knowledge domains in a study field for integrating more scholarly relationships at the same time.



## 1. Introduction

As an important method in bibliometrics proposed by White and Griffith (1981), author cocitation analysis (ACA) aims to utilize cocitation relationships between author pairs to map the intellectual structures of knowledge and research domains (Jeong, Song, & Ding, 2014; Kim, Jeong, & Song, 2016). McCain (1990) summarized four main steps of ACA: 1) selecting and retrieving the author set; 2) constructing the raw cocitation matrix; 3) transforming the cocitation matrix into a correlation matrix; and 4) analyzing the data and interpreting the results. ACA works well independently of the specific application domains as evidenced by its popularity for mapping various scientific disciplines (e.g., Chen & Lien, 2011; Chu, Liu, & Tsai,

* Corresponding author at: Fanglibangqin Building 521, Department of Information Management, Peking University, 5 Yiheyuan Road, Haidian District, Beijing 100871, China.
*E-mail address:* huangwb@pku.edu.cn (W.-b. Huang).

http://dx.doi.org/10.1016/j.joi.2017.06.004
1751-1577/

2012). However, past research has noted defects of ACA, such as that its inputs only includes author cocitation count (frequency), which has been criticized for simply using count-based information (Bu, Liu, & Huang, 2016). Despite some improvements by integrating more general descriptive metadata (Bu et al., 2016; Zhao, 2006), ACA remains limited by the cocitation perspective, thus constraining its performance for mapping intellectual structures and knowledge domains.

According to Zhao (2012), there are four main research paradigms in knowledge domain mappings: the traditional paradigm (mainly using manual literature studies and review), the theory paradigm (based on sociology of science founded by Merton (1973)), the bibliometrics paradigm, and the social network analysis paradigm (primarily based on complex network/system theories and technologies). Regarding the "bibliometrics paradigm", researchers have used several scholarly relationship/network analysis methods, e.g., bibliographic couplings analysis (Boyack & Klavans, 2010; Kessler, 1963; Zhao & Strotmann, 2008a), citation analysis (Garfield, Sher, & Torpie, 1964), cocitation analysis (McCain, 1991; Small & Griffith, 1974; White & Griffith, 1981), coauthorship analysis (Beaver & Rosen, 1979), and co-word analysis (Callon, Courtial, & Turner, 1983), each of which provide different perspectives concerning scientific intellectual structures (Ma & Ni, 2012; Zhao & Strotmann, 2008a). The cocitation relationship is not the only way to show scientific intellectual structures, and combining different scholarly networks can provide broader visual thresholds. The method proposed in this article, which combines coauthorship and author bibliographic coupling analyses into ACA, thus aims to improve the performance of knowledge domain mapping. Note that the two terminologies, scholarly networks and scholarly relationships, are represented as matrices with the same meaning in this article.

The outline of this article is as follows. Related work is provided in Section 2. The point of scholarly relationship combinations is described in Section 3. The dataset used in this paper and the methods combining different scholarly relationships are proposed in Section 4. The results of the empirical studies and our observation are illustrated in Section 5. Finally, the conclusion is remarked in Section 6.

## 2. Related work

The basic assumptions of ACA are that each citation plays an equal role in cocitation analysis and that cocitation counts of the author pair are proportional to their relevance (White & Griffith, 1981). Thus, in ACA, two authors are connected if and only if they were cocited at least once, and the more they are cocited, the stronger cocitation relationship they will have. However, traditional ACA, founded by White and Griffith (1981) and standardized by McCain (1990), takes as input the cocitation counts of first authors, resulting in a small amount of useful information and thus negatively impacting the performance of visualization—two branches of study have emerged to explore this problem. The first branch lay in all-author cocitation analysis (AACA), pioneered by Persson (2001). Other scholars followed, classifying several kinds of ACA according to their methods of cocitation counting, such as first-author cocitation analysis (FACA), inclusive AACA, and exclusive AACA (Zhao, 2006). Zhao and Strotmann (2008b) as well as Eom (2008) found that AACA works better to capture all influential researchers in a field and can identify more sub-specialties than FACA. The other branch of research explored general descriptive metadata into ACA. For example, Bu et al. (2016) combined citation published time, citation published venue (e.g., journal, proceeding, etc.), and citation keywords to reveal more details and nuance in mapping knowledge domains.

A debate of how to transform raw cocitation matrix to correlation matrix in ACA is worth noting (Mêgnigbêto, 2013). Although researchers have adopted Pearson's $r$ in ACA since its birth, Ahlgren, Jarneving, and Rousseau (2004) provide a theoretical perspective and rigorous mathematical proof to argue that Pearson correlation coefficient's use in ACA has several drawbacks. Nevertheless, White (2004), a representative of the "Drexel team", found that although Pearson's $r$ might fluctuate in different experiments, clusters based on it "are much the same for the combined groups as for the separate groups" (p. 1250), but he emphasized that he had never disagreed to use other similarity measurements beyond Pearson's $r$ for ACA, such as cosine and Jaccard similarities. Such debates have been ongoing for more than ten years without resolve (Egghe & Leydesdorff, 2009; Van Eck & Waltman, 2008). Nevertheless, during this debate, bibliometricians have had a consensus that bibliometric analyses should be carried out at least using the theoretically most appropriate methods. In this paper, therefore, we are going to employ the cosine similarity to transform our raw matrix to correlation matrix instead of Pearson's $r$, because the method we propose contains author cocitation, author bibliographic coupling, and coauthorship frequencies and we do not aim to obtain a uni-scholarly-network probability like traditional ACA.

Bibliographic coupling analysis (BCA) reveals that the more similar the research topics of two articles, the more references they share, and was first proposed by Kessler (1963). Coupling strength (coupling frequency) in BCA is defined as the number of references two papers share (cocite). BCA is regarded as a static analysis because the coupling strength of coupled papers does not change. Zhao and Strotmann (2008a) proposed author bibliographic coupling analysis (ABCA) by using the author as the object of research and indicated that ABCA reveals the active researchers in a field and provides a perspective of the structure of the research front. Furthermore, an important reflection of Zhao and Strotman (2008a)'s study is that the coupling strength of the coupled authors in ABCA may change while the authors cocite more papers.

ABCA was used in combination with ACA from its inception. Zhao and Strotmann (2008a) argued that ABCA can be regarded as a complement of ACA and their combination will provide a more comprehensive view of the intellectual structures than either of them can provide on its own; specifically, ACA is considered as a look back in time (historical analysis) and ABCA as a view of the presence (current research front). They also pointed out that the extrapolation of major differences between ACA and ABCA could provide a forecasting method—a glimpse of the likely future development of the field under analysis. More recently, Zhao and Strotmann (2014) verified that their forecast made in 2008 had been remarkably good, and