



# Content- and proximity-based author co-citation analysis using citation sentences



Ha Jin Kim, Yoo Kyung Jeong, Min Song\*

Department of Library and Information Science, Yonsei University, 50 Yonsei-Ro, Seodaemun-Gu, Seoul, Republic of Korea

## ARTICLE INFO

### Article history:

Received 27 February 2016

Received in revised form 18 July 2016

Accepted 19 July 2016

### Keywords:

Author co-citation analysis

Citation proximity analysis

Citation content analysis

Bibliometrics

Citation analysis

## ABSTRACT

Author co-citation analysis (ACA) has been widely used for identifying the subject disciplines of authors. Citations can reveal the explicit relationship between authors as well as their subject research fields. However, previous studies have seldom considered citation contents that convey useful implicit information on the authors or the influence of the links between the authors' subject fields by taking citation locations into account. This study aims to reveal the implicit relationship in the authors' subject disciplines by considering both citation contents and proximity. To this end, the researchers propose a new ACA method, called content- and proximity-based author co-citation analysis (CPACA). For the study, we extracted citation sentences and locations from full-text articles in the oncology field. The top 15 journals on oncology in Journal Citation Reports were selected, and 6,360 full-text articles from PubMed Central were collected. The results show that the proposed method enables the identification of distinct sub-fields of authors to represent authors' subject relatedness.

© 2016 Elsevier Ltd. All rights reserved.

## 1. Introduction

Since Author Co-citation Analysis (ACA) was introduced by White and Griffith (1981), ACA has been widely adopted in bibliometric research to analyze the intellectual structures of academic fields. For its clear and simple method, compared with methods adopting interviews and surveys, ACA has been used by academic institutes or funding agencies as a tool for evaluating authors' scholarly activities (He & Hui, 2002). However, with easier access to citation databases provided by the Institute for Scientific Information's (ISI), the traditional ACA relied solely on simple citation counts to measure the author similarity that does not properly reflect the contribution of each author. As the advent of Web of Knowledge and Scopus, ACA researches received much attention and various author credit methods were proposed by modifying reference information (Boyack, Small, & Klavans, 2013; Zhao & Strtmann, 2011).

With the increase in the number of open-access journals and fully accessible databases of full-text articles, such as PubMed Central, recent studies have used various citation-related attributes, including cited location in an article and citation sentences, extracted from full-text papers, in addition to reference information (Gipp, 2006; Jeong, Song, & Ding, 2014; Zhao & Strotmann, 2014). Thus, author names are used along with the bibliometric information linked with the contents, such as author position, frequency of reference, and cited location of citation sentences.

Indeed, existing approaches rely heavily on citation counts, and do not consider the citation contents capturing the comprehensive similarity between authors. Hence, the present study aims to extend the traditional approaches by incorporating concepts of both "content" and "proximity" into ACA. In previous research using full-text papers, author's similarity was measured either among citation sentences at the document level (Jeong et al., 2014) or through their in-text citing location (Gipp, 2006). The present study introduces content- and proximity-based author co-citation analysis (CPACA) that uses citation sentences to capture the "contents" and constraint their in-text citing location at the section level to measure "proximity" of citation sentences. In addition, this study explores how CPACA compares with the traditional ACA approach in a field where the contents of citation sentences and their proximity play a special role. The current work focuses on the oncology research field as a case study for CPACA. Therefore, by considering content and proximity of citation sentences, we explore the following two research questions: (1) Can CPACA identify more sub-disciplines compared to traditional ACA approach? (2)

\* Corresponding author.

E-mail addresses: [hajin\\_228@yonsei.ac.kr](mailto:hajin_228@yonsei.ac.kr) (H.J. Kim), [yk.jeong@yonsei.ac.kr](mailto:yk.jeong@yonsei.ac.kr) (Y.K. Jeong), [min.song@yonsei.ac.kr](mailto:min.song@yonsei.ac.kr) (M. Song).

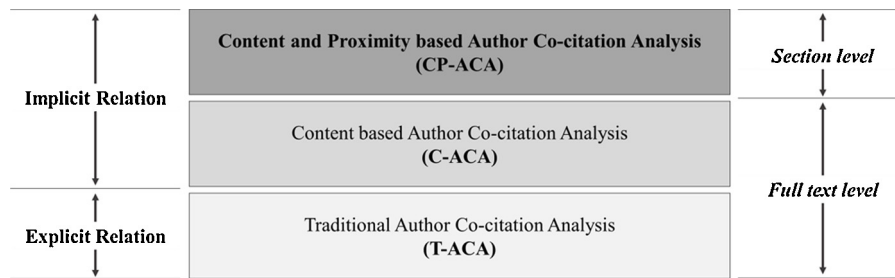


Fig. 1. Research design.

Can CPACA show subject relatedness among authors more noticeably than other approaches? These two questions are to be addressed in research design (Fig. 1) and to be discussed throughout the study.

## 2. Related works

### 2.1. Author co-citation analysis (ACA)

Co-citation analysis allows for different units of analysis, including documents, authors, or institutions (Zhao & Strotmann, 2011). Small (1973) and Marshoakova (1973) introduced document-based co-citation analysis, which relies on the fact that the more frequently two documents are cited together, the closer is the relationship between them. White and Griffith (1981) adapted this analysis into another unit of analysis at the authors' levels and proposed ACA, which is based on basic co-citation analysis. ACA considers the instances in which two authors are cited in the same document (Andrés, 2009; White & Griffith, 1981). ACA can identify the authors' related fields by counting the frequency of the two co-authors' oeuvres. This method can also imply that the more frequently author A and author B are cited in the same paper, the more similar their research fields are likely to be (White & Griffith, 1981).

Many ACA related studies have been performed since 1981. Most of them have attempted to clarify the research fields in perspective of how to measure the relationship between two co-cited authors and the impact of their contributions in the target research fields. White and Griffith (1981) first conducted ACA research by selecting 39 most influential researchers in the information science field. Several follow up studies were conducted according to the methods proposed by White and Griffith (1981). White and McCain (1998) demonstrated changes in the information science field at different points in time by slicing the period. They clarified the specific research fields of the authors, authors' affiliations, and paradigms over time. Ding, Chowdhury, and Foo (1999) conducted ACA in the information retrieval field. They analyzed 39 authors' intellectual structures over two time periods. Recently, Zhao and Strotmann (2008b) studied the frequency of co-cited authors in the information science field from 1996 to 2005 with 120 authors.

ACA has also been applied to various fields outside information science. To take into account researchers' contributions to the decision support systems field, Eom (1996) analyzed ACA in a collection of 944 articles and 23,768 references. He conducted factor analysis on 113 authors to identify the authors' subject areas. Andrews (2003) used ACA in the medical informatics field. He selected the top 50 authors with a high impact factor and then analyzed the field between 1994 and 1998. Applying cluster and factor analyses and multidimensional scaling, he demonstrated that ACA can help predict future research directions.

Previous studies counting co-cited authors have mainly paid attention to first authors. Recent works have introduced last author or all author co-citation analyses (Eom, 2008; Zhao & Strotmann, 2008b; Zhao & Strotmann, 2011; Zhao, 2006). Zhao (2006) and Zhao and Strotmann (2008b) compared the first author with all-author co-citation analysis. Eom (2008) noted the distinction between all-author and first-author citation analyses. The results revealed that all-author analysis was a more efficient way to identify authors' research fields compared with the first-author type.

Most previous studies have been limited by their focus on simple author co-citation frequency counts. These studies only counted the number of citations from reference metadata, and suggested that authors with a high co-citation frequency were related to one another and worked in related research fields. Consequently, only the explicit relation between two authors was examined, ignoring the contents of the citations. To tackle this limitation, Jeong et al., 2014 recently proposed a form of content-based ACA using text-mining techniques from full-text documents. Using *Journal of the Association for Information Science and Technology* to analyze full-text citation sentences content, they compared traditional ACA with content-based ACA. However, few studies have conducted citation content analyses at the full-text level, as the new concept of content-based ACA has only been introduced. The present study also takes into consideration the implicit relationship between two co-cited authors through the citation content. The goal is to identify the authors' research fields and the researchers' contributions from full-text documents, rather than simply considering author co-citation counts from a surface metadata level.

### 2.2. Citation proximity analysis (CPA)

Gipp (2006) introduced CPA in his doctoral thesis. CPA takes co-citation analysis into account but further exploits the citations' location within the full-text documents (Gipp, 2006, 2014). CPA assumes that the closer the citation sentences are to one another within the full-text, the closer the relationship between the two sentences (Gipp & Beel, 2009; Gipp, 2006; Gipp, 2014; Liu & Chen, 2011, 2012). CPA measures the degree of closeness between two citation sentences using the citation proximity index (CPI), which is calculated from the number of citations and their location (Gipp & Beel, 2009; Gipp, 2006, 2014). Gipp and Beel (2009) subsequently calculated CPI measures by counting pairs of co-cited references at four levels: sentence, paragraph, section, and article levels. When two citation sentences were located within the closest area in the article, they had the closest relation in the text. Further, Liu and Chen (2011) conducted a preliminary study of CPA using three different BMC journals, namely, *BMC Bioinformatics*, *BMC System Biology*, and *BMC Biology*, and compared the method

Download English Version:

<https://daneshyari.com/en/article/6934357>

Download Persian Version:

<https://daneshyari.com/article/6934357>

[Daneshyari.com](https://daneshyari.com)