



Content-based author co-citation analysis



Yoo Kyung Jeong^a, Min Song^{a,*}, Ying Ding^b

^a Department of Library and Information Science, Yonsei University, Seoul, Republic of Korea

^b Department of Information and Library Science, Indiana University, Bloomington, IN, USA

ARTICLE INFO

Article history:

Received 20 August 2013

Received in revised form 27 October 2013

Accepted 2 December 2013

Available online 21 December 2013

Keywords:

Author co-citation analysis

Citation content analysis

Bibliometrics

Information science

Citation analysis

ABSTRACT

Author co-citation analysis (ACA) has long been used as an effective method for identifying the intellectual structure of a research domain, but it relies on simple co-citation counting, which does not take the citation content into consideration. The present study proposes a new method for measuring the similarity between co-cited authors by considering author's citation content. We collected the full-text journal articles in the information science domain and extracted the citing sentences to calculate their similarity distances. We compared our method with traditional ACA and found out that our approach, while displaying a similar intellectual structure for the information science domain as the other baseline methods, also provides more details about the sub-disciplines in the domain than with traditional ACA.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Since author co-citation analysis (ACA) was introduced in 1981 by White and Griffith, it has been a key method used in bibliometrics research. ACA is used to identify, trace, and visualize the intellectual structure of an academic discipline by counting the frequency with which any work of an author is co-cited with another author in the references of citing documents (Bayer, Smart, & McLaughlin, 1990). The primary goal of ACA is to identify the intellectual structure of a scientific knowledge domain in terms of the groupings formed by accumulated co-citation trails in the scientific literature. The traditional ACA process constitutes roughly the following six steps (McCain, 1990): (1) select authors, (2) retrieve co-citation frequencies, (3) compile a raw citation matrix, (4) convert it to the correlation matrix, (5) apply multivariate analysis of the correlation matrix, and (6) interpret and validate the results.

However, existing ACA approaches do not focus on identifying the intellectual structure of a target domain based on the citing content of the cited paper. They equally weight all citations without considering the variation of citing content. For example, following two sentences are cited in the same paper (White, 2003a, 2003b), but the purpose and the location of citing are different from each other.

- “The first goal of ACA mapping is to epitomize a field of learning through meaningful arrangements of its key authors’ names (White & McCain, 1989).”
- “The Kamada–Kawai spring embedder in Pajek placed the nodes freely from a circular starting position (Kamada & Kawai, 1989).”

* Corresponding author. Tel.: +82 221232405.

E-mail addresses: yk.jeong@yonsei.ac.kr (Y.K. Jeong), min.song@yonsei.ac.kr (M. Song), dingying@indiana.edu (Y. Ding).

The first sentence is to explain the general purpose of the ACA in the introduction section, and the second sentence is in the methodology section to describe the usage of a layout algorithm for network visualization. Though both citation sentences are located in the same paper, the citing purpose and content in the paper are different from each other.

In this paper, we further extend the current author co-citation analysis method by incorporating citing sentence similarity into citation counts. We use citing sentences to obtain the topical relatedness between the cited authors instead of traditional author co-citation frequency, and citing sentence similarity is measured by topical relatedness between two citing sentences. The basic assumption of this study is that citations should be assigned different weights under different contents. Since sentences in the full-text can describe the subject of an article at a more fine-grained level, using a sentence as a unit of analysis can be used to reveal a specific latent structure of a discipline. We present a bottom-up approach to ACA by mining full-text journal articles.

This paper was organized as follows. Section 1 introduces the topic. Section 2 outlines related works, Section 3 presents the proposed methods. Section 4 analyzes the results and discusses the impact. Section 5 concludes the article by pinpointing the limitations and future research.

2. Related works

2.1. Author co-citation analysis

A citation reflects an author is influenced by the work of another author, but usually does not explicitly indicate the strength or direction of that influence. Conventionally, it is assumed that each reference makes equal contribution to the citing article. In Small's study (1973) which first introduced co-citation analysis, the document co-citation analysis (DCA) quantifies the relationship between co-cited documents with the assumption that more frequently co-cited documents exhibit greater co-citation strength. After Small's study, White and his colleagues analyzed and mapped the information science domain using author co-citation analysis (White & Griffith, 1980; White & McCain, 1998). Furthermore, White (2003b) adopted a new network algorithm Pathfinder Networks (PFNETs), and demonstrated that PFNETs gave an advantage for ACA over other techniques in terms of computational cost. ACA methods have been widely applied to many domains including information retrieval, international management, strategic management, and e-learning. (Acedo & Casillas, 2005; Chen & Lien, 2011; Ding, Chowdhury, & Foo, 1999; Ma, Dai, Ni, & Li, 2009; Nerur, Rasheed, & Natarajan, 2008; Zhao & Strotmann, 2011).

Some researches focused on the advancement of methodology for ACA. He and Hui (2002) proposed a mining process to automate ACA based on the Web Citation Database. Their mining process used the mining technique, agglomerative hierarchical clustering (AHC), for author clustering and multidimensional scaling (MDS) for displaying author cluster maps. Chen, Ibekwe-Sanjuan, and How (2010) introduced a multiple-perspective co-citation analysis method for characterizing and interpreting the structure and dynamics of co-citation clusters. The multiple-perspective method integrates network visualization, spectral clustering, automatic cluster labeling, and text summarization.

While most studies have applied the general steps and techniques of classic ACA to different research domains with minor or no modifications, some studies have proposed new techniques to map author clusters (White, 2003b) or to process co-citation counts statistically (Ahlgren, Jarneving, & Rousseau, 2003). Persson (2001) attempted to compare first-author and all-author co-citation analysis with a small set of Web of Science citation data. Zhao and Logan (2002) suggested that all-author co-citation is a better measure of the connectedness between authors than first-author co-citation. Zhao (2006) compared the results of two different types of co-citation counting: first author co-citation versus all author co-citation with full-text articles in the field of XML. Schneider, Larsen, and Ingwersen (2009) also compared the first and all-author co-citation counting, and proposed the new matrix generation approach by extracting all-author information from a corpus of full text XML documents. Eom (2008) compared the differences of first-author counting and all-author counting for ACA to capture all influential researchers in a field. Recently, Zhao and Strotmann (2011) introduced last-author citation counting and compared it with traditional first-author counting and all-author counting.

2.2. Citation content analysis

While traditional ACA focused on quantitative measures, a few less prevalent studies investigated the citation content. Tradition citation analysis is mainly quantitative (e.g. citation frequency) and pays less attention to the actual content, while classical content analysis (CA) is essentially qualitative (e.g. codebook categories) and rarely applied to citation data.

MacRoberts and MacRoberts (1984) dissected negative citations and concluded they are usually disguised as perfunctory citations or citations combined with a positive description of the same work. Giles et al. (1998) used citing context for enhancing bibliographic records which later led to CiteSeer. The methodology developed to extract and represent citing sentences in CiteSeer is complex and requires a significant computational effort. While it performs with decent accuracy as an online tool, CiteSeer does not perform any analysis of citing sentences; it simply provides them to the user, allowing them make their own inferences about the nature of the citation. On the other hand, McCain and Salvucci (2006) did citation content analysis to understand "the diffusion of ideas in scholarly communication." He et al. (2010) built a context-aware citation recommendation system. This system not only recommended citation related papers, but matched the recommendations to specific parts of the paper under analysis.

Download English Version:

<https://daneshyari.com/en/article/523973>

Download Persian Version:

<https://daneshyari.com/article/523973>

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