

Content-based Author Co-citation Analysis

Yoo Kyung Jeong, Min Song*, Ying Ding

Yoo Kyung Jeong, Dept. of Library and Information Science, Yonsei University, Seoul, Republic of Korea, yk.jeong@yonsei.ac.kr

Min Song, Department of Library and Information Science, Yonsei University, min.song@yonsei.ac.kr

Ying Ding, Department of Information and Library Science, Indiana University, Bloomington, IN, USA. dingying@indiana.edu

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.

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 et al., 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, 2003), 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).”

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 an 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 and Griffith, 1980; White and 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; Ding et al. 1999; Ma et al. 2009; Nerur et al. 2008; Zhao & Strotmann 2011; Chen & Lien 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 et al. (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 et al. (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.

MacRobert and MacRobert (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.

Citation content also has been proposed as a useful construct for classifying articles and automatically generating abstracts or summaries of articles (Callahan et al. 2010). Nanba and Okumura (1999) used reference areas (single or multiple sentence sections related to a citation) from multiple papers to generate article summaries, extract the relationships described between papers, and classify reference types or reasons for a citation. Nanba and Okumura (2005) used the same method to identify survey or review articles, which contain a high proportion of citations that are considered influential or related papers in a field. Nakov et al. (2004) coined the term *citances* to refer to citing sentences, or the sentences that contain a citation, and their research indicated a number of functions *citances* can serve in citation analysis. Teufel (2001) and Ritchie et al. (2006) demonstrated the use of text windows (comparable to Nanba and Okumura’s references area) to assign index terms to articles and generate summaries of articles that establish the relevance of an article to a subject area almost as effectively as reading the entire article. Elkiss et al. (2008) introduced the application of collaborative citation summaries, the set of all sentences that cite a document, and also considered the role of co-citation in citation summaries. The most significant finding of their study for contextual co-citation is that “papers co-cited in the same sentence tended to be more similar than papers co-cited in the same paragraph.” Small (2010) used cue words extracted from the citing context, the text surrounding references, to identify interdisciplinary links. The results showed that the citation contexts play a crucial role in interpreting interdisciplinary links, and they are associated with the structure of the scientific map (Small, 2011). In recent years, Zhang et al. (2012) proposed a new framework for Citation Content Analysis (CCA) for syntactic and semantic analysis of citation content that can be used to better analyze the rich socio-

cultural context of research behavior. They proposed only procedures but did not report experiments.

2.3. Citation Proximity Analysis (CPA)

The growing interests in citation content extended to the study of citation location in the article (e.g. same section or paragraph). In Citation Proximity Analysis (CPA), the proximity of citations in full-text is used to calculate the Citation Proximity Index: the proximity among pairs of citations is examined with the assumption that the strength of contextual co-citation in the same sentence is greater than contextual co-citation in the same section of an article (Gipp & Beel, 2009).

Elkiss et al. (2008) found that papers co-cited “within the same sections, paragraphs, or sentences are more similar to each other than papers co-cited at the article level.” Gipp and Beel (2009) proposed a measure of citation proximity analysis to identify related work. They demonstrated the utility of citation proximity for identifying potentially related works. Callahan and Hockema (2010) introduced contextual co-citation analysis that is very similar to that of Gipp and Beel’s (2009) study. The main difference between the two studies is that Callahan and Hockema (2010) used a discrete set of fixed values to quantify proximity among citations while Gipp and Beel (2009) used the values that reflect the structural complexity of the citing document. More recently, Liu and Chen (2011) investigated “the effects of co-citation proximity on the quality of co-citation analysis” through experiments of co-citation instances found in full-text scientific publications. The results showed that sentence-level co-citation preserves the structure of the traditional co-citation network and forms a smaller subset of the entire co-citation. Recent research by Boyack et al. (2013) used normalized proximity for improving the accuracy of co-citation clustering. They compared the results of the traditional co-citation clustering using only proximity between reference pairs with their method and reported their approach increased the textual coherence and clustering accuracy. Lu and Wolfram (2012) presented static and dynamic word-based approaches using vector space model as well as a topic based approach based on Latent Dirichlet Allocation (LDA) for mapping author research relatedness.

The major difference in our approach compared to other studies lies in our similarity measure between cited authors. Previous studies use a single measurement such as quantitative (co-citation frequency and proximity) or qualitative content. Since the citing sentences give us a better idea of why a paper is cited in content, our approach incorporates rich contents of citation content into citation analysis using citing sentences, as opposed to using citation frequencies for the similarity measure. We assert that our approach reflects the sub-structure in domain analysis at a more granular level than provided with other traditional ACA approaches.

3. Methodology

In this section, we provide details of a novel co-citation analysis using citing sentences. Citing sentences in a scientific article may contain information about the cited research and cited authors' research area. Previous work has shown the importance of citing sentences in scientific domains to be used for quantitative analysis of textual relationships with the potential applications in summarization and information retrieval (Elkiss et al., 2008; Nakov et al., 2004). As far as we know, this is the first research that utilizes the contents of citing sentences in determining the relationship between the authors referenced. Citing sentences could be an important factor in ACA. Therefore, our approach of incorporating citing sentences into co-cited relations (co-citation strength) among authors in ACA is a novel extension to the traditional ACA. Whereas traditional ACA uses co-citation frequency for author counting, our approach uses the citing sentence similarity to consider topical relatedness (citation content). To evaluate our new method, we applied it to the information science domain from the point of view of comparison, which has been widely studied (White & McCain, 1998; Zhao & Strotmann, 2008; Chen et al., 2010). One of the important tasks for this analysis is to extract the citing sentences. For our analysis, we chose the *Journal of the American Society for Information Science and Technology (JASIST)*, which is considered the most prominent journal in information science, and the full-text papers containing citation sentences were collected. To extract the citing sentences in a full-text article, we made use of the specific

APA citation style, adopted in *JASIST* and specified in the html syntax. For factor analysis and other preprocessing tasks, we used R (<http://www.r-project.org>) as the statistical programming language and Gephi (Bastian et al., 2009) as the visualization tool.

3.1. Data Collection

The datasets used in this study were gathered from 1,420 full-text articles in the *JASIST* digital library and Web of Science over 10 years (from January 2003 to June 2012). The datasets collected consist of two data categories, the full-text data and bibliographic metadata. The 1,420 collected documents have 60,068 references. The number of collected articles was 1,436. However, 16 articles did not include citation links in the html file and had different citation styles like a square bracket style. These were excluded from the analysis. We extracted the following elements from the full-text article to form a citation index: 1) cited authors (in author order) in the cited document, 2) the title of the cited articles or books, and 3) citing sentences.

3.2. Extending the Traditional Procedure

Since our approach is based on extracting citation information from citing sentences found in full-text articles, we need to extend the traditional ACA procedure. Fig. 1 outlines this extension of ACA.

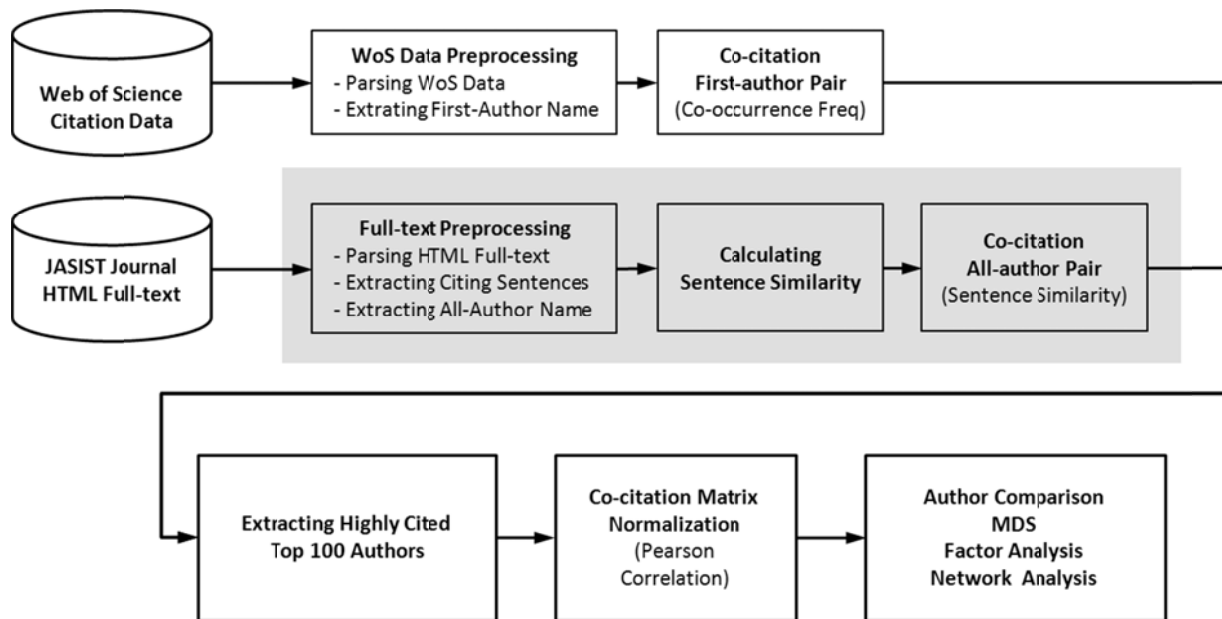


Fig. 1. Experimental overview.

Our experiment consists of four parts: data gathering, data preprocessing, generating the co-citation matrix by Pearson’s correlation, and data analyses using multivariate analyses and network analysis. We implemented the traditional author counting approach in order to compare our sentence similarity results. For the first-author counting method we retrieve the citation data from the Web of Science.

In the data preprocessing stage, we extract the citing sentences from journal papers in html format. By relying on html syntax, we avoid some of the errors encountered by other automatic citation indexing techniques (e.g., Giles, Bollacker, & Lawrence, 1998). Giles and his colleagues’ approach to extracting reference data from PDF files like those of Zhao (2006) had to deal with many problems of segmentation and disambiguation of data from the raw PDF files. To recognize the citing sentences in the html full-text article, we parse the sentences of the full-text documents then we detect the citation styles of *JASIST* citation in the sentences using regular expressions.

Fig. 2 shows a paragraph in a *JASIST* paper containing 2 citing sentences. The first sentence contains references to 12 papers that are hyperlinked to the reference section containing anchor tags (<a>

using href attributes. It is from these tags, author names are extracted.


Introduction Jump to...

At the heart of relevance lies topical relevance. Although topical relevance is widely recognized as the most important factor in selecting information, our understanding of the notion is limited and merits further explication building on work by Cooper (1971), Wilson (1973, 1978), Saracevic (1975), Rees and Saracevic (1966), Green and Bean (1995), Bean and Green (2001), Huang and Soergel (2004, 2006), Huang and White (2005), and Huang (2009a, 2009b). The growth of interest in "user relevance" has shifted attention away from topical relevance (Hjørland, 2010). Topicality tends to be treated as a primitive (undefined, self-explicating) concept and is rarely discussed in depth. With the exceptions cited above, the underlying mechanism of topical relevance is taken for granted, treated as a black box. A main purpose of this article is to further open the black box and shed light on relevance in general and topicality in particular.

References Jump to...

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology*, 61(2), 217–317.

Hjørland, B. (2002). Epistemology and the socio-cognitive perspective in information science. *Journal of the American Society for Information Science and Technology*, 53(4), 257–270.

 [Abstract](#) | [Full Article \(HTML\)](#) | [PDF \(138K\)](#) | [References](#) | [Web of Science@Times Cited: 82](#)

Huang, X., & Soergel, D. (2004). Relevance judges' understanding of topical relevance types: An explication of an enriched concept of topical relevance. In Proceedings of the 67th Annual Meeting of the American Society for Information Science and Technology (ASIS&T 2004). Medford, NJ: Information Today.

Huang, X., & Soergel, D. (2006). An evidence perspective on topical relevance types & its implications for task-based retrieval. *Information Research*, 12(1).

Huang, X., & White, R.W. (2005). Policy capturing models for multi-faceted relevance judgments. In Proceedings of the 68th Annual Meeting of the American Society for Information Science and Technology (ASIS&T 2005). Medford, NJ: Information Today.

Huang, X. (2009a). Topical relevance, rhetoric, and argumentation: A cross-disciplinary inquiry into patterns of thinking and information structuring. Unpublished doctoral dissertation, University of Maryland, College Park, MD.

Huang, X. (2009b). Developing a cross-disciplinary typology of topical relevance relationships as the basis for topic-oriented information architecture. In Proceedings of the ASIST SIG-CR (Special Interest Group on Classification Research) 20th Workshop.

Fig. 2. The html links with citing sentences in JASIST papers.

In traditional ACA, Salton's cosine similarity, Pearson's r correlation coefficient, Jaccard index, or direct co-citation counts are used to quantify the relationship between two documents. These statistical measures are based on binary values (Shneider & Borland, 2007) such that a co-occurrence is counted as 1, no matter where in the document the two citations occur together. However, when taking into account the content of citation, using binary values to represent the presence of co-citation is not sufficient (Callahan et al., 2010). Co-citation strength is a concept that has existed since co-citation analysis was introduced (Small, 1973). However, it has been previously used in a very limited manner as a function of the proximity of co-citations. Our intention of a non-binary strength is similar to Callahan's (Callahan et

al., 2010) approach, but we use the citing sentence similarity instead of using the proximity function. To calculate the similarities between citing sentences that are cited in the same paper, we remove stop words and stem sentences using SnowballStemmer (Hornik, 2007), and use the cosine similarity measure to calculate the sentence similarity because of its prevalent usage.

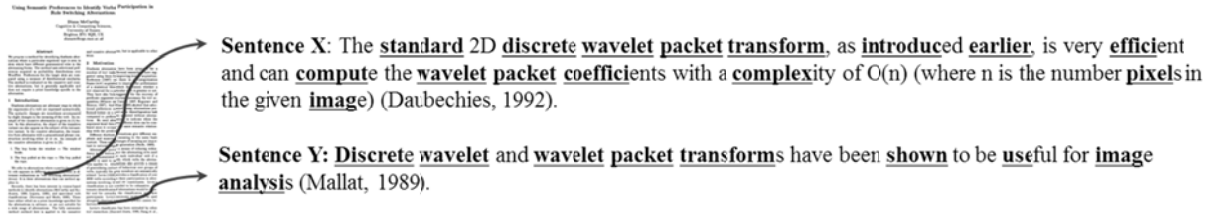


Fig. 3. Examples of citing sentences

As shown in Fig. 3, the sentences cited in the same article are extracted, and the cosine similarities (Equation 1) among sentences are calculated. In the traditional approach of calculating the co-citation similarity, any author pair is counted as 1. But in our approach, the author pair is weighted by the similarity of sentences that these two authors were cited in the full-text article.

$$Sim(\vec{x} \cdot \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|} = \frac{\sum_{i=1} x_i y_i}{\sqrt{\sum_{i=1} x_i^2 \sum_{i=1} y_i^2}} \dots (1)$$

For instance, the co-citation similarity of the author pair, Daubechies (in Sentence X) and Mallat (in Sentence Y), is 1 in the traditional approach whereas in our approach, it is 0.623.

- Author pair (Daubechies, Mallat) co-citation frequency=1 (traditional approach)
- Author pair (Daubechies, Mallat) co-citation cosine similarity = 0.623 (sentence similarity)

The content similarity of the sentence vectors, 0.623, is calculated by cosine similarity (Equation 1).

	analysi	complex	discret	effici	introduc	pixel	standard	use	coeffici	comput	earlier	imag	packet	shown	transform	wavelet
X	0	1	1	1	1	1	1	1	1	2	1	0	1	1	0	2
Y	1	0	0	0	1	0	0	1	0	1	0	1	0	1	1	2

Fig. 4. An example of sentence vectors.

As shown in Fig. 4, the similarity of 0.623 is based on the sentence vectors converted from the two example sentences in Fig. 3 after stemming and removing stop words (see the bold text in Fig. 3). In the case of no commonly shared words in two citing sentences, the similarity value is 0. For authors who are cited in the same paper, if there are no commonly shared words in the citing sentences, they are not treated as being co-cited. Thus, with our approach, high similarity values are assigned to authors who share similar citing sentences, which is different from existing approaches. In the situation where the same author is cited multiple times in the paper, it produces multiple values of the sentence similarity of the author, and we assign the maximum value to the similarity. The rationale for this is to compare with citation frequency-based values that are counted 1 as the default adopted by traditional ACA and to identify which citation in reference has the highest impact on the citing paper. After all, we accumulate the similarity values which are assigned to the pair of authors.

On the other hand, the similarity between the authors cited in the same sentence is 1 if all the words contained in the same sentence are equally applied to all authors in the sentence. Fig. 5 shows the distribution of citing sentence similarity values over the collected data. Although authors are co-cited in the same paper, their content similarity is low. Most of sentence similarity values are below 0.5 (91.4%). The proportion of the sentence similarity for authors who are cited in the same sentence is 7.9% (similarity value is 1).

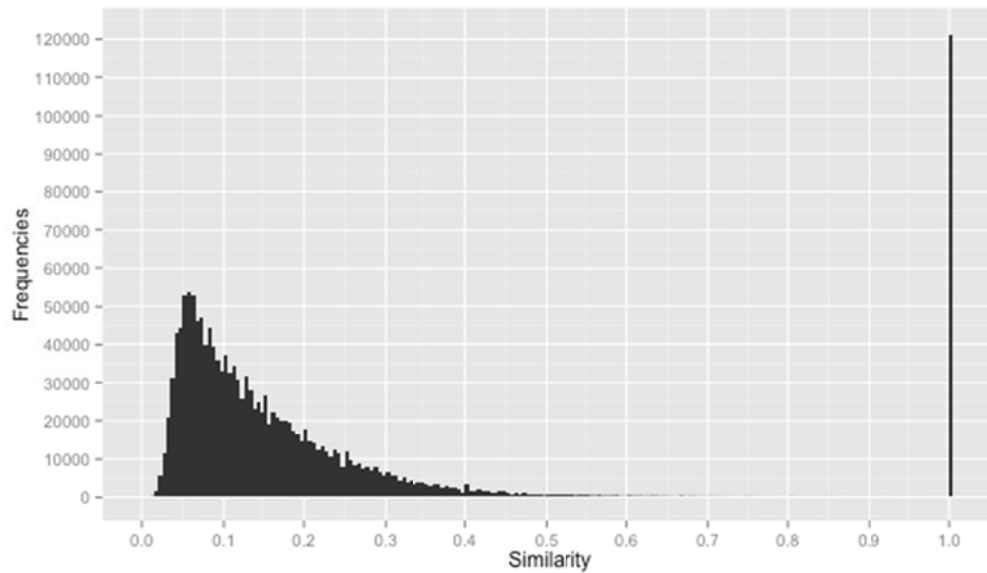


Fig. 5. Distribution of the content-based ACA values.

The sentence similarity value does have the advantage of showing two citing sentences may share the topicality. Since we are interested in the impact of various ACA methods on mapping the knowledge structure of a given field, we investigate differences in traditional ACA and our content-based ACA methods. Once, we construct matrices of co-cited authors, the next step is to represent the co-citation matrices as network graphs. To identify specialties and scholarly communities in terms of co-citation sub-disciplines, we use multivariate analysis -factor analysis.

To identify the important sub-disciplines and authors, we select the highly cited top 100 authors from our data set and disambiguate author names manually. Since there are no strict rules regarding thresholds for citation-based author selection in ACA studies (McCain, 1990), the present study selected 100 authors to be included since the results of White and McCain (1998) reveal that 100 authors are the most canonical authors among the submitted 120 authors during long periods in the information science domain and Zhao (2006) also selected the top 100 most cited authors for multivariate analysis. While the threshold selection of the top authors is still arbitrary, we've decided to the same threshold with previous studies to reduce variability in the results.

To compare traditional ACA and content-based ACA, we chose the same set of 100 authors. Since only 63 authors appeared in top 100 ranks in both traditional ACA and content-based ACA, we merged the entire set of authors and selected top 100 authors by citation frequency.

The author co-cited similarity matrix is converted to the author correlation matrix (Pearson's correlation matrix) for multivariate analysis to identify the sub-disciplines. The reason for using Pearson's correlation matrix is two-folds: 1) many ACA studies have used it (White, 2003a; Zhao & Strotmann 2007), and 2) we want to use normalized matrix as input for factor analysis. We also ran the cosine normalization method used in Ahlgren et al.'s study (Ahlgren, Jarneving, & Rousseau, 2003) and found the results to be very similar to the results gained by Pearson correlation matrix. Finally, as in the traditional ACA, we conduct factor analysis. For network analysis, we visualize the co-citation matrix using a line (edge) connecting two items (author as nodes) represents a co-citation link. The thickness of a line is proportional to the strength of co-citation. The color of the nodes represents the topic areas that are the results of modularity (Newman, 2006) which is used to measure how modular a network is.

4. Results and Discussion

4.1. Basic Statistics

Table 1 shows the basic statistics of extracted authors. The total numbers of distinct authors in our dataset is 22,913 and 32,095, respectively. The traditional ACA counting method uses only the first author of references and the content-based ACA considers all-authors counting. The content-based ACA excludes the author that is only cited in references and not in the body of the full-text. These authors, the very small proportion, are not detected during the process of extracting citing sentences, and the value of the sentence similarity is 0.

Table 1 Comparison of the number of author.

	Traditional ACA	Content-based ACA
The number of authors	22,913	32,095
The number of pairs	1,032,828	1,376,116

Table 2 lists the top 10 scientists drawn from these three methods in information science. These highly cited authors define two main research fields of *JASIST*, information retrieval (defined by Gerard Salton and Amanda Spink) and bibliometrics (defined by Eugene Garfield and Loet Leydesdorff). Citation frequencies of the authors in Table 2 are at least more than 100 times and are used to rank the authors in the Table. Of the top 10 authors, 5 authors (shown in bold) are found in both ACA methods but are not listed in the same priority order. One interesting observation is that traditional ACA counting identifies only well-known authors in traditional areas of Library and Information Science. Overall, citation frequencies in the all-author counting approach in content-based ACA are bigger than in traditional ACA. However, the authors at 1st and 2nd rank (Salton and Garfield) in content-based ACA have lower citation frequencies than ones in traditional ACA. This is caused by the authors who do not have the sentence similarity values. In fact, the citation frequencies by the simple all-author count method are 183 for Salton and 173 for Spink.

Table 2 Top 10 authors based on citation frequency.

Rank	Traditional ACA	Content-based ACA
1	G. Salton (178)	G. Salton (156)
2	E. Garfield (157)	A. Spink (147)
3	A. Spink (133)	R. Rousseau (145)
4	B. Cronin (126)	T. Saracevic (142)
5	L. Leydesdorff (126)	E. Garfield (139)
6	M. J. Bates (123)	P. Ingwersen (130)
7	L. Egghe (118)	H. Chen (124)
8	T. Saracevic (110)	L. Leydesdorff (124)
9	H. D. White (110)	A.F.J Van Raan (113)
10	N. J. Belkin (108)	H.F. Moed (110)

We selected the top 100 highly cited authors and built raw co-citation frequencies matrices. The resulting matrices were converted to Pearson's r correlation matrices that were in turn used as input to factor analysis.

4.2. Factor Analysis

Factor analysis is widely used in ACA to find latent structures buried in the mapping results. Traditionally, major factors in ACA are interpreted as research specialties (White & McCain, 1998). Zhao and Strotmann (2008) identified 11 specialties based on 120 most-cited authors in 2001-2005. They manually labeled these specialties by examining each specialty's member. Determining the number of specialties is a key issue in a co-citation analysis (Chen et al., 2010). In factor analysis, an oblique rotation was chosen because it is often more appropriate than an orthogonal rotation when it is expected theoretically that the resulting factors would in reality be correlated (Hair, et al., 1998), and the eigenvector is used for selecting the number of factors. If the eigenvalue is 1 or greater, the eigenvector is treated as common, which is known as the Kaiser rule (Kaiser, 1960).

Table 3 presents the factor analyses extracted from traditional ACA and content-based ACA matrices along with their model fits. It shows that the model fit is very good in both ACA methods, i.e., each account for over 81% of the variations in the highly aggregated author co-citation matrix. For example, a factor analysis of the traditional ACA co-citation matrix resulted in a 10-factor model, which explains 81.5% of total variance.

Table 3 The Result of factor analysis.

Input co-citation matrix	# of factors	Total variance explained
Traditional ACA	10	0.815
Content-based ACA	22	0.884

Content-based ACA produces a slightly better model fit than others when it uses many factors. It is apparent that there is a huge difference between the factor solutions within these two ACA methods. The results of traditional and content-based ACA methods are similar to a study done by Zhao (2006) which observed that a smaller set of factors explained the majority of variance in the dataset, and supported the assumption that the latent structures in the dataset are more explicable and visible. In our study, the factor solution shows similar results as those of Zhao's for the traditional ACA method and reinforces Zhao's findings through the content-based ACA results.

In the traditional ACA, 10 factors explain 81.5% of variance. In the content-based ACA method, 22 factors explain 88.4% of variance in the matrices and first 10 factors explain about 60.2% of variance. Although the factor structure is relatively weak with less total variance explained, this result can be interpreted as the information science domain consists of a sub-disciplines and there are more factors that influence the factor structure than the first 10 factors.

To identify more specific sub-disciplines as specialties in the research field, we use factor labeling through extracting keywords from the authors' paper titles. To extract keywords that represent factors, we collected paper titles of the top 100 authors from the reference list and extracted keywords that frequently appear in titles. These keywords were used to determine which factor is related to which subfield. The words such as "science", "research", "analysis", "approach", "information", and "study" are removed to better detect specialties.

Table 4 Factor labels, number of authors, highest loadings.

FL Number	Factor	First-author counting		Sentence similarity	
		No. of authors	Highest loading	No. of authors	Highest loading
F1	Information retrieval	34	1.001		
F2	Information seeking behavior			16	1.202
F3	Interactive information retrieval, relevance feedback			6	1.222
F4	Language model, query, clustering	13	1.040	10	0.967
F5	Classification algorithms			2	1.060
F6	Information seeking behavior, contextual approach			4	1.022
F7	Text mining, machine learning	6	1.016	4	0.882
F8	User interface	3	0.731	2	0.832
F9	User acceptance of information technology			1	1.108
F10	Digital library, data mining			1	0.823
F11	Information retrieval, information usage			4	0.832
F12	Information systems			1	0.987
F13	Electronic Journal, open access			3	0.916
F14	Domain analysis, information behavior			2	0.943
F15	Multimedia information retrieval			1	0.703
F16	Bibliometrics				
F17	Evaluation indicator, index	15	1.102	12	1.058
F18	Webometrics	9	1.070	6	1.057
F19	Visualization, mapping			5	1.012
F20	Journal indicator, evaluation			7	0.767
F21	Scholarly communication	2	0.897	5	0.934
F22	Journal citation analysis, interdisciplinarity	15	1.025	4	0.917
F23	Network analysis	2	0.788	3	0.963
F24	Bioinformatics	1	0.674	1	1.110

Table 4 shows the 24 different factor labels found in the papers of the top 100 authors. It analyzed for each ACA method how many authors were associated with a label and what was the highest loading. Three coarse-level sub-disciplines revealed by factor analysis for the information science domain are information retrieval (F1), bibliometrics (F16), and bioinformatics (F24), which are displayed in Table 4 in order of relevancy based on our data. Information retrieval and bibliometrics are the major sub-disciplines and are drawn from the traditional ACA factor labels (see bold rows in Table 4). In these sub-disciplines, traditional and content-based ACAs identified 56 authors and 57 authors for information retrieval respectively. Traditional ACA methods identified 43 authors in the bibliometrics sub-discipline and 42 authors in content-based ACA. There is an author who moved from the different top-level sub-discipline by the content-based ACA method than in the other method. Ronald E. Rice belongs to the factor of information retrieval in the traditional ACA method, but in the content-based ACA method, he belongs to the factor of bibliometric. The major research field of Rice is scholarly communication which is more relevant to bibliometrics. This indicates that the content-based ACA method reflects the research field of an author more accurately than the traditional ACA method does. In some cases, as in bibliometrics, many of these were classified in a higher sub-discipline rather than a more detailed factor label. This supports Zhao's results (2006).

With respect to information retrieval, most of authors belong to information retrieval, and factors related to language model, query, and clustering (F4) and factor related to user interfaces (F8) appear in both methods consistently. In particular, with respect to the factor label related to "language model, query, and clustering", all 10 authors belonging to the factor in the content-based ACA method also belong to the factor in the traditional ACA methods. In the traditional ACA, the factor of user interface, Jakob Nielsen, Ben Shneiderman and Fred D. Davis are belonging the factor of user interfaces. In the content-based ACA method Nielsen and Shneiderman, the representative researchers in the field of user interface, are included and the factor splits to another factor (F9) which includes Fred D. Davis.

With respect to the factor of general information retrieval, it is observed that the sentence

similarity method consists of factors of various sub-disciplines of information retrieval. It should be also noted that the only traditional ACA method factor label that is not used by sentence similarity is the generic sub-discipline, “information retrieval”. In general, content-based ACA method discovered new sub-disciplines within information retrieval and bibliometrics which were used to distribute authors with a finer granularity. For instance, in traditional ACA method there was a factor label defined as “text mining, machine learning (F7)” which contained 6 authors (Dumais, S., Joachims, T., McCallum, A., Sebastiani, F., Witten, I.H., and Yang, Y). When content-based ACA factor labels are generated, this label still exists with only 4 authors (Joachims, T., McCallum, A., Sebastiani, F., and Yang, .Y.) but the other label is created called “digital library, text mining (F10)” and it contains one authors (Witten, I.H). Susan Dumais and Hao Chen are located in another factor label defined as “classification algorithms”.

Another interesting characteristic in factors is that both Marcia J. Bates and Birger Hjørland belong to the one “Domain analysis, information behavior” factor. Both authors belong to the factor of information retrieval in traditional ACA method, but Bates and Hjørland have an exchange in the pages of *JASIST* about nature of information science. The reason that these two researchers belong to this factor is because they are frequently co-cited in the citing sentences although the particular research interests of both researchers are different. Therefore, sentence similarity method supports the identification of associations between authors based on content similarity of the citing sentences. A similar pattern is observed with the factor of bibliometrics with webometrics (F18) and network analysis (F23) factors being found in both methods. Again, the content-based ACA method generates more factors representing sub-disciplines than the other two methods but it is interesting to note that traditional and content-based ACA methods have some common comparisons.

The bioinformatics and statistic topic areas did not show any sub-disciplines, most likely due to the small number of co-cited authors. Note that bioinformatics is not in the scope of *JASIST*. In the bioinformatics factor, Don Swanson is the only author that appears in a different factor in both methods. In the traditional ACA method, Cohen belongs to general information retrieval factor, but in the content-

based ACA method, it is observed that he is also cited together with Cohen's kappa as evaluation coefficient that is related to machine learning.

4.3. Network analysis

4.3.1 Author co-citation network structure

To examine whether there are structural differences in two ACA methods, we built the author co-citation network by the list of authors who received more than one co-citation count (Fig. 11 and 12). In the case of network mapping by the total number of authors regardless of co-citation count, 70.2% of the authors have one co-citation count for the traditional ACA method and 69.3% in the content-based ACA method. In addition, there are a number of authors who do not belong to main three topics (information retrieval, bibliometrics, and bioinformatics). This makes it difficult to identify the outstanding structure of the fields. Therefore, in the case of assigning edge weights based on frequency for traditional ACA method, we map authors who have more than a weight of 5. In case of the content-based ACA method, because edge weight is relatively lower than the traditional ACA method, we map authors with similarity of two or higher. 1,099 authors and 2,457 authors are selected in the traditional ACA and the content-based ACA methods respectively. For the node size on the network, node weight is calculated by degree centrality. In addition, we used the modularity algorithm optimized on large network (Blondel et al., 2008) for clustering that was designed to measure the strength of division of a network into modules (or clusters, communities).

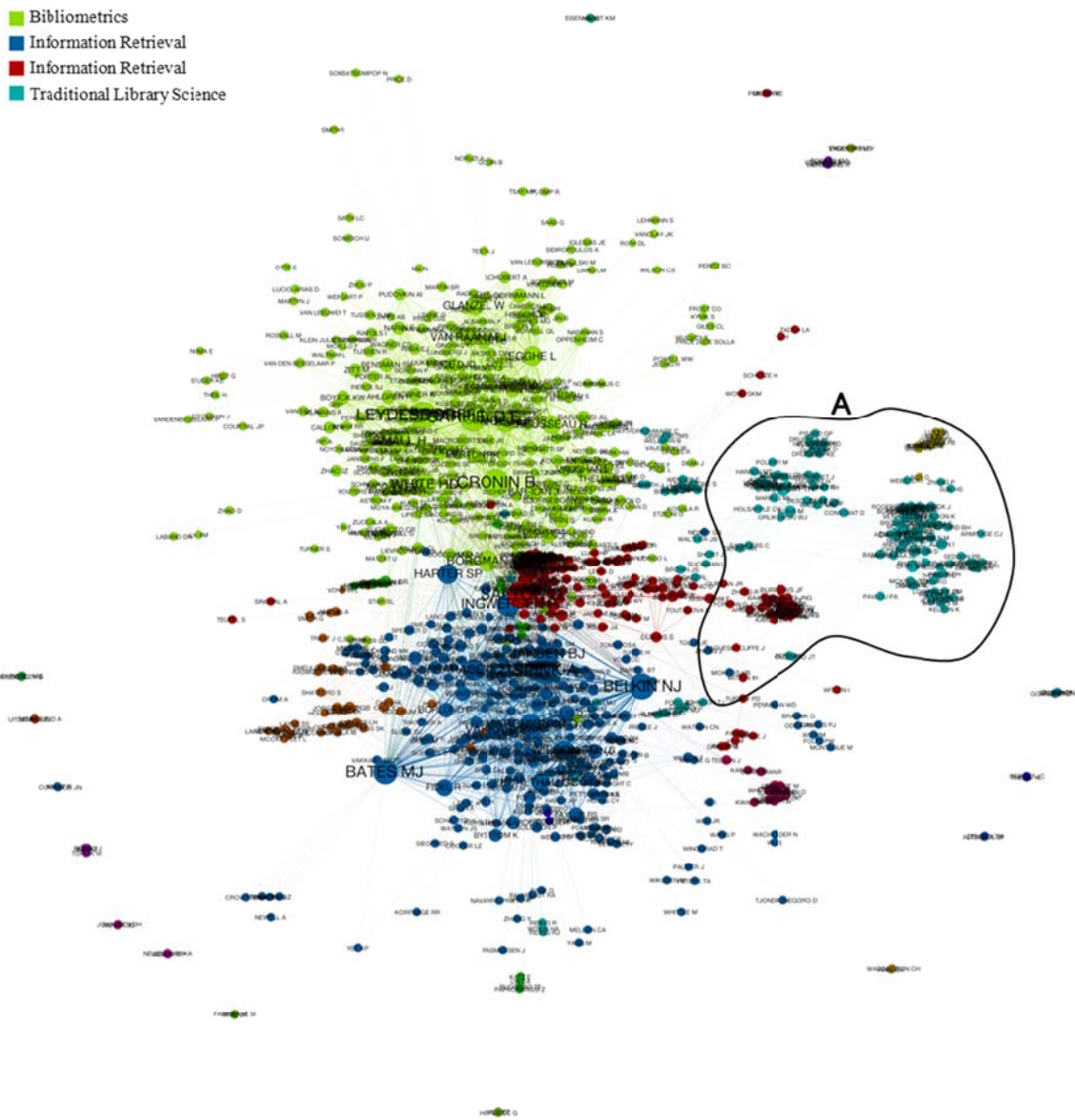


Fig. 6. Author co-citation map of the traditional ACA method.

Fig. 6 shows the mapping results of the traditional ACA method. This is similar with the result of factor analysis that consists of author clusters of bibliometrics and information retrieval. The bibliometric cluster is located in the upper and the central part of the map (in green). Information retrieval is split into lower two clusters (in blue and red). In Fig. 6, there are sub-disciplines newly identified by the author co-citation map that are not revealed by factor analysis based on top 100 authors. That is, the right side

cluster (A) in Fig. 6 shows sub-disciplines such as traditional library science, knowledge management, and library management. These sub-disciplines are formed as a distinct cluster that is differentiated from main *JASIST* components (e.g., information retrieval and bibliometrics). Fig. 7 shows the map of the content-based ACA method. Compared with Fig. 6 more clusters are spread out over the map.

Cluster (B) that is located in the lower part in Fig. 7 was not identified in Fig. 6. The sub-discipline “text mining” was spun off as an independent cluster. Cluster (C) is related to user behavior analysis and Cluster (D) is about social network analysis.

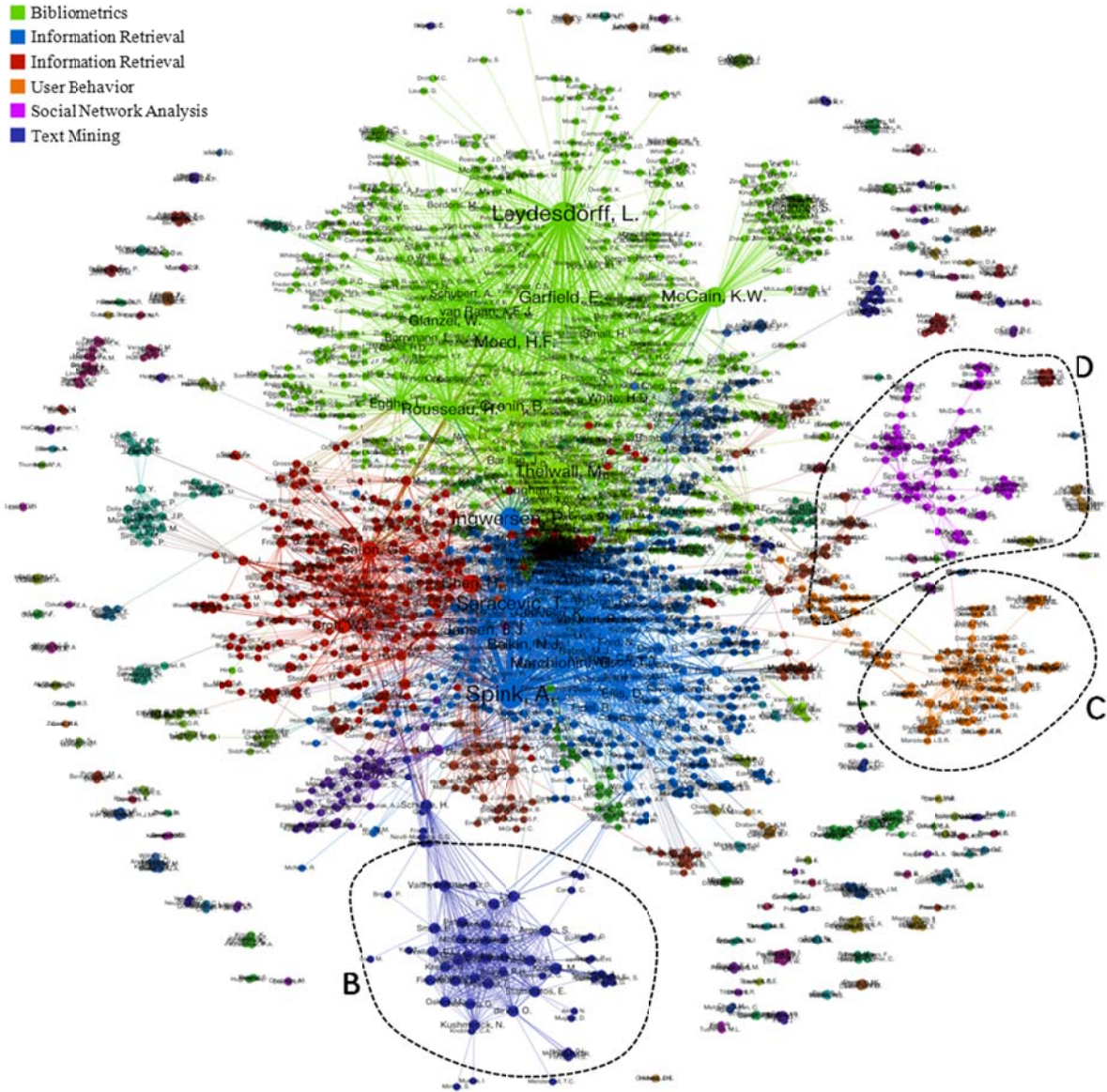


Fig. 7. Author co-citation map of the content-based ACA method.

4.3.2 Network map with the list of top 100 authors

The citation network based on traditional ACA with the top 100 authors is shown in Fig. 8. Pearson's correlation matrix is inappropriate for network analysis in that it consists of a few zero values in most cases. If the matrix is fully connected as a complete network, it is hard to interpret the structure of a specific discipline. To prevent this problem, previous studies invented link reduction algorithms. White

and McCain (1998) only kept loadings that are 0.3 or higher in the map. Zizi and Beaudouin-Lafon (1994) imposed a link weight threshold and only included links with weights above the threshold. We accepted the Zizi and Beaudouin-Lafon's method and extracted the co-citation matrix from Pearson's correlation matrix with a value over 0.75 to investigate the sub-structural feature. To identify the main research area of the authors associated with clusters, we extract keywords from cited paper titles that belong to a certain community. Keywords are tokenized and stop words are removed so that we keep keywords with high frequencies. Because the input data, Pearson correlation matrix, are the same between two, network analysis shows the similar result with factor analyses.

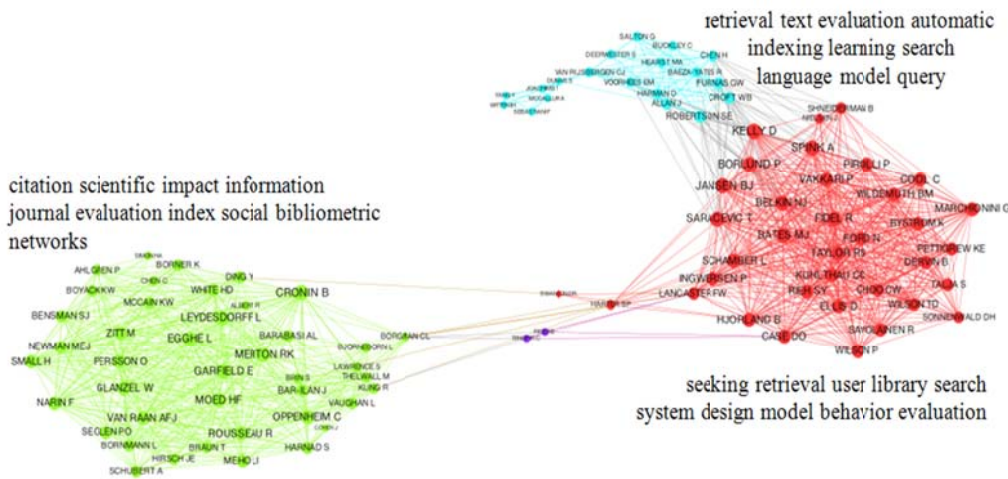


Fig. 8. Map of traditional ACA method.

Fig. 8 illustrates an extremely simple map whose values of link weights are at least 0.75. It shows roughly two parts with 4 modularity classes. As shown in Fig. 8, information retrieval and bibliometrics are two major research areas in *JASIST*, information retrieval along with information seeking behavior and information systems are located at the bottom, and the authors related scholarly communication are located between two major groups. There are only 2 authors who belong to these clusters (Rice and Tenopir). Bibliometrics turns out to be a core constituent in information science. The result of traditional ACA simply shows the mainstream topics of *JASIST* for the last 10 years (Fig. 8). This is in accordance with the result of White and McCain's study (1998) that was based on well-known

researchers in information science.

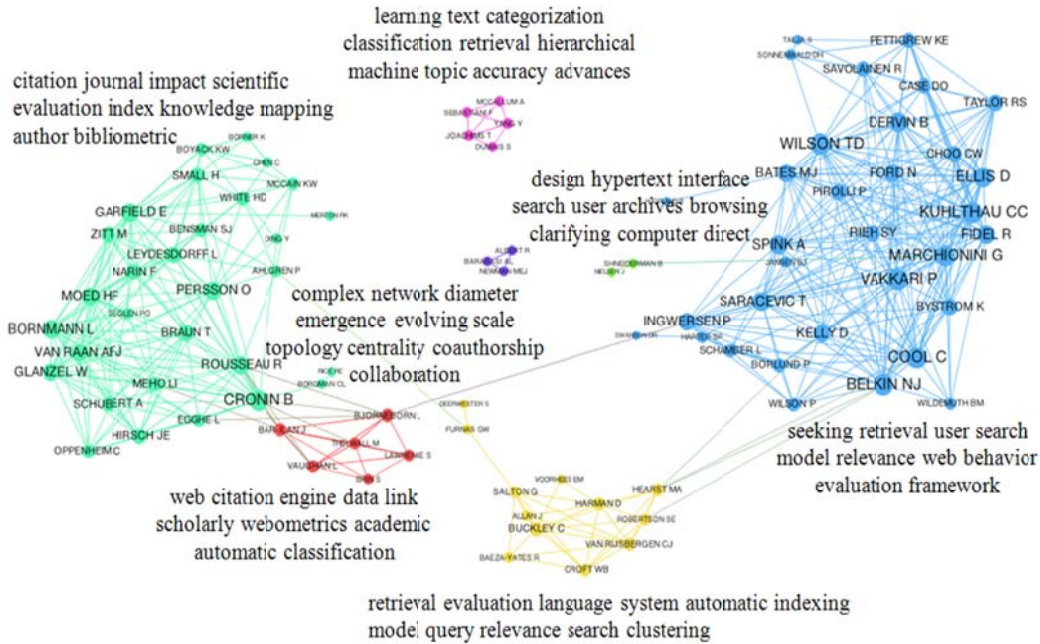


Fig. 9. Map of the content-based ACA method.

The proposed content-based ACA method (Fig. 9) shows a clear difference from previous network maps, and it also represents the research areas more specifically and vividly. Fig. 9 shows the group that is related to information retrieval is fragmented into three groups: information seeking behavior, text mining (machine learning, text categorization, and classification), language model and user interface. The authors grouped by information seeking behavior formed the main group. Neilson and Shneiderman are leader researchers in the group related to user interface. The authors such as McCallum, Sebastian, and Yang are grouped in one cluster more explicitly. The group related to bibliometrics is also separated into three parts 1) a (green) cluster including journal citation analysis, evaluation indicator, and visualization, 2) webometrics (red), and 3) network analysis that authors who are related to network analysis such as Barabasi and Newman, and they are located in the center of the network as an independent cluster (dark blue). With respect to the field of information retrieval, it is split into 1)

information seeking behavior (light blue) located in the right side of the network and 2) language model and query cluster (yellow) located in the lower part of the network. Unlike Fig. 8, a cluster that is related to machine learning and classification is located in the above center of the network as an independent cluster. In addition, user interface (green) is located in the left side of information seeking behavior which is similar with the factor analysis result.

The following observations also seem to be well reflected compared to the other two results: 1) author relation in the citation context is revealed by factor analysis, 2) the relationship between Bates and Hjørland shows a finer granularity and indicates a bridge between the two main fields (bibliometric and information retrieval), and 3) the cluster that Rice belongs to is actually a separate smaller discipline that utilized techniques from one field (information retrieval) but overlaps the other (bibliometrics).

Table 5 ACA network statistics.

	Traditional ACA		Content-based ACA	
	Overall ACA Map (Fig. 6.)	Top 100 Authors (Fig. 8.)	Overall ACA Map (Fig. 7.)	Top 100 Authors (Fig. 9.)
Network Diameter	8	8	15	8
Graph Density	0.010	0.274	0.002	0.115
Clustering Coefficient	0.545	0.819	0.413	0.672
Average Path Length	3.349	2.871	4.612	3.495

Table 5 illustrates the basic statistics of the co-citation networks. Overall, the traditional ACA network shows more comprehensive network structure than content-based ACA does. This is attributed to the calculation of graph density and clustering coefficient of the traditional ACA network are larger than the content-based ACA's. That is, these diameters reveal that the proposed content-based ACA network is somewhat sparse and is split into sub-components contrary to the traditional ACA network.

Consequently, these results show that our approach is more appropriate to reveal the sub-structure in domain analysis at a more detailed level than traditional approaches and also enables us to identify a large number of sub-disciplines.

5. Conclusion

Although ACA has proven to be an effective method for eliciting a bird's eye view of the intellectual structure of a research field, there are certain limitations to ACA as with any methodology. Recent research attempts have been made to seek better alternatives to some or all of its classic components. The present study contributes to this research trend. This study examines a novel methodology measuring the similarity between two cited authors in a research paper using citing sentences in the full-text. We introduced the content-based ACA method, and compared traditional ACA method in the information science field by using the electronically available issues of *JASIST*, a key journal in the information science field.

The results of the experiment show that citation ranking is sensitive to different types of citation counting, especially between traditional ACA and content-based ACA. However, overall these two methods produced maps that share similar patterns, but many differences were observed at the detail level. With respect to the structure of network, a clear distinction by the content-based ACA method exists between traditional ACA. The results show the content-based ACA method reveals more specific subject fields than the traditional ACA. The main difference from traditional approach and content-based ACA method is that we use citing sentences (citation content) as the unit of analysis. This in turn enables us to reflect the topical relatedness (citation content) in citation analysis more precisely and thus allows for more specific domain analysis. From the macro perspective, in addition to finding of authors who are subjectively related, it becomes useful to identify the context of author co-citation by analysis of citing contents in a precise manner. The results of factor analysis affirm that our method is more suitable for identifying the sub-structure in domain analysis at a more detailed level than current approaches.

Although this study proposes a novel method of ACA, there is still much work to be done. The values of the simple cosine similarity measure we use range between 0 and 1. However, the values are either close to 0 or are exactly 1. For example, if an author cited exactly the same sentence, then the value

of the similarity is 1. In other words, the authors in the same sentence are most similar in the paper, and these authors receive highly weighted values. To gain more specific and accurate structure, we need to adjust or normalize these similarity values.

Nevertheless, our results indicate that citing sentences are considerably useful for traditional ACA because they help discover the essential structural components of the corresponding traditional co-citation network. The methods proposed in this paper extracted citing sentences from HTML-formatted full-text articles, and they can be used as an adequate method for the study of ACA in information science and any other fields. We expect that this method provides rich interpretability by revealing the detailed structure in a research domain. These findings also help improve an understanding of citation content that may influence the outcome of co-citation analysis.

As a future study, we plan to include other key conferences in the field of information science such as *Information Processing and Management* and *Journal of Informetrics* to map a broader picture of the field. In addition, we will take into consideration the citation position (proximity) alongside citation content with the underlying assumption that the closer the citations are the more assimilated. In conducting the present study, we observed that citing sentences in the Introduction section tend to have multiple citations. Also, since the citation position provides richer contexts for the reason for citation, it is worth further investigating the impact of the citation position in ACA.

Acknowledgments

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012-2012S1A3A2033291) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2012033242).

References

Ahlgren, P., Jarneving, B., and Rousseau, R. (2003). Requirements for a cocitation similarity measure,

with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54 (6), 550–560.

Acedo, F.J., and Casillas, J.C. (2005). Current paradigms in the international management field: An authorco-citation analysis. *International Business Review*, 14, 619–639.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proc 3rd Intl ICWSM Conf*.

Boyack, K. W., Small, H., and Klavans, R. (2013). Improving the Accuracy of Co-citation Clustering Using Full Text. *Journal of the American Society for Information Science and Technology*, 64(9), 1759-1767.

Bayer, A. E., Smart, J. C., and McLaughlin, G. W. (1990). Mapping Intellectual Structure of a Scientific Subfield through Author Cocitations. *Journal of the American Society for Information Science*, 41(6), 444-452.

Blondel, D. V., Guillaume, J., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10, 1000.

Callahan, A. and Hockema, S. (2010). Contextual Cocitation: Augmenting Cocitation Analysis and Its Applications. *Journal of the American Society for Information Science and Technology*, 61(6), 1130-1143.

Chen, C., Ibekwe-SanJuan, F., and How, J. (2010). The Structure and Dynamics of Cocitation Clusters: A Multiple-Perspective Cocitation analysis. *Journal of the American Society for Information Science and Technology*, 61(7), 1386-1409.

Chen, L. C. and Lien, Y. H. (2011). Using author cogitation analysis to examine the intellectual structure of e-learning: A MIS perspective. *Sceintometrics*, 89, 867-886.

Ding, Y., Chowdhury, G., and Foo, S. (1999). Mapping the intellectual structure of information retrieval studies: An author co-citation analysis, 1987–1997. *Journal of Information Science*, 25(1), 67–78.

Elkiss, A., Shen, S., Fader, A., Gunes, E., David, S., and Dragomir, R. R. (2008). Blind men and

- elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1), 51–62.
- Eom, S. (2008). All author cocitation analysis and first author cocitation analysis: A comparative empirical investigation. *Journal of Informetrics*, 2, 53-64.
- Giles, C. L., Bollacker, K. D., and Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. In E. Witten, R. Akscyn, & F.M. Shipman III (Eds.), *Digital Libraries 98: The Third ACM Conference on Digital Libraries* (pp. 89–98). New York: ACM Press.
- Gipp, B., and Beel, J. (2009). Citation Proximity Analysis (CPA)—A new approach for identifying related work based on co-citation analysis. In B. Larsen & J. Leta (Eds.), *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)* (pp. 571–575). Leuven, Belgium: International Society for Scientometrics and Informetrics.
- Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, C. L. (2010). Context-aware Citation Recommendation. *Proceedings of the 19th international conference on World wide web*, April 26-30, 2010, Raleigh, North Carolina, USA.
- He, Y., and Hui, S. C. (2002). Mining a Web Citation Database for author cocitation analysis. *Information Processing and Management*, 38, 491-508.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569–16572.
- Hornik, K. (2007). Snowball: Snowball Stemmers. R package version 0.0-1.
- Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage
- Liu, S. and Chen, C. (2011). The Effects of Cocitation Proximity of Cocitation Analysis. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of ISSI 2011-The 13th International Conference on Scientometrics and Informetrics* (pp. 474–484), Durban.

- Lu, K. and Wolfram, D. (2012). Measuring Author Research Relatedness: A Comparison of Word-Based, Topic-Based, and Author Cocitation Approaches. *Journal of the American Society for Information Science*, 63(10), 1973-1986.
- Persson, O. (2001). All author citations versus first author citations. *Scientometrics*, 50(2), 339–344.
- R Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Ma, R., Dai, Q., Ni, C., and Li, X. (2009). An author co-citation analysis of information science in China with Chinese Google Scholar search engine, 2004–2006. *Scientometrics*, 81(1), 33-46.
- MacRoberts, M.H., and MacRoberts B.R. (1984). The negational reference: Or the art of dissembling. *Social Studies of Science*, 14, 91–94.
- McCain, K. W. (1990). Mapping Authors in Intellectual Space: A Technical Overview. *Journal of the American Society for Information Science*, 41(6), 351-359.
- McCain, K. W., and Salvucc, L. J. (2006). How influential is Brooks' Law? Alongitudinal citation context analysis of Frederick Brooks' The Mythical Man Month. *Journal of Information Science*, 32, 277–295.
- Nakov, P. I., Schwartz, A. S., and Hearst, M. A. (2004). In Citances: Citation sentences for semantic analysis of bioscience text. Paper presented at the SIGIR 2004 Workshop on Search and Discovery in Bioinformatics, Sheffield, UK.
- Nanba, H. and Okumura, M. (1999). Towards multi-paper summarization using reference information. *IJCAI '99: Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 926–931.
- Nanba, H., and Okumura, M. (2005). Automatic detection of survey articles. *In Research and Advanced Technology for Digital Libraries, 9th European Conference*, 391–401.
- Nerur, S., Rasheed, A. A., and Natarajan, V. (2008). The intellectual structure of the strategic management. *Strategic Management Journal*, 29, 319–336.

- Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582.
- Ritchie, A., Teufel, S., and Robertson, S. (2006). How to find better index terms through citations. *Proceedings of COLING/ACL Workshop on How Can Computational Linguistics Improve Information Retrieval (COLING/ACL)*, Sydney.
- Schneider, J. W., and Borland, P. (2007). Matrix comparison, part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science*, 58(11): 1586-1595.
- Schneider, J. W., Larsen, B., and Ingwersen, P. (2009). A comparative study of first and all-author cocitation counting, and two different matrix generation approaches applied for author cocitation analyses. *Scientometrics*, 80(1), 103-130.
- Small, H. (1973). Co-citation 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.
- Small, H. (2010). Maps of science as interdisciplinary discourse- co-citation contexts and the role of analogy. *Scientometrics*, 83(3), 835–849.
- Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, 87(2), 373–388.
- Teufel, S. (1999). *Argumentative Zoning: Information Extraction from Scientific Text*, PhD thesis, School of Cognitive Science, University of Edinburgh, UK
- White, H. D. and Griffith, B. C. (1980). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 163–171.
- White, H. D., and McCain, K. W. (1998). Visualizing a discipline: An author cocitation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, 49, 327–355.
- White, H. D. (2003a). Author cocitation analysis and Pearson's r . *Journal of the American Society for Information Science*, 54(13), 1250–1259.

- White, H. D. (2003b). Pathfinder Networks and Author Cocitation Analysis: A Remapping of Paradigmatic Information Scientists. *Journal of the American Society for Information Science*, 54(5), 423–434.
- Zhang, G., Ding, Y., and Milojevic, S. (2012). Citation Content Analysis (CCA): A Framework For Syntactic and Semantic Analysis of Citation Content. (Submitted.)
- Zhao, D. (2006). Towards all-author co-citation analysis. *Information Processing & Management*, 42(6), 1578–1591.
- Zhao, D. and Logan, E. (2002). Citation analysis of scientific publications on the Web: A case study in XML research area. *Scientometrics*, 54, 449–472.
- Zhao, D. and Strotmann, A. (2007). Can Citation Analysis of Web Publications Better Detect Research Fronts? *Journal of the American Society for Information Science and Technology*, 58 (9), 1285–1302.
- Zhao, D. and Strotmann, A. (2008). Comparing all-author and first-author co-citation analyses of information science. *Journal of Informetrics*, 2, 229-239.
- Zhao, D. and Strotmann, A. (2011). Counting first, last, or all authors in citation analysis: Collaborative Stem Cell Research Field. *Journal of the American Society for Information Science and Technology*, 62 (4), 654–676.