Journal of Informetrics 10 (2016) 954–966

Contents lists available at ScienceDirect

Journal of Informetrics

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

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.

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 (H.J. Kim), yk.jeong@yonsei.ac.kr (Y.K. Jeong), min.song@yonsei.ac.kr (M. Song).

http://dx.doi.org/10.1016/j.joi.2016.07.007
1751-1577/

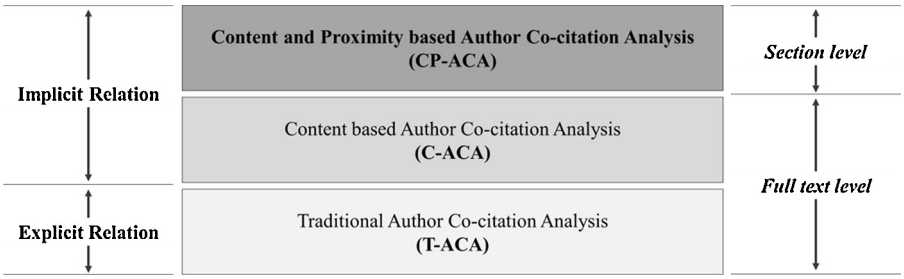

Fig. 1. Research design.

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

with traditional co-citation analysis. They defined four levels of proximity, namely, sentence, paragraph, section, and article. Liu and Chen (2012) extended their 2011 study by utilizing 22 BMC journals to verify the pattern of CPA. Through their 2011 and 2012 studies, they delineated the distribution of co-citation proximity, presenting the percentage of co-citation proximity at the sentence, paragraph, section, and article levels. Moreover, they clarified the distribution of sections and different proximity levels in the background and introduction sections. The authors then generalized the distribution of CPA by extending the number of journals (2012).

Building on Gipp (2006) and Gipp and Beel (2009) and previous CPAs, Callahan, Hockema, and Eysenbach (2010) proposed a new method called co-citation strength. Instead of considering the position of the citation sentences, they took into account the weights attributed to citation sentences when located in close proximity to one another. They then used the BioMed Central database; their proposed co-citation strength analysis was more effective than simple co-citation analysis for identifying the relations between citation sentences. Eto (2013) also assumed that two citations located farther from one another were weaker than those within enumerations located closer to one another. He attempted to enhance the performance of co-citation retrieval by utilizing CPA. In his 2013 study, he used CiteSeer metadata to evaluate performance using six different co-citation methods. Furthermore, Boyack et al. (2013) conducted citation proximity analysis in different ways. They proposed and used a new distance measure incorporating byte counts on full-text into a scoring function to compute co-citation strength between two reference pairs.

Most CPAs have considered document structure and assumed that two citation sentences' level of relatedness increases with their proximity to each other. However, those studies were limited to numerical values as they only counted the frequency or the weight based on the location. As such, the present work suggests a content- and proximity-based ACA. The assumption is that the more closely located authors cite other authors' works in a document, the more subject relatedness there must be between the two cited authors by analyzing citation content. This method can help unveil the implicit relationships in authors' relatedness through third parties from the citation placement and researchers' contributions by analyzing the content and proximity between authors co-cited in sentences.

2.3. Citation content analysis

Most studies of citation content analysis have focused on citation classifications, summaries, or sentiment analyses (Angrosh, Cranefield, & Stanger, 2010; Di Marco & Mercer, 2004; Elkiss et al., 2008; Hayes, Andersen, Nirenburg, & Schmandt, 1990; Nanba, Kando, & Okumura, 2011; Teufel, Siddharthan, & Tidhar, 2006a, 2006b; Yu, 2013).

Citation classification analyses were used to reveal authors' reasons for citing other documents (Garfield, 1955; Hayes et al., 1990; Moravcsik & Murugesan, 1975). Most of them relied on a citation category or codebook used to identify the motivations behind citations. In addition, Teufel et al. (2006a) presented an enhanced citation classification method from 320 transcripts of conference articles and classified 548 citation sentences.

Recently, Zhang, Ding, and Milojević (2013) proposed a new citation content analysis method based on semantic and syntactic analysis. Semantic-based citation analysis is a qualitative form of analysis used to discover the citation motivation and classification. Syntactic-based citation analysis considers the citation location and frequency, revealing the hidden relations between authors from the metadata of documents. Ding et al. (2014) developed the method combining semantic and syntactic types of citation content analyses. They proposed a theoretical methodology based on content citation analysis.

However, previous studies on citation classification and citation categories were limited to unveiling the hidden author's intention, which depended on the codebook or categorization. Previous studies on citation content did not pay much attention to author co-citation analysis. To broaden citation content analysis to include author co-citation analysis, our study proposes a new method, CPACA, which considers citation content with citation location to identify subject disciplines on authors.

3. Research design and method

The theoretical framework in ACA is realized with three different approaches (Fig. 1); Traditional Author Co-citation Analysis (TACA), Content-based Author Co-citation Analysis (CACA), and Content- and Proximity-based Author Co-citation Analysis (CPACA).

Among the major problems in bibliometrics is how to map out the intellectual structure of a research field. Only identifying the explicit relation between two authors through TACA, based on the frequency of author mentions of a cited reference, does not represent the topical relation of the two authors' research fields. TACA has a narrow view on suggesting the most actively cited authors in the research field. Thus, this method only conveys the explicit relationship between authors at the document surface. Unlike TACA, CACA and CPACA are based on citation content. CPACA considers the citation's location. Both citation content and proximity play an important role in identifying sub-disciplines of authors' implicit relationships from citation content and location.

The present study extends the work of Jeong et al., 2014, a CACA study to the CPACA study. All citation sentences in a document are counted as at the same level within full-text articles by calculating the similarity between two citations. Subsequently, CPACA was performed. CPACA analyzes the subject disciplines among authors at a fine-grained level. Facing the limitation on CPA, which mainly focuses on quantitative measures, the researchers attempted to grasp authors' relationships at the citation location level with citation content analysis. CPACA counts the pair of citation sentences within the same section.

3.1. Methodology

The study examines whether and how TACA, CACA, and CPACA differ. The overall flow is illustrated in Fig. 2. The details of each step are: (a) querying the PubMed Central database for data collection; (b) extracting citation sentences and cited reference information from the collected data and location information at the pre-process stage; (c) counting the author co-citation for first author only; (d) delineating two different approaches for ACA, namely, CACA within a full-text level and CPACA within the section level; and then (e) interpreting the different author co-citation results.

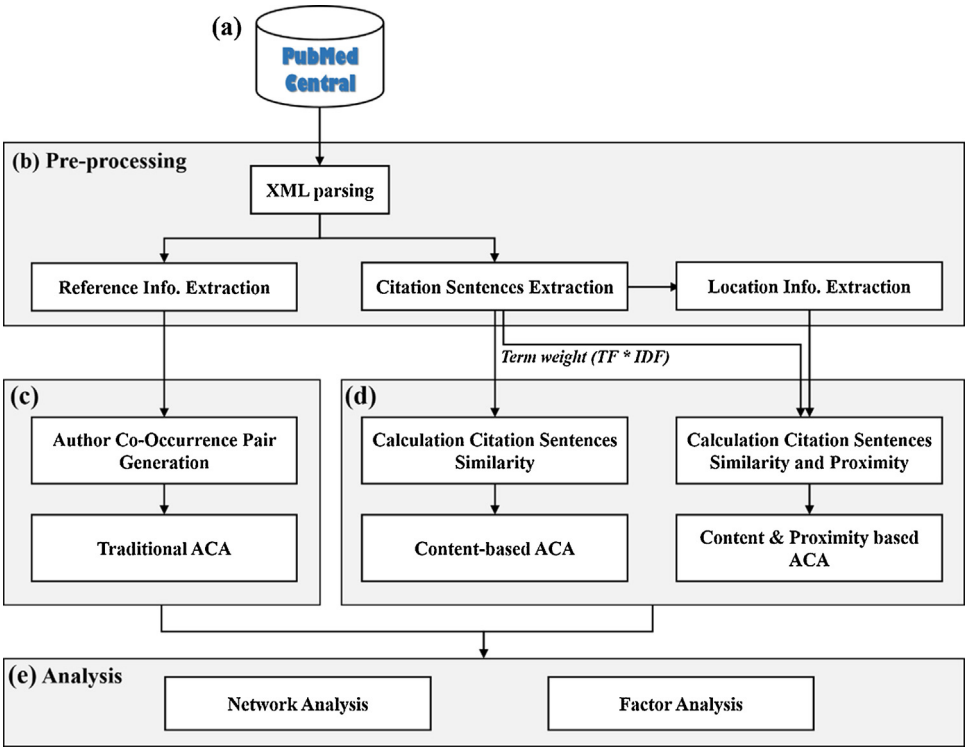

Fig. 2. Workflow.

Table 1
Journal list.

Journal title	Number of full-text articles	Percentage of full-text articles	Impact Factor
Molecular Cancer	1,547	98.30%	4.257
Oncotarget	1,671	88.30%	6.359
OncoImmunology	835	78.50%	6.266
Clinical Epigenetics	148	56.20%	4.543
Breast Cancer Research	674	30.70%	5.49
Oncogene	753	21.60%	8.459
Journal of the National Cancer Institute	125	16.30%	12.583
Leukemia	285	15.50%	10.431
Stem Cells	107	13.50%	6.523
Journal of Thoracic Oncology	32	9.90%	5.282
Annals of Oncology	105	6.60%	7.04
Neuro-Oncology	30	3.40%	6.776
Pigment Cell & Melanoma Research	21	3.30%	4.619
Molecular Oncology	7	1.80%	5.331
Cancer Cell	20	1.80%	23.523

3.1.1. Data collection in oncology research field

The dataset was collected from PubMed Central (http://www.ncbi.nlm.nih.gov/pmc), which provides full-text articles in the biomedical field. The researchers focused on "Oncology," with consideration for the recent surge in number of publications in this field. *Stem cells*, one of the subfields of oncology, has been at the forefront of medicine (Zhao & Strotmann, 2011). Selection of the top 15 oncology journals was based on JCR (https://jcr.incites.thomson-reuters.com/) and the impact factor among those highly ranked journals (up to 30). Fifteen journals were retrieved, yielding a total of 11,636 records in xml format. Table 1 shows the final 15 oncology journals containing 6,360 full-text articles out of 11,636 from 1999 and 2015, the years used in this study. Table 1 also shows the proportion of full-text articles per journal in the dataset. The total percentage of full-text is about 34% of articles published in the selected journals for the experiments. As shown in Table 1, in PubMed Central, the field of oncology is not evenly distributed across the specific journals and we have relatively small coverage on the full-text articles in the oncology field. This is due to the characteristics of an open access repository. Like our study, previous studies had also limitations to collect the full-text within the open source database (Boyack et al., 2013; Elkiss et al., 2008; Zhao & Strotmann, 2014). Since the main purpose of our study is to propose a novel method on author co-citation analysis using full-text, we followed the full-text collection strategy used in previous studies. To conduct the proposed study, we collected as much full-text data as possible from PubMed Central in the oncology field.

3.1.2. Pre-processing

3.1.2.1. Extraction of citation sentence and reference information. To extract citation sentences from collected xml data, the reference style presented in citation sentences was first considered. Most citation sentences have a certain reference style and are kept in the following formalized format: (author, year), (reference number), and [reference number] at the end of the citation sentences.

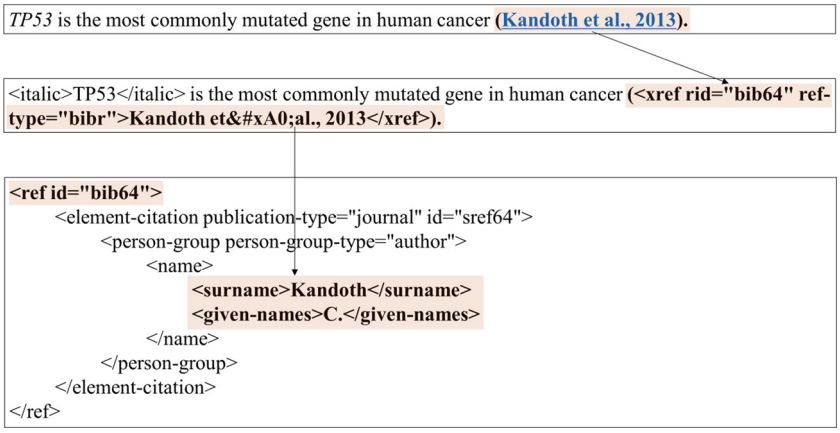

Fig. 3. Example of reference style and citation sentence in XML format.

Table 2
Example of parsed data information from citation extraction.

PubMed Central ID	4023863
Authors	Khan Irum, Huang Zan, Wen Qiang, Stankiewicz Monika J., Gilles Laure, Goldenson Benjamin, Schultz Rachael, Diebold Lauren, Gurbuxani Sandeep, Finke Christy M., Lasho Terra L., Koppikar Priya, Pardanani Animesh, Stein Brady, Altman Jessica K., Levine Ross L., Tefferi Ayalew, Crispino John D.
Title	Akt is a therapeutic target in myeloproliferative neoplasms
Journal title	Leukemia
Section	Introduction
Citation sentence	A number of other JAK inhibitors are in varying stages of pre-clinical and clinical development (22, 23).
Cited author ID (reference ID)	[R22 R23]
Name of cited author	Pardanani A

Given the example of xml data shown in Fig. 3, citation sentence adopted the following format: "(<xref rid='bib64' ref-type="bibr">Kandoth et al., 2013</xref>)." Thus, a regular expression technique to parse and extract the citation sentences in the tag <xref rid=> was applied, and </xref> was found in the citation sentences after parsing the xml data with the Java-based SAX parser.

We extracted the last and first names of the authors cited in the paper from the reference data. The metadata with the cited authors' names in the reference section were parsed using the following tags: <surname> and <given-names> inside the <person-group>. Fig. 3 gives an example of the way the names of cited authors were extracted from the reference information section.

3.1.2.2. Extraction of location information for citation sentences. Previous studies such as those using TACA count mention frequency from the reference information when authors A and B are co-cited. They provide a mere relation of authors as they only account for simple counts of the authors' oeuvres. Unlike previous studies, the present study sought to refine the interpretation of the relation between two citation sentences implicitly by considering not only the similarities between the pair of citation sentences but also the different locations of the author co-citation, using CACA and CPACA.

With the section tag, the full-text articles in xml were parsed to extract the location information for CPACA. The tag <sec-type> on <sec sec-type = "Method"> referred to the section in which the citation sentences were located at the section level.

3.1.2.3. Final parsed data collection of citation information. Table 2 shows a set of elements needed for analysis, such as citation sentences, cited authors' reference information, and location information of the citation sentences. The researchers collected the PubMed Central (PMC) ID, PMC authors, title, journal title, and section location, as well as the citation sentences, reference ID, and cited author information. The example of a paper by Khan et al. (2013) is provided in Table 2. Khan Irum is a first author of "Akt is a therapeutic target in myeloproliferative neoplasms." and has a PubMed Central ID of 4023863. Khan's paper, published in *Leukemia,* has 17 co-authors. For this paper, the researchers extracted the following citation sentence: "A number of other JAK inhibitors are in varying stages of pre-clinical and clinical development (22, 23)." This citation sentence is located in the introduction section. The numbers 22 and 23 represent the reference IDs and the two authors cited by Khan. The author names for IDs 22 and 23 were then extracted from the reference section; the citation sentence referred to Pardanani A's work.

3.1.3. Calculation of citation sentence similarity

The present study expanded Jeong et al., 2014 by incorporating proximity into citation sentence similarity. They used simple term frequency to calculate the similarity between co-cited authors. However, the present study took proximity into consideration; citation sentences' closer proximity tends to reflect closer subject relatedness between cited authors. The cosine similarity of citation sentences was computed with the term frequency-inverse document frequency (tf-idf) weighting of citation sentences. ACAs were conducted at the

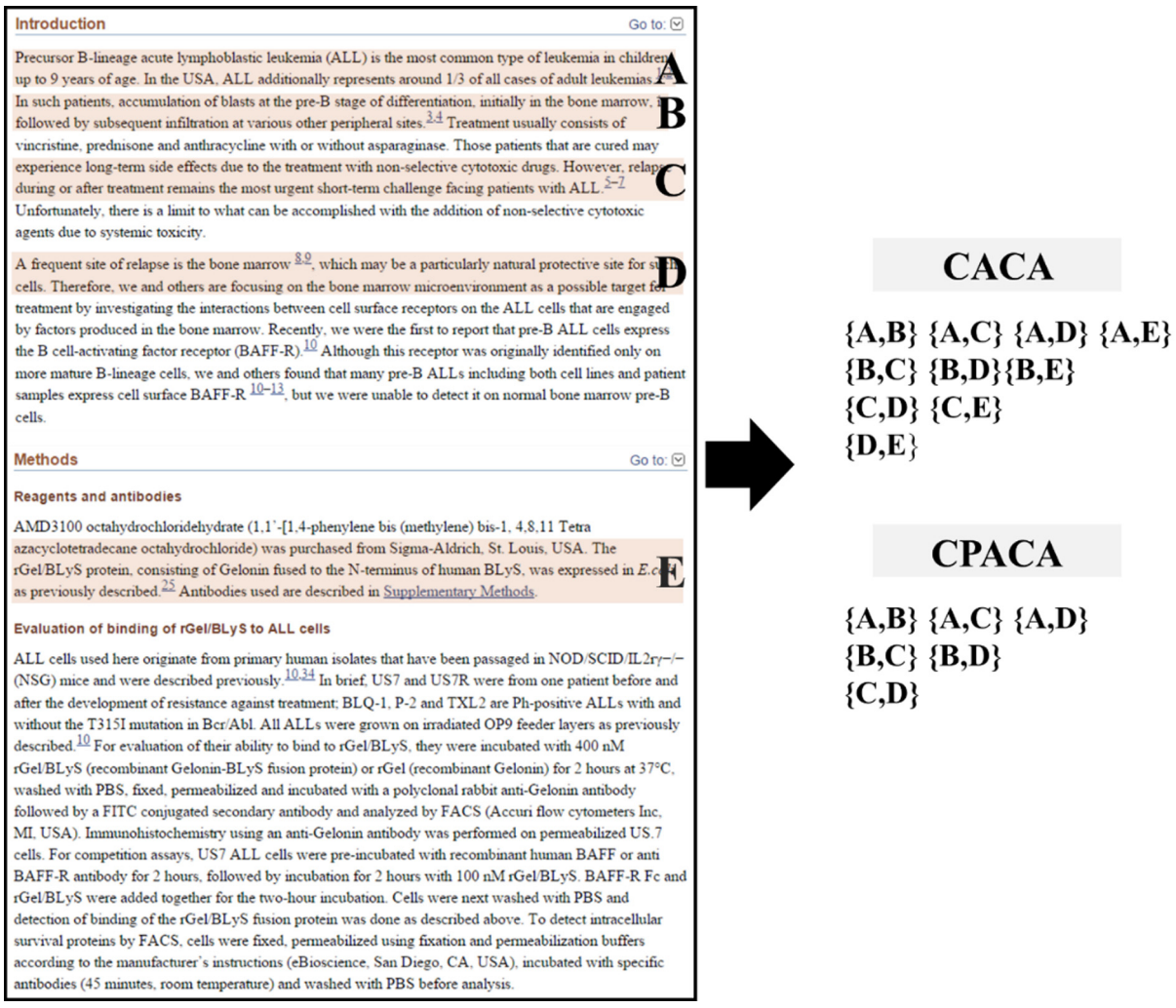

Fig. 4. Example of pair generation for CACA and CPACA.

full-text level as CACA and section level as CPACA. As CPACA counted the two citation sentences within the same section, it considered the proximity of citation sentences. The researchers calculated the two citation sentences, X and Y, as means of cosine similarity as follows:

$$similarity\,(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_i X_i Y_i}{\sqrt{\sum_i X_i^2}\sqrt{\sum_i Y_i^2}}$$

where Xi and Yi are tf-idf values of each word in the citation sentences X and Y, respectively. The cosine similarity is measured by using the weighted word. In computing the cosine similarity measures, if the documents share any words, the cosine similarity score is higher. Stopwords were removed and the citation sentences lemmatized for computed similarity measures. The cosine similarity was calculated according to word co-occurrences. The researchers selected the highest similarity score between duplicate cited authors and added a similarity score when the same pair of authors was cited. When the similarity score is 1 and only appeared once, the similarity score was excluded. Although a similarity score of 1 assumes that two citation sentences have the highest relatedness in terms of subject field, it indicates that the two authors were only cited at one time.

3.1.4. ACA

All three author co-citation approaches were based on first authors in the present study. The similarity scores between two citation sentences were applied to the CACA and CPACA. The proximity of the citation sentences was considered at section level as CPACA.

Following previous studies (Andrews, 2003; Eom, 2008; Zhao & Strotmann, 2011; Zhao, 2006), in the case where two authors were cited by another author, the number of reference occurrences of authors was counted for TACA. First-author reference information in the reference section was extracted and the citation frequency for two cited authors counted.

For CACA, the researchers first considered every citation sentence as a unit at the full-text level. All citation sentences were counted equally based on similarity scores. For CPACA, the similarity of the citation sentences was counted at the section level.

Fig. 4 illustrates how the citation sentences were assessed at each location level. Suppose that there are five citation sentences A–E in a document. At the section level, four citation sentences A–D appear in the introduction, and one citation sentence E in the methods section. When ACA is considered for CACA at the full-text level, the following pairs are generated: {A,B}, {A,C}, {A,D}, {A,E}, {B,C}, {B,D}, {B,E},

Table 3
Example of citation pattern.

To date, these projects have resulted in 29 primary research publications [25], ranging from reports on the cumulative risk of breast cancer in families with BRCA1 and BRCA2 mutations to studies of psychological morbidity in these families [8,19,26–37]

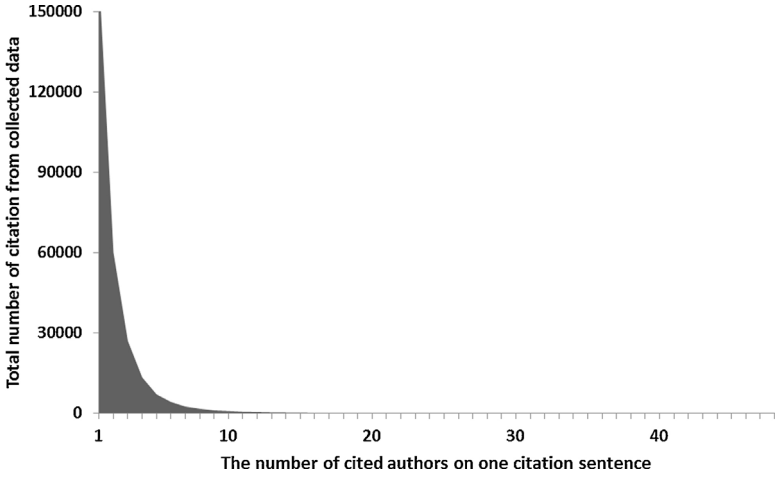

Fig. 5. Citation count within one citation sentence.

Table 4
Number of author co-citation pairs for the three approaches.

	TACA	CACA	CPACA
Number of author co-citation pairs	7,105,990	5,558,458	3,088,096

{C,D}, {C,E}, and {D,E}. For CPACA at the section level, {A,B}, {A,C} {A,D}, {B,C}, {B,D}, and {C,D} are generated in the introduction and methods section.

3.1.5. Analysis

Network analysis was performed with different approaches. In addition, we conducted factor analysis for interpreting research fields revealed in each network. We computed the frequency of word occurrence in the title of papers and Medical Subject Headings (MeSH) to determine the subject fields on author groups. Unlike previous studies using the titles of either their papers or highly cited articles (Zhao & Strotmann, 2008b, 2008a; Jeong et al., 2014), we used both the title of papers and MeSH to interpret topical topology of the author groups. In the medical field, since MeSH terms are used as a domain-specific knowledge resource, it is an excellent option for our case. To extract frequently mentioned MeSH terms, we filtered terms by MeSH hierarchy. Since the terms in MeSH category B (Organisms) and E (Analytical, Diagnostic and Therapeutic Techniques and Equipment) were not suitable to recognize the sub-disciplines of oncology, we excluded the terms in these categories. Furthermore, we conducted expert analysis. As mentioned in Zhao and Strotman (2011) and Andrews (2003), human judgements are needed to interpret each research area on authors. For our study, three experts helped us to interpret each field on authors' clusters. The three experts are one professor, Ph.D student, and a B.A. student either in biology or physiology major.

3.2. Data description

3.2.1. Citation count within one citation sentence

Citation behavior or citation pattern has a difference between the areas of research fields in the sciences, humanities, and liberal arts (Andrés, 2009). When taking citation into consideration, there are noticeable differences on the number of citations and authors. Journals in the sciences tend to be rapidly circulated; hence, researchers need to use the most recent studies. The oncology field is not an exception.

This present study covered oncology, which has similar citation patterns on journals. As shown in Table 3, a citation sentence (Mann et al., 2006) published in *Breast Cancer Research* can give an example of citation patterns. One sentence has 15 cited references, which means the paper's researchers cited more than 15 other authors in one sentence. Thus, the number of citations can be overvalued in science fields.

The number of citations in each citation sentence was analyzed. In Fig. 5, the x-axis denotes the number of cited authors in each citation sentence and the y-axis, the total number of citation in the dataset. In the collected data, the highest numerical value was 163,076 depicting the citation frequency for one author in a single sentence. The number of citations in which two authors are cited in one sentence was 60,097. Thus, the occurrence of up to two authors in a single sentence takes up 78% of all results. The most number of authors mentioned in a citation sentence is 47.

3.2.2. Number of author co-citation pairs

All three approaches were conducted by considering only first authors. Table 4 shows the number of author co-citation pairs by three different approaches. TACA had 7,105,990 pairs between two authors cited, whereas CACA and CPACA had 5,558,458 and 3,088,096 author pairs, respectively. These numbers indicate that a pair of author co-citations reduced as the scope of analysis unit was narrowed down.

4. Result

This section examines the intellectual structures of the oncology field generated by the different author co-citation approaches using network analysis. Nodes on network denote the name of authors, whereas the node size is based on degree centrality. For edge weight, TACA was based on citation frequency; CACA and CPACA relied on the content similarity measure between citation sentences. Modularity algorithms (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) were applied to interpret each author cluster in the network. For factor analysis, we used the similarity matrices of highly cited authors. We extracted 98 authors who were cited more than 300 times in the oncology research area. The factors were extracted by principal component analysis (PCA) with oblique rotation. To identify the subject areas in author clusters, frequently appeared words were analyzed by collecting the title of the authors' papers located in the same cluster. We also used MeSH terms assigned to these articles. MeSH terms, controlled by expert indexers, represent the topics of the articles rather than specific entities used in the research. Therefore, by adopting both title words and MeSH terms as a unit of analysis, we were able to grasp both subject headings of research topics and the specific key terms extracted from titles. In the TACA network, authors in the main cluster were highly cited by other researchers. The TACA Network is shown based on degree centrality (over 6). CACA and CPACA were conducted to overcome the limitation of TACA. CACA was implemented in Jeong et al., 2014, and CPACA was proposed by considering the location of citation sentences with content analysis. The similarity between two citation sentences was measured according to cosine similarity. Unlike TACA, CACA and CPACA do not rely on reference co-occurrence represented by the binary values of 0 and 1; both instead use the sentences' similarities in measuring how two citation sentences are semantically similar to each other while considering contents.

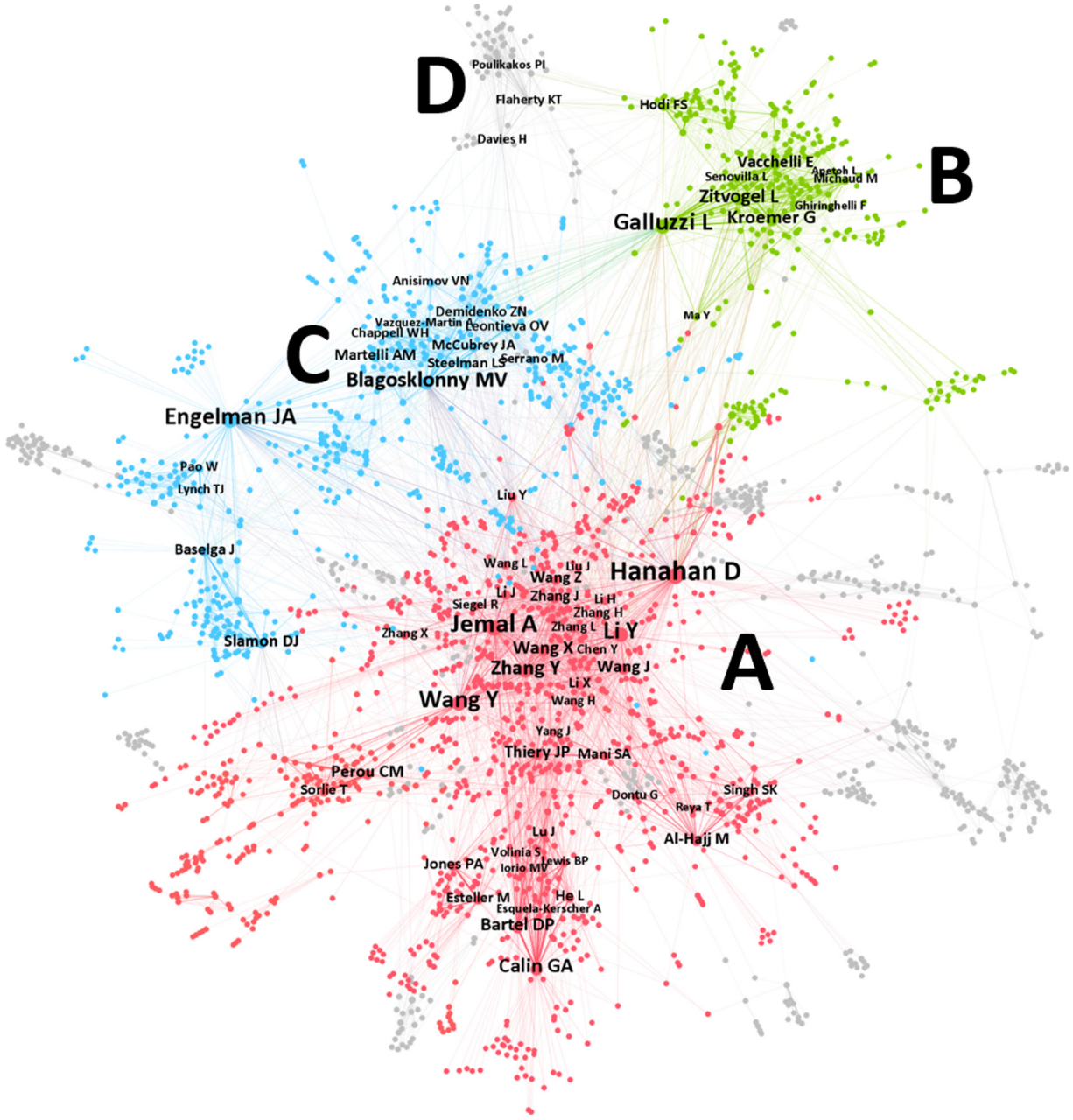

Fig. 6. Author network for TACA.

CACA only counted the citation similarities at the full-text level, whereas CPACA counted the citation sentences within the same section level. A network of CACA and CPACA was mapped, and edge weight was based on the similarity measures. The network with edge weight of two or higher was presented.

Fig. 6 shows the TACA network (number of nodes: 2,315; number of edges: 7,338). The modularity algorithm yielded four distinguished components on TACA. Cluster A in Fig. 6 is mainly related to the gene expression in cancer cell. D. Hanahan, A. Jemal, and Y. Zhang are the most influential authors based on degree centrality with respect to those studying the broad field of tumor cell and gene expression. Cluster B can be categorized as the cancer immunology field; L. Galluzzi, L. Zitvogel, and G. Kroemer are the most highly cited authors in the group. These authors are experts in metabolism of immunologic response on cancer. On the left side of the cluster, cluster C focuses on the oncogene pathway field. The leading authors include M.V. Blagosklonny, J.A. Engelman, and J.A. McCubrey. In cluster D, K.T. Flaherty and H. Davies are the leading authors who are experts in melanoma.

Fig. 7-a demonstrates CACA (number of nodes: 669; edges: 1,222) and Fig. 7-b shows CPACA (number of nodes: 967; edges: 1,546). When compared with TACA, the most remarkable difference of CACA and CPACA is that they demonstrated sub-areas in oncology. The CACA approach shows different network structures from TACA, which is each author's divergent composition in each clustering. Cluster A on TACA was segmented into different clusters A-1–A-3. Cluster A-1 is pertinent to stem cell based on cancer cell expression. Cluster A-2 can be categorized as breast cancer. Cluster A-3 is related to cancer cell metastasis with targeted therapy. As a result, subject fields from clusters A-1–A-3 are subdivided from cluster A on TACA. S.K. Singh, L. Ricci-Vitiani, and M. Al-Hajj are located in Cluster A-1 and are interested in stem cell based on cancer cell expression. Researchers such as T. Sorlie, C.M. Perou, and C. Sotiriou are in cluster A-2; they have influence in the breast cancer field. The research area of the author group A-3, where leading authors include L.M. Weiner, A. Mantovani, and G.L. Semenza is closely associated with cancer cell metastasis with targeted therapy. As mentioned in Jeong et al. (2014), CACA can identify more sub-disciplines compared with TACA as the former approach considers the content of citation sentences between two authors. Subject fields in clusters B and C have similar results with those in TACA. Cluster B is related to cancer immunology and cluster C can be the field of oncogene pathway.

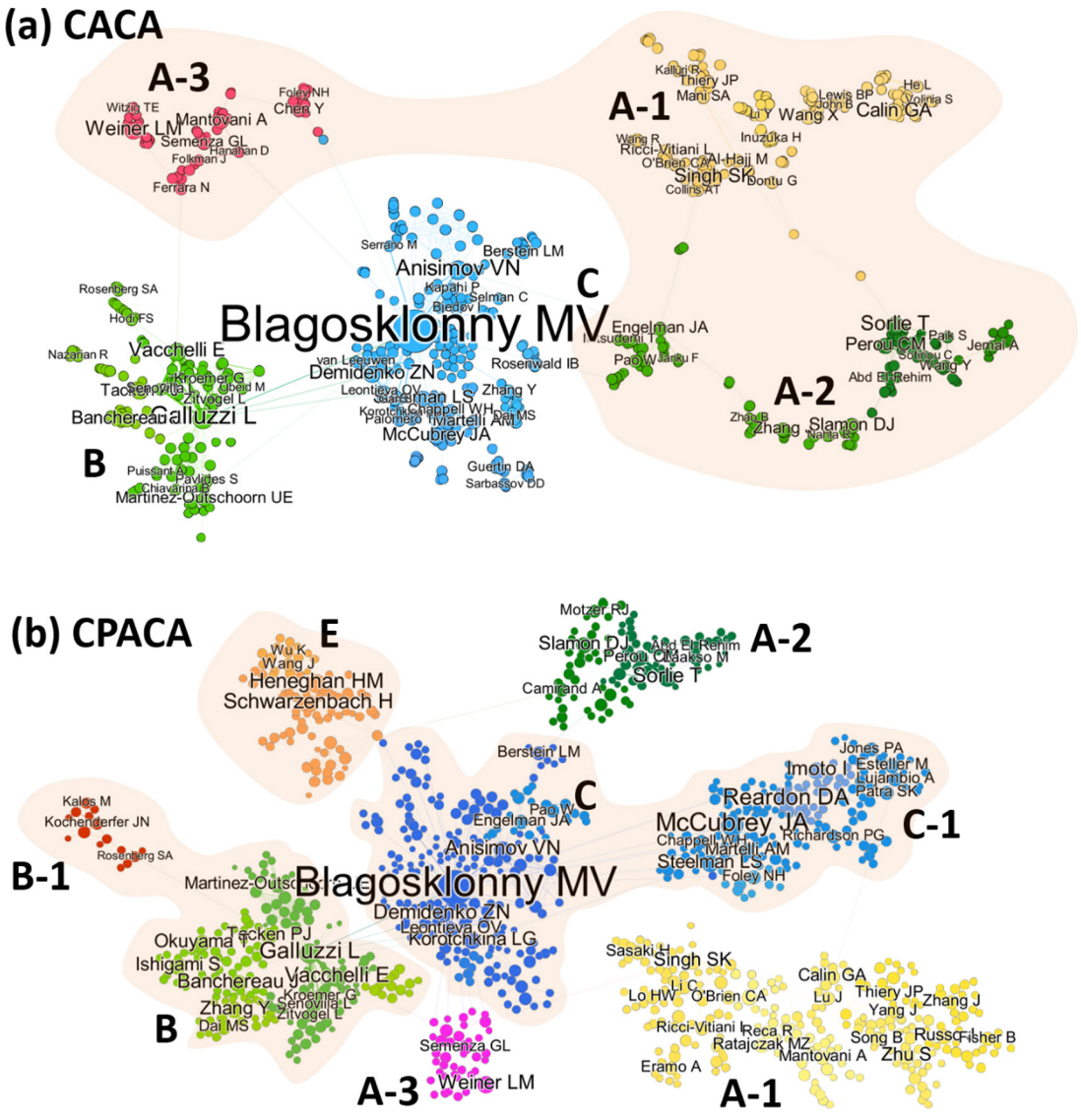

Fig. 7. Author networks of CACA (a) and CPACA (b).

With consideration for the location of citation sentences, the network of CPACA is similar to that of CACA, as shown in Fig. 7-b. However, there are certain unique features that can only be induced by CPACA. CPACA presents more sub-disciplines than CACA: clusters B and C are segmented. Cluster B-1 on CPACA can refer to the sub-discipline of immunology called molecular mechanism of cancer immunology, which is from the main topic of cluster B on CACA. Cluster B-1 contains top authors including S.A. Rosenberg, R.J. Brentjens, and M. Kalos. Their papers are closely related to cancer immunotherapy and cancer antigens for specific diseases. For example, "Cancer immunotherapy: moving beyond current vaccines" (Rosenberg, Yang, & Restifo, 2004) and "T cells with chimeric antigen receptors have potent antitumor effects and can establish memory" (Kalos et al., 2011) are highly cited articles in the cancer immunology field. CPACA's Cluster C-1 presents more detailed subject fields diversified from cluster C on CACA. Cluster C-1 is about the molecular pathway of oncogenesis. The authors J.A. McCubrey, L.S. Steelman, and W.H. Chappell, who are top authors in C-1, have a frequently cited paper entitled "Roles of the Raf/MEK/ERK pathway in cell growth (McCubrey et al., 2007)." Further, Cluster E is newly identified in the CPACA network. H.M. Heneghan and H. Schwarzenbach are influential authors in cluster E, which are related to the microRNA field.

Network analyses allow us to discover the differences among TACA, CACA and CPACA. Authors in TACA are more famous and frequently cited by other authors. The highly cited authors have a more influence on the network than less cited authors do due to the fact that TACA uses simple co-cited pairs. In CACA method, the connection between two cited authors is made by common terms in the citation sentences. Since CACA considers the relationship between authors in terms of common terms or concepts, it is more effective to identify a small unit of sub-disciplines (Jeong et al., 2014). However, the results of CPACA are quite different from the other two approaches. CPACA identifies more topically related authors than TACA and CACA. This may be attributed to the fact that we consider both proximity and contents in author co-citation analysis. Proximity on citations can reveal how closely related are the citation contents, which is realized by CPA (Gipp & Beel, 2009). By adopting citation sentences, our method also has the merit of identifying implicit thematic relatedness between co-cited authors, because citation sentences contain key concepts of cited articles. Therefore, CPACA reveals more sub-fields and authors tend to be grouped per sub- field.

Following the results from network analyses, Table 5 shows how the sub-disciplines are differently presented according to ACA approaches. In Table 5, the differences in sub-fields are explained with frequently appearing words extracted both from titles and MeSH terms.

The TACA approach identified four distinguished subject disciplines: gene expression in cancer cell, cancer immunology, oncogene pathway, and melanoma. As the TACA approach only considers citation frequency, TACA results can explicitly identify the research areas and provide the macroscopic view on oncology. Meanwhile, CACA and CPACA clusters are divided into sub-disciplines. Cluster A is separated into "stem cell based on cancer cell expression," "breast cancer," and "cancer cell metastasis with targeted therapy" from both CACA and CPACA results.

More specifically, CPACA has more clusters divided into sub-disciplines than CACA. Clusters B-1, C-1, and E contain more words specific. Cluster B is related to cancer immunology, whereas Cluster B-1 identifies molecular mechanism of cancer immunology, having "immunotherapy," "metastatic," and "antigen." Cluster C focuses on the oncogene pathway field, including "pathway," "mTOR," and "p53." By MeSH terms, broader topics such as "Cell line," "Tumor cells," and "RNA" are identified. Cluster C-1 has more words relevant to the

Table 5
Subject disciplines for each ACA approach.

	TACA	CACA	CPACA	Keywords	MeSH
Cluster A	Gene expression in cancer cell			cancer, cell, expression	Cell line, Tumor, Gene expression regulation
Cluster A-1		Stem cell based on cancer cell expression	Stem cell based on cancer cell expression	stem, epithelial-mesenchymal, epithelial	Gene expression regulation, Carcinoma, Cell proliferation
Cluster A-2		Breast cancer	Breast cancer	breast, expression, carcinoma	Breast neoplasms/pathology, Breast neoplasms/genetics, Breast neoplasms/drug therapy
Cluster A-3		Cancer cell metastasis with targeted therapy	Cancer cell metastasis with targeted therapy	macrophage, angiogenesis, hypoxia-inducible	Immunohistochemistry, Hypoxia-Inducible Factor 1, Macrophages/immunology
Cluster B	Cancer immunology	Cancer immunology	Cancer immunology	therapy, immune, anticancer	Tumor, Antigens, Antibodies
Cluster B-1			Molecular mechanism of cancer immunology	immunotherapy, metastatic, antigen	Antigens, Immunotherapy, T-Lymphocytes
Cluster C	Oncogene pathway	Oncogene pathway	Oncogene pathway	pathway, mTOR, p53	Cell line, Tumor cells, RNA
Cluster C-1			Molecular pathway of oncogenesis	Akt, mTOR, pTEN	Molecular sequence data, Signal transduction, Receptors
Cluster D	Melanoma			melanoma, BRAF, resistance	Melanoma/drug therapy, Drug resistance, Melanoma/pathology
Cluster E			microRNA	microRNA, gene, circulating	Gene expression regulation, RNA, MicroRNAs

Table 6
Factor analysis for each ACA approach.

Label	TACA (5 Factors)			CACA (7 Factors)			CPACA (11 Factors)		
	Factor	Highest loading	# of Authors	Factor	Highest loading	# of Authors	Factor	Highest loading	# of Authors
Gene expression in cancer cell	F1, F4, F5	0.899	84				F4	0.956	8
Stem cell based on cancer cell expression				F4, F6	0.935	40	F10	0.731	17
Breast cancer				F3, F7	0.758	26	F3	0.932	5
Cancer cell metastasis with targeted therapy				F5	0.688	8	F5	0.781	11
Cancer immunology	F2	0.928	16	F1	0.929	10	F2	0.953	9
Molecular mechanism of cancer immunology							F11	0.562	17
Oncogene pathway	F3	0.812	22	F2	0.854	20	F1, F9	0.909	15
Molecular pathway of oncogenesis							F7, F8	0.904	8
microRNA							F6	0.662	16

Table 7
Top 10 authors for each approach by degree centrality.

	TACA	CACA	CPACA
1	Hanahan D	Blagosklonny MV	Blagosklonny MV
2	Galluzzi L	Galluzzi L	Galluzzi L
3	Wang Y	Anisimov VN	McCubrey JA
4	Jemal A	Martinez-Outschoorn UE	Anisimov VN
5	Blagosklonny MV	Demidenko ZN	Heneghan HM
6	Li Y	Vacchelli E	Schwarzenbach H
7	Engelman JA	Sorlie T	Martinez-Outschoorn UE
8	Zhang Y	Pavlides S	Demidenko ZN
9	Wang X	Chiavarina B	Vacchelli E
10	Zitvogel L	Puissant A	Pavlides S

molecular pathway of oncogenesis, including “Akt,” “mTOR,” and “pTEN.” The new cluster E represents topics on microRNA containing “microRNA,” “gene,” and “circulating.” As a result, although CACA identifies more sub-fields than TACA, CPACA, the present study’s novel approach, shows more subject specialties with top frequency words presenting details, such as chimeric, MEK, pTEN, and microRNA. While CACA allowed for a relationship between authors in terms of conceptual similarity, CPACA can consider the structure of documents by measuring proximity in the section level (Gipp & Beel, 2009). Since an author may cite topically similar works in a same section, the cited literatures have a high co-citation strength (Callahan et al., 2010). Therefore, our approach has an advantage of grasping more topically related authors and identify more detailed sub-disciplines than CACA does.

For in-depth analysis of authors, we also performed factor analysis. Table 6 shows the results of factor analysis. The number of factors extracted was determined based on Kaiser’s rule of eigenvalue greater than 1. In Table 6, we combined the factors with similar topics from 98 authors and displayed the highest loading of an author from each factor. The number of authors was defined as the maximal loadings above 0.20 in the pattern matrix.

As shown in Table 6, the results of factor analysis are similar to our network clustering analyses. Comparing the results from three approaches, we identified that the network by CPACA was well dispersed into sub-discipline of oncology with 11 factors. The total variance of CPACA was 0.731, greater than 0.647 of CACA’s, which indicates that our proposed approach achieves a statistically significant result in terms of the amount of variance explained by factor models.

Further, most of the top authors identified by CPACA based on degree centrality represent more diverse disciplines on oncology compared with TACA and CACA; the implicit relationship among authors is considered from the location of citation sentences. As shown in Table 7, top authors, including D. Hanahan, Y. Wang, A. Jemal, and Y. Li, in TACA are related to the gene expression of cancer cells. The authors are generally highly cited in the comprehensive oncology field.

Authors who are in the gene expression field mostly occupied the top author list of TACA. However, the compositions of top authors are different in CPACA. Authors who are related to the molecular pathway of oncogenesis and microRNA are on the top 10 under CPACA. J.A. McCubrey is related to the field of molecular pathway of oncogenesis, which is more detailed than the field of oncogene pathway. H.M. Heneghan and H. Schwarzenbach are related to microRNA, and they newly appeared on the top 10 list in CPACA. As a result, new authors who evidently appear in CPACA are identified in more segmented subject disciplines.

Strong subject relatedness is revealed by the CPACA approach. For example, H.M. Heneghan and H. Schwarzenbach are shown on the top 10 authors list in CPACA. A citation sentence from these two authors are shown within the same section level: “previously, our group, as well as others, have compared the profiles of circulating microRNAs between breast cancer patients and healthy controls, and attempted to identify circulating microRNA-based breast cancer detection biomarkers (Shen et al., 2014).” Two authors are co-cited together within same section. Closely located citation sentence shows more specific terms, such as microRNA, biomarker, and breast cancer. Authors tend to write subjectively relevant sentences within close locations.

This new finding is also confirmed by the cluster-level comparison. Table 8 shows the top 10 authors based on degree centrality of sub-fields in CACA and CPACA. These clusters in CACA and CPACA have the same sub-disciplines on oncology, but the composition of authors in CPACA is more suitable for recognizing subject relatedness within the same subject areas compared with CACA.

As shown in Table 8, the composition of the top authors in clusters of CPACA is different when compared with CACA. Certain authors persistently appear in the top 10 authors in each CACA cluster, whereas others who appear on the top list of CPACA do not come up on

Table 8
Author composition for CACA and CPACA.

	Stem cell based on cancer cell expression (A-1)		Breast cancer (A-2)		Cancer cell metastasis with targeted therapy (A-3)	
	CACA	CPACA	CACA	CPACA	CACA	CPACA
1	Singh SK	Singh SK	Sorlie T	Sorlie T	Weiner LM	Weiner LM
2	Ricci-Vitiani L	Li C	Perou CM	Laakso M	Mantovani A	Ferrara N
3	Al-Hajj M	O'Brien CA	Wang Y	Abd El-Rehim	Semenza GL	Semenza GL
4	Dontu G	Sasaki H	Jemal A	Perou CM	Chen Y	Folkman J
5	Mani SA	Ricci-Vitiani L	Abd El-Rehim	Wang Y	Ferrara N	Hicklin DJ
6	O'Brien CA	Eramo A	Paik S	Sotiriou C	Zhong H	Zhong H
7	Collins AT	Collins AT	Laakso M	Reis-Filho JS	Hanahan D	Jubb AM
8	Thiery JP	Lo HW	Sotiriou C	Rakha EA	Folkman J	Carmeliet P
9	Yang J	Wang R	Howe HL	Paik S	Foley NH	Tsuzuki Y
10	Li C	Beier D	Nielsen TO	Parker JS	Hicklin DJ	Kim JW

the top list of CACA. J.S. Reis-Filho, E.A. Rakha, and J. S. Parker, who do not appear on the top 10 list in CACA, appear on the top 10 authors in cluster A-2. As researchers cite these authors more frequently within close locations, the three authors come up on the top 10 list. J.S. Reis-Filho, E.A. Rakha, and J.S. Parker are often cited in the discussion section within documents, and their citation sentences have more specific words related to the breast cancer field. By counting the similarity within near proximity and citation content, the authors who are in CPACA are more suitable for identifying subject relatedness in research fields. Moreover, H. Sasaki and A. Eramo, who are more closely related to stem cell, are shown on the top 10 authors list in CPACA. Another example can be shown in cluster A-3. Y. Chen is on the list in CACA; however, A.M. Jubb newly appears in cluster A-3 in CPACA. The importance of Y. Chen is decreased when counting similarity within close locations; A.M. Jubb features more actively within sections by other researchers. Moreover, A.M. Jubb's works are mainly published in *Oncotarget*, a journal that covers molecules and pathways in cancer cell and immunology and microbiology. A.M. Jubb is associated with cancer cell metastasis over targeted therapy. These results indicate that authors in CPACA reveal subject relatedness in research fields.

5. Conclusion

Although author co-citation analyses have long been studied to explore research disciplines, there were limitations in previous approaches of ACA. First, traditional methods on ACA were limited to measuring the relationship of two authors who were co cited in the same article using only bibliographic data. Second, conventional CPA was conducted by a simple counting of the proximity of citation sentences, which provides a narrow view to recognize the authors' subject relation.

To tackle these limitations, the present study proposed a new methodology on ACA, called CPACA, considering the concept of content and proximity of citation sentences to identify the topical relationship among authors in a more explicit manner. This study examined ACA not just using bibliographic data but exploring citation contents and citation locations in full-text, which can reveal authors' relationships in the field of oncology.

The study results showed that CPACA provided a new, distinct perspective from the other two, TACA and CACA. Even though CACA with citation content yields more sub disciplines compared with TACA, CPACA produced most segmented sub-fields as shown in the result section. In addition, CPACA is more appropriate than the other two in studying the authors' subject relatedness from the perspective of identifying implicit relationships among authors via citation. By considering both citation content and proximity, closely cited authors within a section are more suitable to represent the specific field on the group. In addition, in CPACA, authors who do not belong to the top author group in TACA and CACA, newly appear on top, presenting more thematically related authors to a specific sub-field. Since the previous approaches adopted simple co-citation counts, research topics of popular authors have a strong influence on identifying subject disciplines. Unlike previous ones, our approach showed fine-grained structures of sub-disciplines.

Although this study provides a different view of author co-citation analysis, there is still much work to be done. As a follow-up study, we plan to apply CPACA to compare a social science field such as political science with a science field such as computer science. In addition, we consider further partition of the co-citation unit such as paragraph or sentence window. Since this study focused on proposing a new method in author co-citation analysis, we plan to conduct a comprehensive comparison study between the proposed work and other ACA researches.

Acknowledgements

This work was supported by the Bio-Synergy Research Project (NRF-2013M3A9C4078138) of the Ministry of Science, ICT and Future Planning through the National Research Foundation and (in part) by the Yonsei University Future-leading Research Initiative of 2015 (2015-22-0119).

References

Andrés, A. (2009). *Measuring academic research: how to undertake a bibliometric study*. Elsevier.
Andrews, J. E. (2003). An author co-citation analysis of medical informatics. *Journal of the Medical Library Association*, *91*(1), 47.
Angrosh, M. A., Cranefield, S., & Stanger, N. (2010). Context identification of sentences in related work sections using a conditional random field: Towards intelligent digital libraries. In *In Proceedings of the 10th annual joint conference on Digital libraries* (pp. 293–302).
Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.
Boyack, K. W., Small, H., & 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.

Callahan, A., Hockema, S., & Eysenbach, G. (2010). Contextual cocitation: Augmenting cocitation analysis and its applications. *Journal of the American Society for Information Science and Technology*, *61*(6), 1130–1143.
Di Marco, C., & Mercer, R. E. (2004). Hedging in scientific articles as a means of classifying citations. *In working notes of the American association for artificial intelligence (AAAI) spring symposium on exploring attitude and affect in text: theories and applications*, 50–54.
Ding, Y., Chowdhury, G., & 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.
Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, *65*(9), 1820–1833.
Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (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. B. (1996). Mapping the intellectual structure of research in decision support systems through author cocitation analysis (1971–1993). *Decision Support Systems*, *16*(4), 315–338.
Eom, S. (2008). All author cocitation analysis and first author cocitation analysis: A comparative empirical investigation. *Journal of Informetrics*, *2*(1), 53–64.
Eto, M. (2013). Evaluations of context-based co-citation searching. *Scientometrics*, *94*(2), 651–673.
Garfield, E. (1955). Citation indexes for science; a new dimension in documentation through association of ideas. *Science*, *122*(3159), 108.
Gipp, B., & 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). Rio de Janeiro (Brazil): International Society for Scientometrics and Informetrics.
Gipp, B. (2006). Citation proximity analysis—A measure to identify related work. In *Doctoral proposal*. Germany: Otto-von-Guericke University.
Gipp, B. (2014). Citation-based document similarity. In *Citation-based plagiarism detection*. Springer.
Hayes, P. J., Andersen, P. M., Nirenburg, I. B., & Schmandt, L. M. (1990). Tcs: A shell for content-based text categorization. *Artificial intelligence applications* (1990) (pp. 320–326).
He, Y., & Hui, S. C. (2002). Mining a web citation database for author co-citation analysis. *Information Processing & Management*, *38*(4), 491–508.
Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics*, *8*(1), 197–211.
Kalos, M., Levine, B. L., Porter, D. L., Katz, S., Grupp, S. A., Bagg, A., & June, C. H. (2011). T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. *Science Translational Medicine*, *3*(95), 95ra73–95ra73
Khan, I., Huang, Z., Wen, Q., Stankiewicz, M. J., Gilles, L., Goldenson, B., . . . & Lasho, T. L. (2013). AKT is a therapeutic target in myeloproliferative neoplasms. *Leukemia*, *27*(9), 1882–1890.
Liu, S., & Chen, C. (2011). The effects of co-citation proximity on co-citation analysis. *Proceedings of ISSI*, 474–484.
Liu, S., & Chen, C. (2012). The proximity of co-citation. *Scientometrics*, *91*(2), 495–511.
Mann, G. J., Thorne, H., Balleine, R. L., Butow, P. N., Clarke, C. L., Edkins, E., . . . & Giles, G. G. (2006). Analysis of cancer risk and BRCA1 and BRCA2 mutation prevalence in the kConFab familial breast cancer resource. *Breast Cancer Research*, *8*(1), R12.
Marshoakova, I. V. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy*, *6*, 3–8.
McCubrey, J. A., Steelman, L. S., Chappell, W. H., Abrams, S. L., Wong, E. W., Chang, F., . . . & Stivala, F. (2007). Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance. *Biochimica Et Biophysica Acta (BBA)-Molecular Cell Research*, *1773*(8), 1263–1284.
Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*(1), 86–92.
Nanba, H., Kando, N., & Okumura, M. (2011). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, *11*(1), 117–134.
Rosenberg, S. A., Yang, J. C., & Restifo, N. P. (2004). Cancer immunotherapy: Moving beyond current vaccines. *Nature Medicine*, *10*(9), 909–915.
Shen, J., Hu, Q., Schrauder, M., Yan, L., Wang, D., Medico, L., . . . & Qin, M. (2014). Circulating miR-148b and miR-133a as biomarkers for breast cancer detection. *Oncotarget*, *5*(14), 5284–5294.
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.
Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). An annotation scheme for citation function. *Proc. of SIGDial-06*.
Teufel, S., Siddharthan, A., & Tidhar, D. (2006b). Automatic classification of citation function. *Proceedings of the 2006 conference on empirical methods in natural language processing*, 103–110.
White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, *32*(3), 163–171.
White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, *49*(4), 327–355.
Yu, B. (2013). Automated citation sentiment analysis: What can we learn from biomedical researchers. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–9.
Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, *64*(7), 1490–1503.
Zhao, D., & Strotmann, A. (2008a). Information science during the first decade of the web: An enriched author cocitation analysis. *Journal of the American Society for Information Science, Journal of Informetrics*, *59*(6), 916–937.
Zhao, D., & Strotmann, A. (2008b). Comparing all-author and first-author co-citation analyses of information science. *Journal of Informetrics*, *2*(3), 229–239.
Zhao, D., & Strotmann, A. (2011). Counting first, last, or all authors in citation analysis: A comprehensive comparison in the highly collaborative stem cell research field. *Journal of the American Society for Information Science and Technology*, *62*(4), 654–676.
Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology*, *65*(5), 995–1006.
Zhao, D. (2006). Towards all-author co-citation analysis. *Information Processing & Management*, *42*(6), 1578–1591.