Applying Content-based Similarity Measure to Author Co-citation Analysis

Yoo Kyung Jeong[1], Min Song[1]
[1]Yonsei University

Abstract
This study proposed a novel author similarity measure in author co-citation analysis (ACA). Unlike other ACA studies, we used citing sentences to reflect topical relatedness of authors. In our research, we extended traditional approaches by adopting Word2Vec, one of deep learning methods, to measure author similarity. We also conducted in-depth network analysis of author maps. The results of Word2Vec-based author map revealed more specific sub-disciplines and the important authors in perspective of topical influence than traditional approach does. Our method allows for more sophisticated analysis than the traditional ACA approach by providing an in-depth understanding and the specific structure of a discipline.
Keywords: Author Co-citation Analysis; similarity measure; Word2Vec; content analysis
doi: 10.9776/16212

Acknowledgements: This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046711).
Contact: yk.jeong@yonsei.ac.kr, min.song@yonsei.ac.kr

1 Introduction

Author co-citation Analysis (ACA), which was introduced by White and Griffith (1981), has been widely used in bibliometrics researches to identify and trace the intellectual structure of an academic discipline (He & Hui, 2002). In ACA, traditional approaches relied on the co-citation frequency of cited authors in the reference section. Thus, one of the main topics in ACA was methodological discussion of what kind of measure is appropriate and relevant for calculation of author similarities (Leydesdorff, 2005; van Eck & Waltman, 2007). Existing approaches based on co-citation frequencies such as Pearson correlation coefficient and Salton's cosine similarity, however, do not capture the citation content. Thus, some recent researches used the full-text to obtain the topical relatedness between the cited authors (Jeong, Song, & Ding, 2014; Zhao & Strotmann, 2014). They analyzed the authors mentioned in the full-text and incorporated contents related with cited authors into ACA.

This study extended the traditional approach and considered citing sentence in measuring author similarity. In previous researches using full-text, the similarity of authors was measured among in-text citation data (Zhao & Strotmann, 2014) or the citing sentences (Jeong et al. 2014) in a document, which can be called "local similarity" of authors. In this study, however, we used all citing sentences of the author in whole dataset to compute "global similarity" of authors. To this end, we propose the novel author similarity measure by adopting the Word2Vec technique, one of deep learning methods. Since the proposed approach considered all words related with cited authors, it can be used to measure the global relatedness reveal latent structure of a discipline.

2 Method

2.1 Experimental Design

In this section, we provide details of the author similarity measure based on the Word2Vec technique. In the traditional approaches, the unit of analysis is the counts of author co-citation. Thus, the frequency based ACA approaches could not reflect the topical relatedness of authors. Moreover, since traditional approaches only measure the relatedness of authors in a single document, they are inadequate to encompass the whole oeuvre to identify the authors' research area. In that sense, cumulated citing sentences of cited authors are able to well represent the cited researches and cited authors' research areas. In addition, these citing sentences are particularly useful for summarization of a research document. Therefore, our approach of incorporating citing sentences into co-citation allows for discovering the topical relatedness of authors by employing the Word2Vec technique to calculate author similarity.

To evaluate the proposed method, we collected data from the information science domain and compared it with previous other works (Chen, Ibekwe-SanJuan, & Hou, 2010; White & McCain, 1998; Zhao & Strotmann, 2008). One of the important tasks of our research is to extract the citing sentences.

Therefore, for our analysis, we selected the *Journal of the Association for Information Science and Technology* (*JASIST*), which is the most prominent journal in library and information science, and collected the full-text containing citation sentences. 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 pre-processing and computing similarity of the citing sentences by Word2Vec, we developed the natural language processing modules in Java and used Gephi (Bastian, Heymann, & Jacomy, 2009) as the visualization tool. Figure 1 shows the overall system flow of our approach.

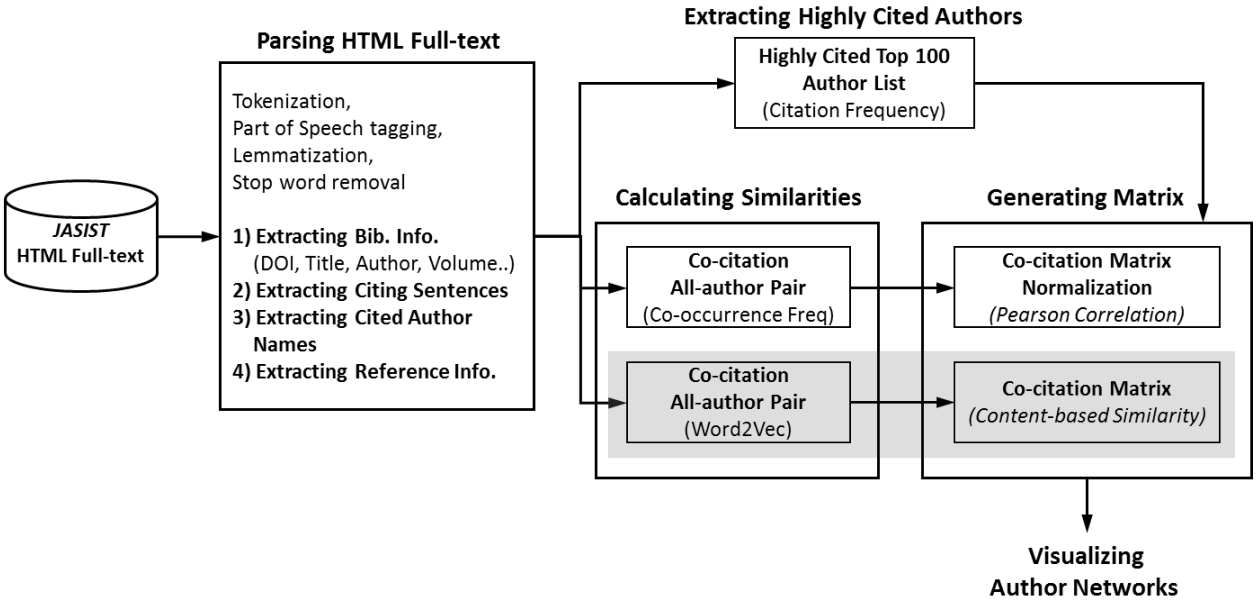

Figure 1. Experimental Overview

In traditional ACA approach, the bibliographic database such as Web of Science or SCOPUS was used to data acquisition. For content analysis, however, we collected full-text research articles of *JASIST* in HTML format. Through the HTML parsing process, we extracted the metadata (title, author name, year, DOI and abstract), citation information (citing sentence, and reference id), and reference information (author name, year, title, and journal). To compare our method to traditional ACA, we computed author-pairs in both approaches. In Word2Vec-based method, the full-text data, first, are splitting into sentences. In second step, matching the citing sentences with reference id in reference section, we separated the citing sentences and other general sentences. Then, citing sentences are preprocessed in the following steps: tokenization, POS tagging, lemmatization of the tokenized sentence, and stop word removal. From these data, we trained Word2Vec model for calculating author similarity and generated author-author similarity matrix. To compare the previous research, traditional author counting approach, we also construct co-citation matrix based on citation counts. Since we preprocessed full-text including all reference information, these matrices considered all cited authors. To evaluation, we selected top 100 authors which are highly cited in both methodology, and conduct network analysis through visualizing author maps.

2.2 Dataset

The data was gathered from 1,910 full-text articles in the *JASIST* digital library over 12 years (from January 2003 to June 2015). The 1,910 collected documents have 77,408 references. We extracted elements from the full-text article: 1) citing sentences from the body of the article, 2) the references information, and 3) all cited authors. Table 1 shows the basic statistics of collected data.

# citing sentences	# of references	# of cited authors	# of co-cited pairs
32,096	77,408	230,005	6,528,731

Table 1. Data Description

2.3 Author Similarity Measure: Word2Vec

Word2Vec models, one of the neural network approaches, are able to carry semantic meanings and turns text into a numerical form that deep-learning nets can understand (Mikolov et al., 2013). Based on a large amount of plain text, Word2Vec trains relationships between words automatically. Word2Vec spatially encoded a word meaning and the relationship between words, which was originally applied to word clustering or synonym detection (Wolf et al., 2014). We applied Word2Vec into author similarity measure regarding cited author names as a word in plain text.

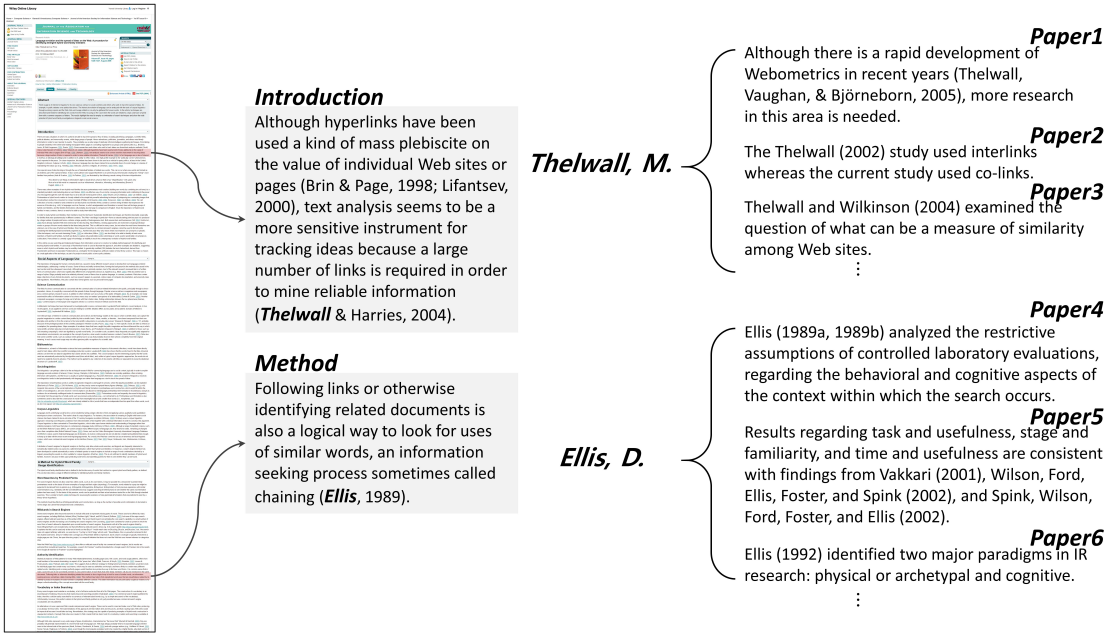

Figure 2. Example of Citing Sentences

In the traditional approach, if two authors, i.e. Thelwall and Ellis, are co-cited in a paper, the co-citation count of these authors increments by 1 regardless of the location or content similarity. On the other hand, Word2Vec trains a large amount of citing sentences while preserving the order of words. Thus, the proposed method overcomes the limitation of the frequency based approach by using citing sentences as contents and training its' sequence. Since authors' oeuvre was represented as the citing sentences in research articles, the Word2Vec-based method could consider various topics of the author. It also covered the co-authorship as well as co-citation. If the number of co-authors is less than four in APA style, a citing sentence includes all names of the co-author list. Since the Word2Vec model trains citing sentences including cited author names, co-author names mentioned in the same citing sentence is incorporated into measuring similarity. For example, the citing sentence of Paper1 in Figure 2 contains the author names: Thelwall, Vaughan, and Björneborn. In the traditional ACA approach, these authors are not counted as co-citation. In the proposed approach, however, the author names are also trained as words in a same citing sentence. Therefore, the similarity between two authors in the Word2Vec-based method reflects both topical relatedness and collaborations. Table 2 shows top 10 pairs by the traditional ACA method (Pearson correlation based similarity) and the Word2Vec based approach respectively. About the half of pairs resulted from the Word2Vec approach are the co-author relationship.

Rank	Pearson correlation (Normalization of co-citation frequency)		Word2Vec-based similarity	
1	(Moed, H.F., Rousseau, R.)	0.97261	(Moed, H. F., van Rann, A. F. J.)*	0.99996
2	(Persson, O., Rousseau, R.)	0.95070	(Jansen, B. J., Spink, A.)*	0.99531
3	(Garfield, E., Narin, F.)	0.94678	(Baeza-Yatets, R., Voorhees, E. M.)*	0.99451
4	(Moed, H.F., Persson, O.)	0.93873	(Dumais, S., Joachims, T.)*	0.99145
5	(Garfield, E., Rousseau, R.)	0.93560	(McCain, K. W., White, H. D.)*	0.98992
6	(Bensman, S. J., Narin, F.)	0.93528	(Robertson, S. E., Harman, D.)	0.98918

7	(Garfield, E., Persson, O.)	0.92463	(Cool, C., Kelly, D.)*	0.98915
8	(Borlund, P., Cool, C.)	0.92410	(Wilson, T. D., Kuhlthau, C. C.)	0.98856
9	(Ford, N., Kuhlthau, C. C,)	0.91649	(Hearst, M. A., Croft, W. B.)	0.98849
10	(Narin, F., Small, H.)	0.91368	(Talja, S., Case, D.O.)	0.98787

Table 2. Top 10 Pairs of Pearson correlation and Word2Vec-based similarity
(* The relationship of co-author)

As shown in Table 2, the research field of the most authors in terms of the frequency based approach is related with bibliometrics which were led by researchers such as Henk F. Moed, Eugene Garfield, and Ronald Rousseau. This is caused by the characteristics of the traditional approach that is based on citation counts. On the other hand, the authors by the Word2Vec-based approach research are associated with various topics of information science: user theory (Wilson and Kuhlthau), information retrieval (Jansen and Spink), bibliometrics (McCain and White).This results imply that the proposed approach enables to detect wider range of author pairs in perspective of topical relatedness and grasp more diverse research fields of information science.

3 Analysis

To examine whether there are structural differences in two measures of author similarity, we constructed two author networks with top 100 authors. For network visualization, we used PageRank (Brin & Page, 1998) to determine the node size and also adopted the modularity algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) for the community detection. Figure 3 shows the author co-citation map of the traditional approach.

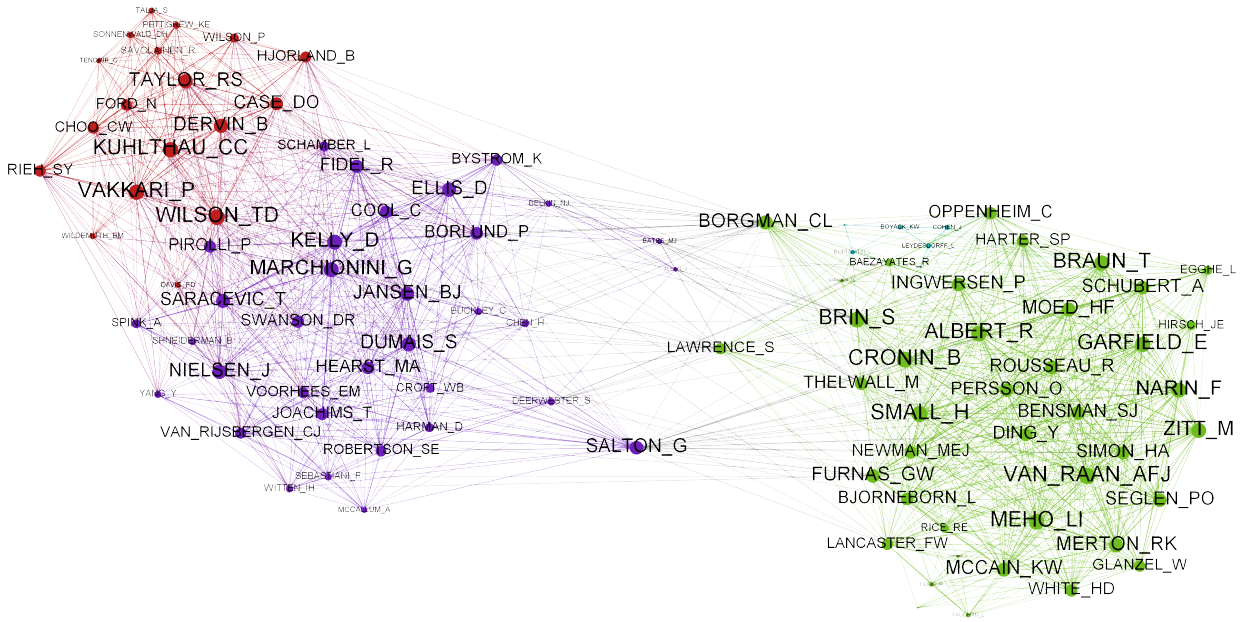

Figure 3. Traditional Author Co-citation Network

Figure 3 illustrates roughly two parts that consist of information retrieval and bibliometrics, two major research areas in *JASIST*. The author group of information retrieval (purple) along with information seeking behavior (red) is located at the left side, and the author group related with bibliometrics is located at the right side. There are only two authors located between two groups (Borgman and Salton), who are traditionally cited authors in the information science field. Borgman studied various topics including information retrieval and scholarly communication and wrote the important books that had won the best information science book from ASIST. Salton's works also received a lot of citations for a long time in the field of information science. This is in compliance with the results of previous works (White & McCain, 1998; Chen et al., 2010) that revealed well-known researchers in information science. The result of the traditional approach shows the mainstream topics of *JASIST*.

The proposed Word2Vec-based author map (Figure 4), on the other hand, shows a different topology that represents the research areas more specifically compared to the traditional approach. The author group related with information retrieval in the left side of the network is split into information seeking behavior (blue) located in the upper side of the network and document retrieval (yellow) located at the bottom side of the network. The group related to bibliometrics is also separated into two parts: (1) a cluster (green) including author analysis and (2) journal citation analysis and evaluation indicator (red). Unlike Figure 3, the communities in the network are connected to each other. Brin is connected with both document retrieval and citation analysis communities. This may be attributed to the fact that the PageRank, developed by Brin and Page (1998), is used in information retrieval and also studied in network analysis to compute node centrality. In bibliometrics, PageRank is adopted as one of the centralities in citation networks (Ding, Yan, Frazho, & Caverlee, 2009). Ingwesen, who is located between information retrieval and bibliometrics, studied information retrieval in earlier works, he extended the research area to network analysis such as webometrics.

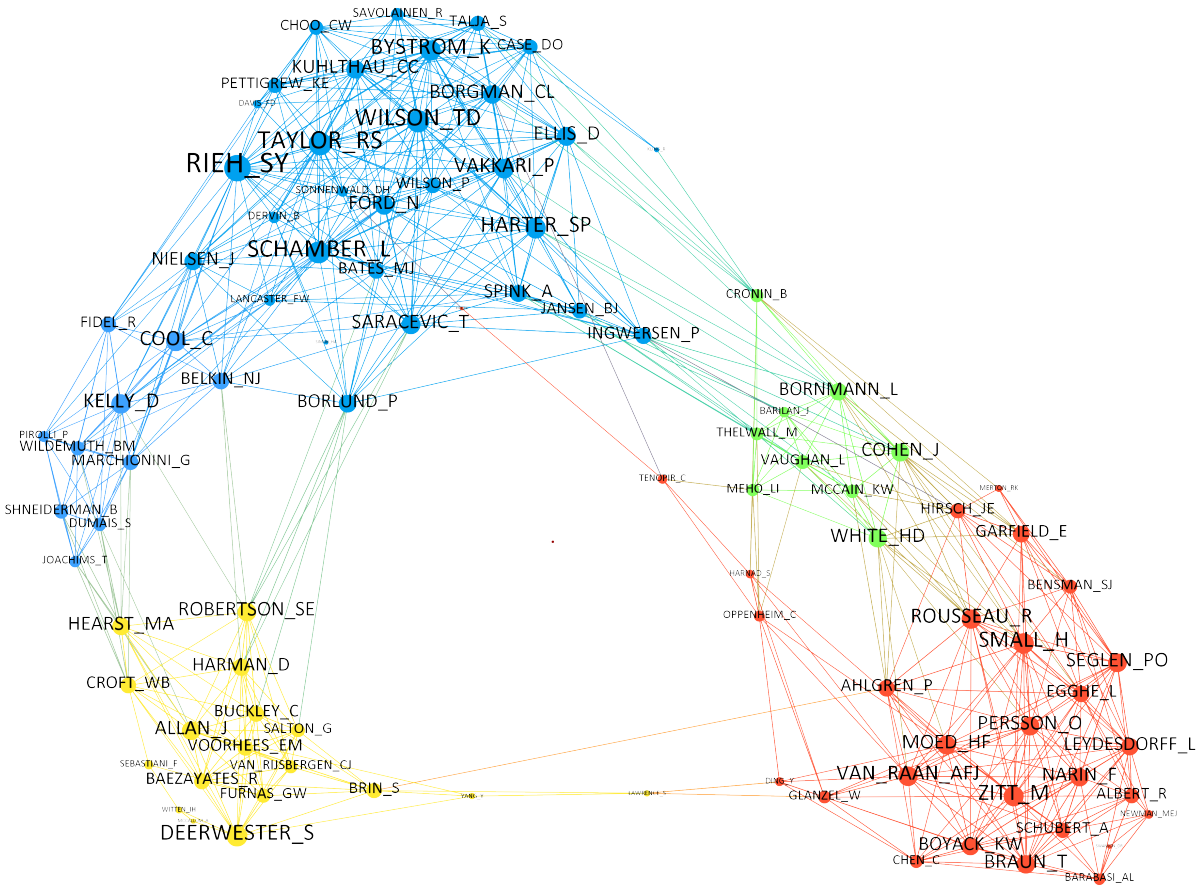

Figure 4. Word2Vec-based Author Network

Table 3 shows the statistics of above two author networks. The larger *graph density* and *clustering coefficient* of the frequency based author network indicated that the authors in this network are densely connected. It is more difficult to capture the sub-research area by the frequency based author network than by the Word2Vec-based author network. In the Word2Vec-based author network, the large *average path length* means that the network topology is sparsely connected. Thus, it supports that the core research areas are well separated than ones by the traditional approach.

	Frequency-based Author Network	Word2Vec-based Author Network
No. of nodes	100	100
No. of edges	1425	633
Graph density	0.294	0.136

Clustering coefficient	0.750	0.612
Average path length	1.986	2.917

Table 3. The Statistics of Author Network

Another difference between two maps is the change of influential authors in the author map. This difference is caused by incorporating topical similarity into building the author map. It implies that the authors linked by citation are topically grouped in the Word2Vec-based author network. For example, Deerwester, introduced latent semantic indexing, had a small portion in Figure 3, whereas he was represented as a relatively bigger node in Figure 4. Thus, we claim that the proposed approach reveals the more influential authors in terms of topicality than the traditional approach.

4 Conclusion and Future Work

This study introduces a new approach to author co-citation analysis by proposing the content-based similarity measure. Unlike other bibliometric studies, we use citing sentences to reflect topical relatedness of authors. In the present paper, we adopt Word2Vec as the measure of author similarity. We also conducted in-depth network analysis of author maps.

We claim that our approach is more appropriate to capture the sub-disciplines and reveals the important authors in perspective of topical influence. Although the dataset is limited to *JASIST*, our method can be applied to other disciplines. A more in-depth understanding and the specific structure of a discipline, enabled by our method, helps advance ACA research.

As a follow-up study, we plan to extend network analysis of all cited authors and construct the author map with all cited authors. We also plan to identify the relationships between authors by analyzing their contribution sources and conduct various statistical analyses such as factor analysis to verify our results.

5 References

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM*, *8*, 361-362.

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.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hyper textual web search engine. In *Proceedings of the seventh international conference on World Wide Web*, 107-117.

Chen, C., Ibekwe-SanJuan, F., & Hou, 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.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, *60*(11), 2229-2243.

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.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Van Eck, N. J., & Waltman, L. (2008). Appropriate similarity measures for author co-citation analysis. *Journal of the American Society for Information Science and Technology*, 59(10), 1653-1661.

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

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.

Wolf, L., Hanani, Y., Bar, K., & Dershowitz, N. (2014). Joint word2vec networks for bilingual semantic representations. *International Journal of Computational Linguistics and Applications*, 5(1), 27-44.

Zizi, M., & Beaudouin-Lafon, M. (1994, September). Accessing hyperdocuments through interactive dynamic maps. In *Proceedings of the 1994 ACM European conference on Hypermedia technology* (pp. 126-135). ACM.