Exploring the leading authors and journals in major topics by citation sentences and topic modeling

Ha Jin Kim[1], Juyoung An[1], Yoo Kyung Jeong[1], Min Song[1]

[1]Department of Library and Information Science, Yonsei University, 50 Yonsei-ro,
Seodaemun-gu, Seoul, South Korea
{hajin_228, anjy, yk.jeong, min.song}@yonsei.ac.kr

Abstract. Citation plays an important role in understanding the knowledge sharing among scholars. Citation sentences embed useful contents that signify the influence of cited authors on shared ideas, and express own opinion of citing authors to others' articles. The purpose of the study is to provide a new lens to analyze the topical relationship embedded in the citation sentences in an integrated manner. To this end, we extract citation sentences from full-text articles in the field of Oncology. In addition, we adopt Author-Journal-Topic (AJT) model to take both authors and journals into consideration of topic analysis. For the study, we collect the 6,360 full-text articles from PubMed Central and select the top 15 journals on Oncology. By applying AJT model, we identify what the major topics are shared among researchers in Oncology and which authors and journal lead the idea exchange in sub-disciplines of Oncology.

Keywords: text mining; citation analysis; topic modelling; bibliometrics

1 Introduction

As the size of data on the web continues to increase in an exponential manner, finding valuable meaning between data becomes of paramount importance in many research areas. In the information science field, citations are challenging, pivotal materials to discover the relationship between academic documents because citations present the description of authors' ideas and the hidden relationship between authors and documents. The earliest works focused mainly on classifying the citation behaviors and discovering the citation reasons with limited data such as the location of citation sentences and the number of references [1,2].

Since the mid-1990s, with the development of computer technology, citation content analysis was elaborated by applying data analysis techniques like text-mining or natural language processing. Zhang et al. [3] present citation analysis based on sematic and syntactic approaches. Semantic-based citation analysis is performed by qualitative analysis to discover the citation motivation and citation classification. On the other hand, syntactic-based citation analysis can be conducted by citation location and citation frequency, which reveals the hidden relation of authors by using meta-data of documents such as journal, venue of publication, affiliation of authors, etc. Following their study, Ding et al. [4] propose a theoretical methodology through content citation analysis. However, these analyses are somewhat limited to the explicit context that primarily represents their own ideas and arguments.

The main goal of the paper is to discover the implicit topical relationships buried in citation sentences by utilizing the citation information from the author's perspective of sharing other authors' point of view. Implicitness of the topical relationship is realized by using citation sentences as the input for the topic modeling technique. In this study, a citation sentence indicates the sentence including citation expression consisting of year and author of the cited work. In general, the citation sentence contains brief content of cited work and opinion that the author of citing work on the cited work. We claim that citation sentences reveal interesting characteristics of scholarly communication such as influence, idea exchange, justification for citer's arguments, etc. We assume that using citation sentences for topic analysis reveals aforementioned characteristics. To explore such intellectual space created by citation sentences, we take both authors and journals into consideration of topic analysis. To this end, we applied Author-Conference-Topic (ACT) model proposed by Tang et al. [5] for our topic analysis in relation with both authors and journals, which is called Author-Journal-Topic (AJT) topic model. ACT model is a probabilistic topic model for simultaneously extracting topics of papers, authors, and conferences. There are a few studies to analyze content of citation sentences. Most of previous studies focus on how the topic of document influences citation and vice versa [6,7,8] using Topic Modeling. Kataria, Mitra, and Bhatia [8] adapt citation to Author-Topic model [9] with the assumption that the context surrounding the citation anchor could be used to get topical information about the cited authors. These studies including Tang et al. [10]'s ACT model are the examples of combining topic modelling methods and citation content analysis. However, most previous studies used metadata of documents. In this work, we focus on identifying the landscape of the oncology field from a perspective of citation. By using citation sentences, our results can indicate which authors are actively cited and which journals lead a certain topic.

The rest of the paper is organized as follows: Section 2 describes the proposed approach. Section 3 analyzes the topic modeling results. Section 4 concludes the paper with the future work.

2 Methodology

2.1 Main idea

The basic assumption of the proposed approach is that citation sentences embed useful contents signifying the influence of cited authors on shared ideas of citing authors. Citation sentences are also considered as an invisible intellectual place for idea exchanging since citations are effective means of supporting and expressing their own arguments by using other works. In the similar vein, Di Marco and Mercer [11] claim that citation sentences play a major role in creating the relationship among relevant authors within the similar research fields. With these assumptions, we are to explore the implicitness of topic relationships resided in citation sentences from the integrated perspective by incorporating the citing authors and journal titles into interpreting the topical relationships.

As shown in Figure 1, we utilized various features including citing authors, citing sentences and journal titles for topic analysis. Authors in Figure 1 mean the citing authors who write a paper and who cite other's work. Citation sentences are the sentences written by the authors when they cite other's work in the paper, and journal titles are the journal names publishing the citing authors' paper. By employing AJT model with these three parameters , we can discover which topics are the most salient ones referred to frequently by researchers and who are the leading authors sharing other authors' ideas in the research field and which journal leads such endeavor.

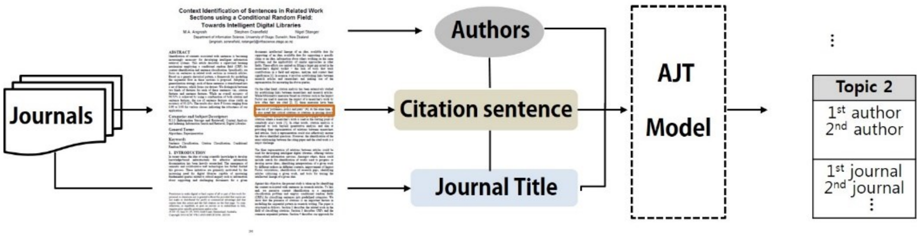


Fig. 1. Three parameters for AJT model

2.2 Data collection

For this study, we compile the dataset on the field of Oncology from PubMed Central that provides the full-text in the biomedical field. We select top 15 journals of Oncology by Thomson Reuter's JCR and journal's impact factor, and from these 15 journals, we are able to collect 6,360 full-text articles.

2.3 Method

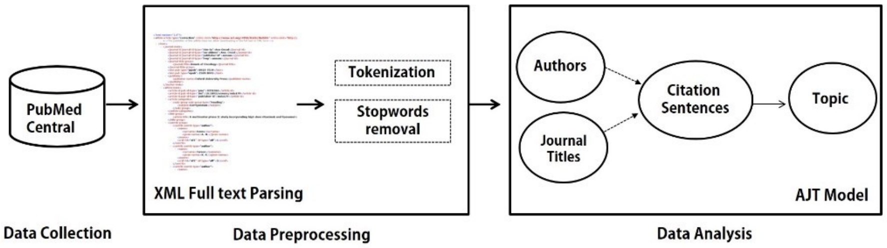


Fig. 2. Workflow

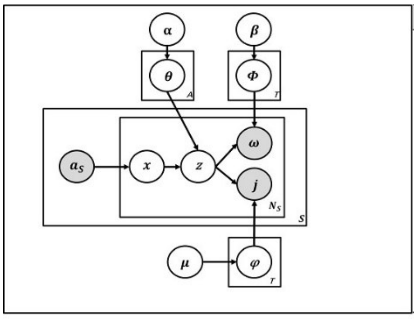
Figure 2 describes the workflow of our study. As mentioned earlier, with the full-text articles collected from PubMed Central, we extract the citation sentences. Most citation sentences are kept in the following format: (author, year), (reference number) [reference number]. An example of such format is "(<xref rid="bib00" ref-

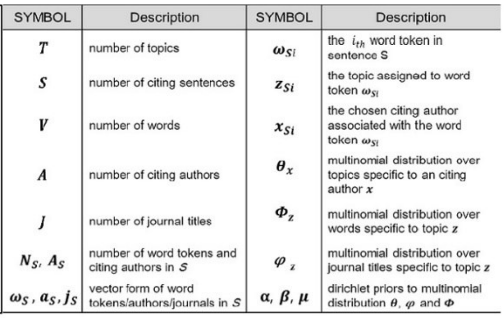
type="bibr">Author name, 2000</xref>)". We use the regular expression technique to parse and extract the citation sentences, when the tag <xref rid=>, </xref> appears on the sentences after parsing XML records with the Java-based SAX parser.

We also parse other metadata for AJT model such as the name of authors and journal titles. The author tags, <surname> </surname> and <given-names></given-names> inside the <contrib-gourp></contrib-group>, denote the list of authors who wrote the paper. For journal, we extract the titles when the journal tags, <journal-title> and </journal-title>, are included in the tag of <journal-meta> and </journal-meta>. We also preprocess extracted sentences by removing both functional and general words and applying the Porter's stemming algorithm to improve the input for AJT Model.

2.4 AJT Model

For our study, we apply ACT [10] model with several metadata such as citation sentences, journal titles and citing authors to develop AJT model. Our AJT model utilizes journal titles and citation sentences instead of conference and abstract on documents. The change of model is needed to analyze most influential topics in Oncology and to find leading authors who frequently mention the active topics and to detect the journals involved in such topics.





SYMBOL	Description	SYMBOL	Description
T	number of topics	ω_{Si}	the i_{th} word token in sentence S
S	number of citing sentences	z_{Si}	the topic assigned to word token ω_{Si}
V	number of words	x_{Si}	the chosen citing author associated with the word token ω_{Si}
A	number of citing authors	θ_x	multinomial distribution over topics specific to an citing author x
J	number of journal titles	Φ_z	multinomial distribution over words specific to topic z
N_S, A_S	number of word tokens and citing authors in S	φ_z	multinomial distribution over journal titles specific to topic z
ω_S, a_S, j_S	vector form of word tokens/authors/journals in S	α, β, μ	dirichlet priors to multinomial distribution θ, φ and Φ

Fig. 3. Graphical representation and Notions of AJT model, which applies ACT model (Applied Tang, J., Jin, R., & Zhang, J., 2008, p.1056, Figure 1, Table 1)

Like ACT model, AJT model assumes that each citing author is related to distribution over topics and each word in citation sentences is derived from a topic. In the AJT model, the journal titles are related to each word. To determine a word (ω_Si) in citation sentences (S), citing authors (x_Si) are consider for a word. Each citing author is associated with a distributed topic. A topic is generated from the citing author-topic distribution. The words and journal titles are generated from a specific topic. AJT model presents (1) the distribution θ of A citing authors-topics, the distribution of ∅ of T topic-words, and the distribution φ of T topic-journal titles; and (2) the following topic z_Si and citing author x_Si for each word ω_Si. The detailed descriptions of the algorithm are provided in the Tang et al.'s paper [5,10].

3 Results and Analyses

For AJT model, we set the number of topics to 15 and finally select 8 topics as major topics. Since we discovered that there are similar topics on our results, we calculated the similarity between 15 topics to select the most representative topics. The topical similarities are measured by each word on topics and we calculated the similarities of two topics where each topic are represented in an array of a term vector. Through this process, we chose 8 topics which have high topical similarities (over 0.5). Each topic presents top 5 words from topic-word distribution, and 5 most related authors and journal titles are displayed along with each topic. By performing several times on the pilot studies, we decided to choose top 5 words which are quite appropriate to describe each topics.

The results of AJT-based topic modeling is shown in Table 1. We label topic 1 "breast cancer" whose top words include breast, expression women, and growth. Since the dataset is compiled with citation sentences, it implies that the topic "breast cancer" is a popular topic where researchers share and exchange ideas and facts related to breast cancer. In relation to the topic "breast cancer", the active authors of breast cancer are Johnston Stephen RD, Colditz Graham A, and Sternlicht Mark D, and they share ideas with others on breast cancer from our results. In terms of journals that provide a common place for idea sharing and communication, the journal "Breast Cancer Research" is the top journal of topic 1, and its impact factor is 5.49. Authors such as Kurzrock Razelle, and Axelrod Haley in group 4 are the leading researchers sharing ideas on the topic "targeted therapy." The topic 4 is associated with the targeted therapy represented by words like mutations, treatments, therapy and disease. The two most influential journals in topic 4 are "Oncotarget" and "Journal of Thoracic Oncology" whose impact factors are 6.36 and 5.28 respectively, which indicates that these two journals are the major journals encouraging authors to share ideas and collaborate with each other on cancer targeted therapy subject area. Authors like Zitgel Laurence, Galluzzi Lorenzo, and Kroemer Guido in the author group 7 are the ones that actively share ideas about the topic "Cancer Immunology." Top concepts that are related to this topic are cell, immune, clinical and antitumor. The top journal of the topic "Cancer Immunology" is Oncoinmmunology whose impact factor is 6.266. Romagnani Paola and Salem Husein K in topic 8 "Stem Cell" are the authors that communicate and share ideas actively with each other in the given field, and the journal "Stem Cells" (impact factor: 6.523) is the leading journal.

Table 1. The Results of AJT-based Topic Modeling in Oncology

Topic1	**Topic2**	**Topic3**	**Topic4**
Breast cancer	Cancer epigenetics	Leukemia	Targeted therapy
breast	methylation	expression	mutations
expression	DNA	mutations	clinical
mammary	expression	AML	treatment
risk	gene	treatment	survival

women	histone	leukemia	resistance
Author group1	**Author group2**	**Author group3**	**Author group4**
Johnston Stephen RD	Gray Steven G.	Tefferi A	Muller Patricia AJ
Colditz Graham A	Mahlknecht Ulrich	Anderson K C	Vousden Karen H
Sternlicht Mark D	Tollefsbol Trygve O.	Ratajczak Janina	Zaravinos Apostolos
Reis-Filho Jorge S	Lichtenstein Anatoly V	Schöffski P	Dienstmann Rodrigo
Esteva Francisco J	Williams David E	Gjertsen B T	Shtivelman Emma
Journal group1	**Journal group2**	**Journal group3**	**Journal group4**
Breast Cancer Research	Clinical Epigenetics	Leukemia	Oncotarget
Annals of Oncology	Oncoimmunology	Pigment Cell & Melanoma Research	Journal of Thoracic Oncology
Cancer Cell	JNCI	Annals of Oncology	Annals of Oncology Cancer Cell
Clinical Epigenetics	Molecular Cancer	Cancer Cell	Clinical Epigenetics
JNCI	Annals of Oncology	Breast Cancer Research	Oncoimmunology

Topic5	**Topic6**	**Topic7**	**Topic8**
Molecular cancer	Oncogene pathway	Cancer Immunology	Stem cell
expression	cell	cell	stem
p53	activity	immune	expression
mutant	activation	expression	differentiation
gene	protein	clinical	MSCs
survival	apoptosis	responses	growth
Author group5	**Author group6**	**Author group7**	**Author group8**
Clarke Paul A	Melino Gerry	Zitvogel Laurence	Romagnani Paola
Workman Paul	Martelli Alberto M	Galluzzi Lorenzo	Salem Husein K
Hoelder Swen	McCubrey James A	Kroemer Guido	Thiemermann Chris
Akhavan David	Blagosklonny Mikhail V	Eggermont Alexander	Lako Majlinda
Cassidy Liam D	Steelman Linda S	Vacchelli Erika	Mellough Carla B
Journal group5	**Journal group6**	**Journal group7**	**Journal group8**
Cancer Cell	Oncotarget	Oncoimmunology	Stem Cells
Neuro-Oncology	Annals of Oncology	Annals of Oncology	Annals of Oncology
Oncotarget	Cancer Cell	Breast Cancer Research	Cancer Cell
Molecular Oncology	Clinical Epigenetics	Cancer Cell	Clinical Epigenetics
Molecular Cancer	Oncogene	Clinical Epigenetics	Molecular Cancer

We visualize topic keywords obtained from results of AJT-based topic model. We construct the co-occurrence network and analyze which topic words play an important role in this domain. Each node in the network represents a topic word, and an edge represents a co-occurrence frequency between keywords. The size of nodes represents degree centrality and the color means network clusters obtained by using modularity algorithm. This network consists of 100 nodes and 1,436 edges. As shown in Figure 4, each topic belongs to a specific community, but shares some important topic keywords. Especially, the topic words positioned at the center is represented core-keywords in Oncology. Figure 4 indicates that these words are the essential concepts of the Oncology domain. Along with the results of AJT-based topic models, we can infer the major journals and authors develop their own research area based on these core-concepts.

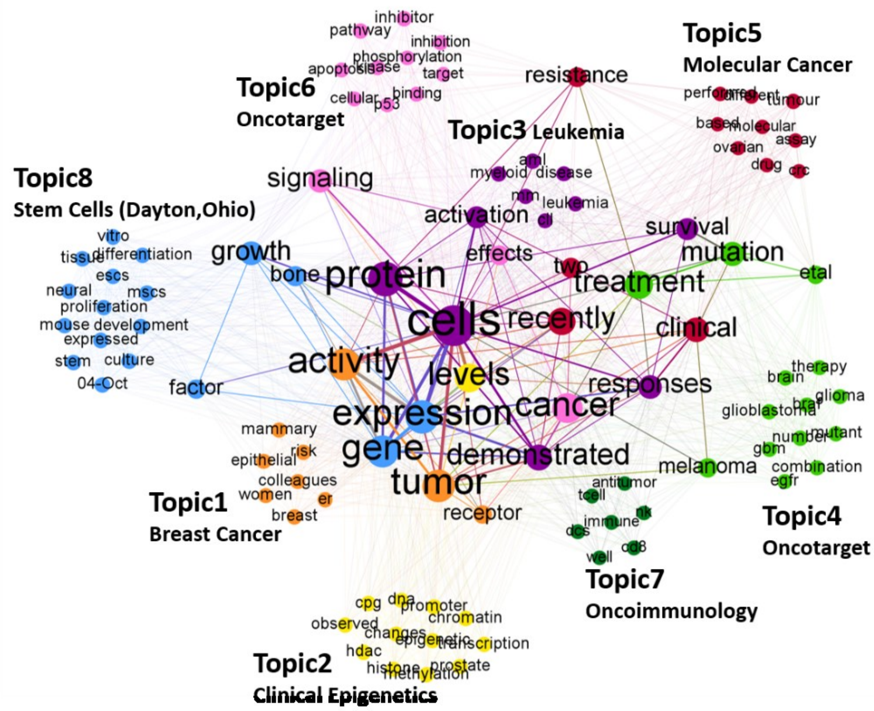


Fig. 4. Network of topic keywords

The above results imply that the proposed approach identifies which topics are frequently shared, who facilitates to exchange ideas, and which journals provide a placeholder for it. Identification of the triple relationship among authors, journals, and topics sheds new insight on understanding the well-discussed topics driven by the leading journals and authors that play a mediator role in the development of Oncology.

4 Conclusion

One of the major research problems in bibliometrics is how to map out the intellectual structure of a research field. The proposed approach tackles such research problem by utilizing citation sentences and AJT model. By using citation sentences as the input for AJT model to find latent meaning, AJT model suggests a new way to detect leading authors and journals in sub-disciplines represented by discovered topics in a certain field. Achieving this is not feasible by traditional frequency-based citation analysis.

One of the interesting observations is that the top-ranked journals in the discovered topics derived from AJT model are not ranked top in terms of JCR. For example, the "Oncotarget" journal is the top-ranked journal in three topics in our analysis, but the ranking of the journal is 20 according to JCR. Since we only report on preliminary results of our approach, we undertake in-depth analysis to investigate why this difference exists. We also conduct various statistical tests on the results. Based on the reported results in this paper, though, we claim that AJT can be used for discovering latent meaning associated citation sentences and the major players leading the field.

As a follow-up study, we will conduct a comparative study that compares the proposed approach with the general topic modeling technique such as LDA. We also plan to investigate whether there is a different impact of using citation sentences and general meta-data such as abstract and title for topic analysis on facilitating idea sharing and scholarly communication. In addition, we would like to consider the window size of citation sentences enriching citation context and to discover the authors' relationships among the neighboring citation sentences.

5 Reference

1. Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. Science, 122(3159): 108–111. doi: 10.1126/ science.122.3159.108.
2. Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. Social studies of science, 5(1), 86-92.
3. Zhang, G., Ding, Y., and 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.
4. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., and 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.
5. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 990-998). ACM.
6. Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In Proceedings of the 24th international conference on Machine learning (pp. 233-240). ACM.
7. Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 542-550). ACM.
8. Kataria, S., Mitra, P., & Bhatia, S. (2010). Utilizing Context in Generative Bayesian Models for Linked Corpus. In AAAI (Vol. 10, p. 1)
9. Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 306-315). ACM.

10. Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 1055-1060). IEEE.
11. 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.