A Trend Analysis of Radiological Research in Korea using Topic Modeling

Dong-Hee Hong*

Department of Radiological Science, Shinhan University

Received: June 08, 2022. Revised: June 24, 2022. Accepted: June 30, 2022.

ABSTRACT

We intend to use topic modeling to identify radiation-themed papers published from 1989 to 2022 and analyze the relevance and weight between topics. This study analyzed topics derived from national subjects for 717 papers published until recently in 2022 to contribute to the revitalization of research in the field of radiation. Through text mining, overall research trends on the subject distribution of the study were analyzed, and five topics were derived through topic modeling.

First, among the papers to be analyzed, a total of 1,675 words were frequency-analyzed through the preprocessing process of key words in a total of 717 papers centered on keywords. Second, as a result of analyzing topics based on the association of constituent words for five topics, it was found that studies focused on minimizing dose in the range that does not degrade image quality in the fields of radiation, image, CT clinical. In addition, it was found that various studies were mainly conducted in the MRI, and the study of ultrasound in various areas of disease analysis was actively attempted.

Keywords: Radiology, Topic modeling, Reaserch trend analysis, LDA(Latent Dirichlet Allocation), Text mining

I. INTRODUCTION

With the recent development of information and communication technology, research using big data is being actively conducted. Text mining methods are used to identify the intellectual structure of the academic using text collected through academic journals or documents in various fields and to examine research trends[1-3]. In Korea, large amounts of literature information are explored and analyzed through text mining in fields such as social science, science and engineering, and health care[4].

Text mining is one of the research areas of data mining, and it is a useful means of grasping the academic development and characteristics of the field by extracting meaningful information from large amounts of literature information and analyzing the distribution of literature among literature[5].

In general, the process of text mining goes through the process of unstructured data collection, data preprocessing, information extraction, and information analysis. Among the topic modeling for information analysis, the Latent dirichlet allocation (LDA) technique can be used to analyze many documents and can effectively analyze other words with the same meaning, the same speech sound, but different words according to context. In addition, the extracted words are easy to determine the topic because the independence between the topics is remarkable. In the collected documents, the weight of each topic can be determined by statistical figures indicating how important a particular topic is[6].

* Corresponding Author: Dong-Hee Hong E-mail: hansound2@hanmail.net Tel: +82-31-870-3415
Address: 95 Hoam-ro, Uijeongbu, Geonggi 11644 Republic of Korea

Data analysis methods to identify research trends can be selected according to research purposes and methods such as statistical analysis, quantitative analysis, expert discussion, and survey, but in this study, Topic modeling which are text analysis and statistical analysis, were applied to the entire document set. Topic modeling is one of the data mining techniques for discovering topics that the entire document means through stochastic calculations of words contained in individual documents. In addition to the representative Latent Dirichlet Allocation(LDA) model, Dynamic Topic Model(DTM) model, Author Topic Model(ATM) model, Pachinko Allocation Model(PAM) model, Topics Over Time(TOT) model were also displayed[7]. The LDA model determines the document-topic-specific distribution value and topic-word-specific distribution value in advance, and then observes and repeats words to stochastically infer which topic the word should belong to.

Therefore, this study aims to examine the future direction of the radiation field by analyzing research trends and academic trends for new academic development and pursuit of clinical value in the radiation field.

To this end, we intend to use topic modeling to identify radiation-themed papers published from 1989 to 2022 and analyze the relevance and weight between topics. The research results will serve as basic data for establishing academic clinical directions in the field of radiation in the future.

II. RESEARCH METHOD

1. Data collection

Academic research provision services include electronic document search services at the National Assembly Library, RISS at the Korea Educational Information Institute, KISS at Korea Academic Information Co., Ltd., and DBpia at Nuri Media, among which RISS has the most papers in Korea. Therefore, this study selected the RISS of the Korea Educational Information Service as a site for data collection, and data was collected with an Excel file in the form of Text through the bibliographic information export function on the website. There are a total of 717 papers collected with radiation keywords, and papers published in a total of 17 academic journals, including the Korean Radiation Science Association, the Korean Radiation Association, and the Korean Radiation Defense Association, were collected.

2. Data Processing

After storing the collected data in the form of a text document, the data preprocessing process was performed to accurately derive the analysis results.

For the collected data, nouns were extracted by performing morpheme analysis. Morpheme analysis was performed by applying the NIADic dictionary, a morpheme analysis dictionary of the KoNLP package provided for R language by the Korea Information Society Agency (NIA), and converted each paper abstract to consist of only extracted common nouns.

For the processing of non-verbal words, the top 3,000 words of frequency were extracted by checking the frequency of words using AntConc ver 4.0.11, a representative corpus analysis program. Among them, the general terms 'study', 'result', 'purpose', 'implementation', 'characteristics', 'method', 'significant', 'use', 'conclusion', 'impact', 'type', 'significance', 'analysis', and 'radiation' were deleted from the data. In addition, it went through a Lamma process of designating a synonym as a single word to add words of the same meaning to the frequency.

3. Data Analysis

LDA analysis was conducted on the data in which the preprocessing process was completed to extract topics and the association of the extracted topics was calculated. In order to model topics by applying the

LDA technique, the number of topics must be determined first. In this study, perplexity was used to determine an appropriate number of topics, and after obtaining perplexity from 5 to 25, the number of topics was determined in a section where the difference in perplexity value was minimized. As a result of calculating the perplexity value with the entire paper to be analyzed, the difference value was minimized in 5 topics. Tomoto (Topic modeling tool) GUI ver0.1 was used for LDA analysis. Fig 1 shows the number of topics (K) was set to 5, a level where the results of the classification could be effectively interpreted, and the number of sampling iterations = 1,000 times, α=0.1, β=0.01 were set.

In the analysis process, 20 major constituent words representing 9 topics were extracted, and the name of the topic was automatically generated around the association of constituent words for each topic. The extracted words were visualized using Voyant Tools based on frequency.


Tomoto - C:₩Users₩user₩Desktop₩topic₩lda.tmt

File About...

Model Info | Topic Result | Document Result

#0 연구, the, 영상	#1 방사선, Radiation	#2 방사선, on the	#3 CT, 선량	#4 초음파, 분석, 성
연구/NNG	방사선/NNG	방사선/NNG	CT/SL	초음파/NNG
the/SL	Radiation/SL	the/SL	선량/NNG	분석/NNG
영상/NNG	on/SL	on/SL	in/SL	성/XSN
Study/SL	in/SL	in/SL	Dose/SL	검사/NNG
on/SL	the/SL	Radiation/SL	the/SL	이용/NNG
A/SL	평가/NNG	and/SL	and/SL	영상/NNG
이용/NNG	for/SL	관하/VV	영상/NNG	Analysis/SL
in/SL	선량/NNG	for/SL	Image/SL	and/SL
D/SL	치료/NNG	Study/SL	검사/NNG	유용/NNG
공명/NNG	영향/NNG	Medical/SL	평가/NNG	진단/NNG
and/SL	Effect/SL	A/SL	to/SL	환자/NNG
자기/NNG	미치/VV	연구/NNG	촬영/NNG	in/SL
관하/VV	Study/SL	Analysis/SL	Quality/SL	for/SL
비교/NNG	and/SL	Radiological/SL	유방/NNG	the/SL
초음파/NNG	X/SL	의료/NNG	Evaluation/SL	CT/SL
MRI/SL	Dose/SL	분석/NNG	for/SL	질환/NNG
평가/NNG	선/NNG	Radiology/SL	측정/NNG	Ultrasonography/SL
성/XSN	연구/NNG	Students/SL	방사선/NNG	MRI/SL
T/SL	효과/NNG	인식/NNG	변화/NNG	on/SL
검사/NNG	조사/NNG	중심/NNG	이용/NNG	Ultrasound/SL

Fig. 1. Tomoto GUI for topic modeling(LDA).

4. Topic Modeling (LDA)

Topic modeling is an algorithm that extracts specific topics from vast amounts of textual data. Typical topic modeling algorithms include Latent semantic analysis and Latent Dirichlet Allocation

There is Blei et al.(2003) proposed LDA (Latent Dirichlet Allocation), an algorithm that can check the topic of documents based on probability techniques[8]. LDA is a probability model that potentially assumes a topic within a given document, generating random documents by iterating the sampling process of probabilistic selection of the topic that constitutes the document, given the probability distribution θ and z of the words that make up each topic[8].

LDA estimates z and θ using parameter values of α, the pre-probability distribution of topics within a given document and predefined document, and β, the pre-probability distribution of words within the topic.Blei (2012) derived topic models using LDA techniques for the 'Science' and 'Yale Law' journals[9], and Keeheon et al. (2015) analyzed research trends in biomedical fields[10], and LDA techniques are often used to analyze technologies and research trends in various fields[11-13].

III. RESULT

1. Keyword Frequency Analysis

In the study of 717 papers in the field of radiation until 2022, a total of 1,675 words were frequency -analyzed through the preprocessing process of key words. Table 1 shows the results of presenting 30 keywords in the order of high frequency. The dose was most commonly used at 94, Ultrasound(US) 49, Computed tomography(CT) 43, development 20, patient 19, digital, abdomen 17, chest 15, image quality 14, breast 13, Magnetic resonance image(MRI), irradiation, contrast medium 12, Positron emission computed tomography(PET), radiologist, effect 11, brain, Monte

Carlo, spatial dose distribution, coronary artery, radiation, low contrast medium 8, blood vessel, protection effect, recognition degree 7, liver, young child, video analysis, phantom 6. Fig. 2 shows the top 700 keywords were visualized using Voyant, which placed the higher the frequency of the keywords.

Table 1. Frequency of keywords

Keyword	Frequency	Keyword	Frequency
Dose	94	Effect	11
US	49	Brain	9
CT	43	Montecalro	9
Develope	20	Shielding	9
Patient	19	Spatial dose distribution	8
Digital	17	Coronary artery	8
Abdomen	17	Radiation therapy	8
Chest	15	Low dose	8
Image quality	14	Cerebrovascular	7
Breast	13	Protective effect	7
MRI	12	Wareness	7
Exposure	12	Liver	6
Contrast	12	Infant	6
PET	11	Image analysis	6
Radiologist	11	Phantom	6

Fig. 2. Keywords in radiation research trend.

2. Keyword Analysis

Table 2 shows the results of analyzing topic modeling to extract research topics and related key vocabulary in the field of radiation. Using an LDA-based topic modeling algorithm, it was presented as five topics, and 10 topics were automatically generated based on the association of topic-specific constituent words, and extracted into 5-7 words by combining the related words.

Topic 1 is radiation and consists of related words such as exposure, dose, effect, impact, irradiation, protection, and medicine. It can be seen that most of these are radiation effects and protection studies.

Topic 2 is an image, and the method, measurement ultrasound, and CT are composed of related words. It can be seen that there are many studies related to image analysis in various clinical fields of radiation.

Topic 3 is an ultrasound examination, and analysis, disease, breast, factors, liver, and patients were composed of related words. It is understood that there are many disease analyses for various local diseases, especially breast and liver, in the field of ultrasound clinical trials of radiation.

Topic 4 is an MRI, and it may be seen that various studies in the field of MRI are attempted by evaluation, device, and magnetic image.

Topic 5 is CT, and dose, image, test, evaluation, and image quality are composed of related words. As it is a device with a high dose in the CT clinical field, it can be seen that research on image quality is being conducted at the same time.

Table 2. Topic and keywords in radiation (LDA)

Cluster	Topic	Words
1	Radiation	Radiation, exposure, dose, effect, effect, investigation, protection, medicine
2	Image	Method, measurement ultrasound, CT
3	Ultrasound	analysis, disease, breast, factor, liver, patient
4	MRI	Assessment, device, magnetic imaging
5	CT	Dose, image, examination, evaluation, quality

IV. DISCUSSION and CONCLUSION

This study analyzed topics derived from national subjects for 717 papers published until recently in 2022 to contribute to the revitalization of research in the field of radiation. Through text mining, overall research trends on the subject distribution of the study were analyzed, and five topics were derived through topic modeling.

First, among the papers to be analyzed, a total of 1,675 words were frequency-analyzed through the preprocessing process of key words in a total of 717 papers centered on keywords. 30 keywords were used most frequently with doses at 94, including 49 ultrasound, 43 CT, 20 development, 19 patients, digital, 17 abdomen, 15 chest, 14 images, 13 breasts, MRI, 12 contrast agents, PET, 11 effects, brain, Monte Carlo, 9 shielding, spatial dose distribution, coronary artery, radiation treatment, 8 low dose, brain vessel, protection effect, 7 recognition degree, liver, young child, video analysis, 6 phantom. As this is a study in the field of radiation, research on dose is being actively conducted, and various studies are actively conducted to reduce dose and research that can expect a protective effect within the range that does not degrade the quality of images. In addition, among the clinical fields, ultrasound and CT MRI are being actively studied, and the abdomen, chest, breast, and liver are considered areas of interest.

Second, as a result of analyzing topics based on the association of constituent words for five topics, it was found that studies focused on minimizing dose in the range that does not degrade image quality in the fields of radiation, image and CT clinical. In addition, it was found that various studies were mainly conducted in the MRI, and the study of ultrasound in various areas of disease analysis was actively attempted.

Topic modeling techniques are being attempted and analyzed in various fields for research trend analysis[14]. However, there is no topic modeling analysis in the field of radiation and health. The purpose of this study was to analyze research trends in the field of clinical radiation and to find out the relevance through topic modeling (LDA). Through the results of this study, it is believed that various keywords will be applied in the future to analyze a wider research topic and continuous research will be actively conducted.

Acknowledgement

This work was supported by the Shinhan University Research Fund, 2021.

Reference

[1] S. Ananiadou, J. Mcnaught, *Text Mining for Biology and Biomedicine*, Artech House Publishers, pp. 135-140, 2005.

[2] R. Feldman, J. Sanger, *The text mining handbook: Advanced approaches in analyzing unstructured data*, Cambridge University Press, pp. 35-40, 2007.

[3] A. Kao, S. R. Poteet, *Natural Language Processing and Text Mining*, Springer-Verlag, pp. 101-110, 2007.

[4] K. W. Cho, S. K. Bae, Y. W. Woo, "Analysis on Topic Trends and Topic Modeling of KSHSM Journal Papers using Text Mining", The Korean Journal of Health Service Management, Vol. 11, No. 4, pp. 213-224, 2017. http://dx.doi.org/10.12811/kshsm.2017.11.4.213

[5] R. N. Kostoff, R. Tshiteya, K. M. Pfeil, J. A. Humenik, G. Karypis, "Power source roadmaps using biliometrics and database tomography", Energy (Oxford, England), Vol. 30, No. 5, pp. 709-730, 2005. http://dx.doi.org/10.1016/j.energy.2004.04.058

[6] H. N. Kim, K. S. Lee, G. S. Jo, "Document classification using weighted associative classifier", Proceedings of the Korean Information Science Society Conference, Vol. 30, No. 2, pp. 154-156, 2003.

[7] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey",

Multimedia Tools and Applications, Vol. 78, pp. 15169-15211, 2019. http://dx.doi.org/10.1007/s11042-018-6894-4

[8] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[9] D. M. Blei, "Probabilistic Topic Models", Communications of the ACM, Vol. 55, No. 4, pp. 77-84, 2012. https://doi.org/10.1145/2133806.2133826

[10] H. G. Eun, S. Min, "Inferring Undiscovered Public Knowledge by Using Text Mining-driven Graph Model", Journal of the Korean Society for information Managemen, Vol. 31, No. 1, pp. 231-250, 2014. http://dx.doi.org/10.3743/KOSIM.2014.31.1.231

[11] B. M. Jeong, T. H. Kim, J. Lee, J. S. Kim, "Twitter Topic Extraction and Topic Category Decision Using LDA Model", Korea Inst. Inf. Sci. Eng. Vol. 12, pp. 787–788, 2015

[12] C. W. Jeong, J. J. Kim, "Analysis of Foresight Keywords in Construction Using Complexity Network Method", Journal of The Korean Digital Architecture Interior Association, Vol. 12, No. 2, pp. 15-23, 2012.

[13] D. H. Park, W. S. Choi, H. J. Kim, S. L. Lee, "Web document classification based on hangeul morpheme and keyword analyses", The KIPS Transactions: PartD, Vol. 19, No. 4, pp. 263-270. 2012. https://doi.org/10.3745/KIPSTD.2012.19D.4.263

[14] D. H. Kim, T. M. Cho, J. H. Lee, "A Domain Adaptive Sentiment Dictionary Construction Method for Domain Sentiment Analysis", Proceedings of the Korean Society of Computer Information Conference, Vol. 23, No. 1, pp. 15-18, 2015.

토픽모델링을 이용한 국내 방사선 학술연구 트렌드 분석

홍동희

신한대학교 방사선학과

요 약

토픽 모델링을 활용하여 1989년부터 2022년까지 출판된 방사선을 주제로 한 논문을 파악하고 주제들 간의 관련성과 비중을 분석하고자 한다. 본 연구는 방사선 분야의 연구 활성화에 기여하기 위하여 2022년 최근까지 출판된 논문 717편을 대상으로 국문제목에서 도출된 토픽들을 분석하였다. 텍스트마이닝을 통해 연구의 주제 분포에 대한 전반적 연구 동향을 분석하였으며, 토픽모델링을 통해 5가지 주제를 도출해냈다.

첫째, 분석 대상 논문 중 키워드 중심으로 총 논문 717편의 연구에서 핵심어를 전처리 과정을 거쳐 최종적으로 선정된 단어는 총 1675개의 단어를 빈도 분석하였다.

둘째, 5개 토픽에 대하여 구성단어의 연관성을 중심으로 토픽을 분석한 결과 방사선, 영상, CT 임상분야에서 영상의 화질을 떨어뜨리지 않는 범위에서 선량을 최소화 하는데 연구가 주를 이루고 있음을 알 수 있었다. 또한, MRI 분야는 다양한 연구가 주를 이루었고 초음파는 다양한 부위의 질환 분석이 연구가 활발하게 시도되고 있음을 알 수 있었다.

중심단어: 방사선, 토픽 모델링, 연구 트렌드 분석, 잠재 디리클레 할당, 텍스트 마이닝

연구자 정보 이력

	성명	소속	직위
(단독저자)	홍동희	신한대학교 방사선학과	조교수