THE JOURNAL OF KOREAN INSTITUTE OF ELECTROMAGNETIC ENGINEERING AND SCIENCE. 2020 October.; 31(10), 839～842.

http://dx.doi.org/10.5515/KJKIEES.2020.31.10.839
ISSN 1226-3133 (Print) · ISSN 2288-226X (Online)

전자파 인체영향 연구논문에 대한 연구형태 자동분류 연구

Research Category Classification for Scientific Literature on Human Health Risk of Electromagnetic Fields

이상우[1] · 권정혁[2] · 김남*[3] · 최형도**[4] · 김의직[5]

Sang-Woo Lee[1] · Jung-Hyok Kwon[2] · Nam Kim*[3] · Hyung-Do Choi**[4] · Eui-Jik Kim[5]

요 약

본 연구는 전자파 인체영향 연구논문 데이터베이스의 활용성과 객관성을 높이기 위한 연구형태 자동분류 기법을 제안한다. 제안하는 기법은 문서 임베딩(embedding) 모델인 Doc2Vec을 사용하여 연구논문의 텍스트 데이터를 벡터로 표현하고, 동물실험, 세포실험, 역학조사에 해당하는지 판단하는 3가지 연구형태 이진분류기를 학습시켜, 전자파 인체영향 연구논문의 연구형태를 자동으로 분류한다. 제안하는 기법의 성능을 검증하기 위하여 전자파 인체영향 관련 연구논문 120개 중 90개를 학습 세트로, 30개를 테스트 세트로 하여 제안하는 기법을 실험하였다. 그 결과, 제안하는 기법이 평균 88 %의 높은 정확도로 전자파 인체영향 연구논문의 연구형태를 분류하는 것을 확인할 수 있었다.

Abstract

This study presents a research category classification method for scientific literature on the human health risk of electromagnetic fields. The proposed method uses the document embedding model, Doc2Vec, to convert the research article text data into a vector, through which it trains three binary classifiers to determine whether the article belongs to animal experiments, cell experiments, or epidemiological studies. Finally, by using the trained binary classifiers, the research articles were automatically classified into one of three research categories. The proposed method was implemented for performance evaluation with 120 research articles on the human health risk of electromagnetic fields. Among the 120 research articles, 90 were used as a training dataset, while 30 were used as a test dataset. The implementation results showed that the proposed method can classify the research articles on the human health risk of electromagnetic fields with 88% accuracy on average.

Key words: EMF Exposure, Document Classification, Machine Learning, Scientific Literature, EMF Databases

「본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신 · 방송 연구개발사업의 일환으로 수행하였음. [2019-0-00102, 복합 전파환경에서의 국민건강 보호기반 구축].」

한림대학교 소프트웨어융합대학(School of Software, Hallym University)
*충북대학교 정보통신공학부(School of Information and Communication Engineering, Chungbuk National University)
**한국전자통신연구원 전파 · 위성연구본부(Radio & Satellite Research Division, Electronics and Telecommunications Research Institute)
1: 석박사통합과정(https://orcid.org/0000-0002-1993-1249), 2: 연구교수(https://orcid.org/0000-0001-6617-6541),
3: 교수(https://orcid.org/0000-0001-8109-2055), 4: 책임연구원(https://orcid.org/0000-0003-2652-7524), 5: 부교수(https://orcid.org/0000-0003-4475-8107)
· Manuscript received July 16, 2020 ; Revised August 27, 2020 ; Accepted September 1, 2020. (ID No. 20200716-003S)
· Corresponding Author: Eui-Jik Kim (e-mail: ejkim32@hallym.ac.kr)

Ⅰ. 서 론

최근, 4차 산업혁명이 진행되며 사람이 일상적으로 접촉하는 전자통신 기기의 수가 급증하였다. 이에 따라, 전자통신 기기가 방출하는 전자파의 인체영향과 안전성을 규명하는 일이 중요한 문제로 대두되었고, 연구자들에 의해 전자파의 인체영향에 대한 수많은 연구와 실험이 수행되었다. 전자파의 인체영향 관련 연구는 연구논문의 형태로 EMF-Portal 및 PubMed와 같은 데이터베이스에 축적되어 왔으며, 이렇게 축적된 전자파 인체영향 연구논문 데이터베이스는 전문가가 전자파의 안전성을 판단하는데 있어서 매우 중요한 자료로 활용되고 있다. 하지만, 연구논문 데이터베이스 활용을 위해서는 전문가가 직접 방대한 양의 연구논문을 검토해야 하기 때문에 긴 시간이 소요되고, 전문가마다 검토한 의견이 다를 수 있다는 문제점이 있다[1].

연구논문 데이터베이스의 활용성과 객관성을 높이기 위한 다양한 연구가 수행되었다. PubTator는 연구논문 데이터베이스의 검토를 돕기 위한 웹기반 검색도구로서, PubMed에 축적된 3천만 건의 연구논문에 대하여 세포, 병명, 종, 약품과 같은 관련 바이오 개체(bioentity)에 대한 주석과 함께 키워드 및 시맨틱 검색을 제공한다[2]. Textpresso Central 또한 웹기반 검색도구로서 키워드 및 연구형태 검색을 제공하며, PubMed Central Open Access Subset 및 WormBase C. elegans bibliography와 같은 연구논문 데이터베이스에서 연구논문 내 전체 텍스트(full-text corpus)에 대한 상세 검색을 제공한다[3]. BioReader는 사용자로부터 키워드, 관심있는 연구논문, 관심없는 연구논문에 대한 정보를 입력받고, PubMed의 키워드 검색결과 중 사용자가 관심이 있을만한 논문 만을 분류하여 제공한다[4]. 하지만, 앞서 언급된 연구의 결과물은 생물학 · 의학의 전반적인 연구논문에 대한 검색을 위해 사용되기 때문에 전자파의 인체영향 및 안전성 평가에 도움이 되는 정보를 제공하지 못한다.

이에, 본 연구는 전자파 인체영향 관련 연구논문 데이터베이스의 활용성과 객관성을 높이기 위한 전자파 인체영향 연구논문의 연구형태 자동분류 기법을 제안한다. 제안하는 기법은 전자파 인체영향 연구논문의 제목 및 초록으로 구성된 텍스트 데이터를 임베딩(embedding)하고, 연구형태 별로 각각 이진분류기를 학습시켜 분류작업을 수행한다. 이때, 연구형태 분류기준은 전자파의 인체영향 및 안전성 평가 관련 전문가의 자문을 받아, 전자파 노출 연구의 가장 일반적인 3가지 연구형태(동물실험, 세포실험, 역학조사)로 선정하였다. 학습된 이진분류기는 연구논문이 해당 연구형태에 포함될 확률을 출력하며, 이진분류기의 출력 값이 0.5 이상이면 연구논문이 해당 연구형태에 속한다고 판단한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 제안하는 기법의 기능적 구조를 설명한다. Ⅲ장에서는 제안하는 기법을 구현하고, 실험을 통해 성능을 검증한다. 마지막으로, Ⅳ장에서는 본 논문의 결론을 내리며 마친다.

Ⅱ. 제안하는 기법의 기능적 구조

그림 1은 제안하는 기법의 기능적 구조를 보여준다. Dataset은 전자파 인체영향 관련 연구논문의 제목과 초록을 포함한 텍스트 데이터로 구성된다. Dataset의 연구논문 텍스트 데이터는 동일한 범주에 속하는 문서를 유사한 값의 벡터로 변환하는 Doc2Vec을 통해 문서의 특징을 나타내는 벡터 값으로 임베딩된다[5].

임베딩된 연구논문 벡터 중 연구형태 정보가 표기된 벡터를 사용하여 3개의 연구형태(동물실험, 세포실험, 역학조사) 이진분류기를 각각 훈련시킨다. 훈련된 이진분류

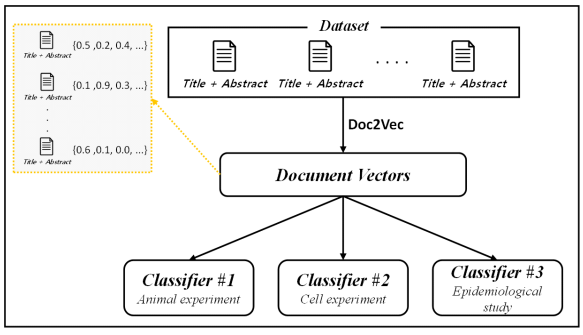

그림 1. 전자파 인체영향 연구논문 연구형태 자동분류 기법의 기능적 구조

Fig. 1. Functional architecture of research category classification for scientific literature on human health risk of electromagnetic fields.

기는 새로운 연구논문 벡터가 입력되면, 연구논문이 해당 연구형태에 포함될 확률을 출력하며, 출력값이 0.5 이상이면 해당 연구형태 분류에 속한다고 판단한다.

Ⅲ. 구현 및 실험

3-1 사용 데이터

그림 2는 제안하는 기법의 구현을 위해서 사용된 전자파 인체영향 관련 연구논문 데이터베이스를 보여준다. 해당 데이터베이스는 전문가가 직접 작성한 연구형태, 연구결과 등과 같은 전자파 인체영향 관련 연구논문의 상세정보와 출판정보를 함께 제공한다.

본 연구에서는 제안하는 기법의 구현 및 실험을 위해서 해당 데이터베이스에 수록된 전자파 인체영향 관련 연구논문 중 동물실험 논문 40개, 세포실험 논문 40개, 역학조사 논문 40개, 총 120개 연구논문의 제목 및 초록 텍스트 데이터를 수집하였다. 이렇게 수집된 120개 연구논문의 텍스트 데이터는 구현과 실험과정에서 이진분류기 훈련 및 제안기법의 성능검증을 위해 사용되었다.

3-2 제안기법 구현

먼저, 전자파 인체영향 연구논문 자동분류를 위해 Dataset의 모든 텍스트 데이터가 다음과 같은 전처리 과정을 거친다. 첫번째로 모든 텍스트 데이터의 알파벳을

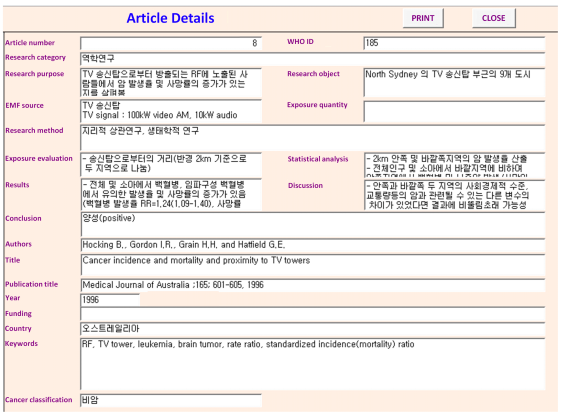

그림 2. 전자파 인체영향 연구논문 데이터베이스
Fig. 2. Scientific literature database for scientific literature on human health risk of electromagnetic fields.

소문자화 한다. 두번째로 텍스트 데이터에서 구두점, 공백, Be동사, 관사 등 의미상 불필요한 텍스트를 제거한다. 세 번째로 표제어 추출(lemmatization)을 수행하여 텍스트 데이터에서 같은 뜻을 내포하고 있는 다양한 형태의 단어를 단일 형태의 명사 혹은 동사로 통일한다. 마지막으로 각 연구논문의 텍스트를 문장 및 단어 단위로 분할한다.

전처리가 끝난 연구논문 텍스트 데이터의 문서번호와 단어는 One-hot Vector로 표현된다. 그리고 같은 문장에서 이웃하게 쓰인 단어들의 One-hot Vector 쌍, 문서번호와 문서에서 사용된 단어 간 One-hot Vector 쌍을 만들어 Doc2Vec 훈련용 데이터를 구성하고, Doc2Vec 모델을 훈련시킨다. 훈련 과정이 끝난 후 Doc2Vec 모델 Hidden Layer의 가중치 행렬이 각 문서 및 단어를 대표하는 벡터로 사용된다. 이러한 텍스트 데이터 전처리 과정 및 Doc2Vec 임베딩은 텍스트정보처리를 위한 Python 라이브러리인 Gensim을 사용하여 수행되었다[6].

임베딩된 연구논문의 벡터를 입력값으로, 데이터베이스에 표시된 연구형태 정보를 목표값으로 하여 3개의 연구형태 이진분류기의 학습을 진행하였다. 앞서 수집한 120개의 연구논문 중 90개를 학습 세트로, 30개를 테스트 세트로 분할하여 실험을 진행하였으며, 여러 알고리즘 중에서 가장 적합한 이진분류 알고리즘을 찾기 위해 3개의 알고리즘(logistic regression, linear support vector machine(SVM), decision tree)을 적용하여 실험을 진행하였다. 이러한 연구논문 이진분류기의 학습 및 검증은 머신러닝을 위한 Python 라이브러리인 scikit-learn을 통해서 수행되었다[7].

표 1은 전자파 인체영향 연구논문 자동분류 기법의 실험결과를 보여준다. 3개의 알고리즘 중에서 Logistic Regression 알고리즘이 가장 높은 성능(F1=0.84, Accuracy=0.88)을 보였으며, Linear SVM이 두번째로 높은 성능(F1=0.79, Accuracy=0.83)을 보였다. 마지막으로 Decision Tree 알고리즘이 가장 낮은 성능(F1=0.59, Accuracy=0.55)을 보였다. 이를 통해 Logistic Regression과 Linear SVM과 같은 선형 알고리즘이 Decision Tree 알고리즘보다 제안하는 기법에 더 적합한 것을 확인할 수 있다. 또한, 이진분류기의 연구형태(동물실험, 세포실험, 역학조사) 별로 비교할 때, Logistic Regression과 Linear SVM 모두 역학조사 이진

표 1. 전자파 인체영향 연구논문 자동분류 실험 결과
Table 1. Evaluation results of research category classification for scientific literature on human health risk of electromagnetic fields.

Algorithm	Classifier	Precision	Recall	F1	Accuracy
Logistic regression	Animal experiment	0.89	0.80	0.84	0.90
	Cell experiment	0.66	0.80	0.72	0.80
	Epidemiological study	0.90	0.90	0.90	0.93
	Average	0.82	0.83	0.82	0.88
Linear support vector machine	Animal experiment	0.72	0.80	0.76	0.83
	Cell experiment	0.61	0.80	0.77	0.76
	Epidemiological study	0.89	0.80	0.84	0.90
	Average	0.74	0.80	0.79	0.83
Decision tree classifier	Animal experiment	0.37	1.00	0.54	0.43
	Cell experiment	0.45	0.90	0.60	0.60
	Epidemiological study	0.47	0.90	0.62	0.63
	Average	0.43	0.93	0.59	0.55

분류에서 가장 높은 성능을 보이고, 동물실험과 세포실험 이진분류에서의 성능은 비교적 낮은 것을 확인할 수 있다. 이는 동물실험과 세포실험 연구논문의 단어사용 패턴이 유사하고, 반면에, 역학조사 연구논문에서의 단어사용 패턴은 동물실험 및 세포실험 연구논문에서의 단어사용 패턴과 뚜렷히 구분되기 때문이다. 또한, 실험 결과와 실제 연구논문의 본문을 검토해본 결과, 비록 데이터베이스에는 동물실험으로 표기가 되었지만, 실제로 동물실험과 세포실험이 함께 병행된 경우도 더러 있었다. 이러한 경우로 인해 세포실험 연구논문 이진분류기의 Precision 항목이 비교적 낮게 도출되었다.

Ⅳ. 결 론

본 논문에서는 전자파 인체영향 관련 연구논문 데이터베이스의 활용성과 객관성을 높이기 위한 전자파 인체영향 연구논문의 연구형태 자동분류 기법을 제안하였다. 제안하는 기법은 전자파 인체영향 연구논문의 텍스트 데이터를 Doc2Vec을 통해 임베딩하고, 동물실험, 세포실험, 역학조사의 3가지 연구형태 이진분류기를 훈련시켰다. 그리고 훈련된 연구형태 이진분류기를 통해 연구논문이 각 연구형태에 포함될 확률을 구하여, 이진분류기의 출력값이 0.5 이상이면 연구논문이 해당 연구형태에 속한다고 판단하였다. 제안하는 기법의 성능을 검증하기 위해 Logistic Regression, Linear SVM, Decision Tree 알고리즘을 이진분류기로 사용하여 실험을 수행하였다. 실험결과에서 가장 높은 성능을 보인 Logistic Regression 알고리즘이 평균 88 %의 높은 정확도를 보였으며, 이를 통해 제안하는 기법의 성능과 효율성을 검증할 수 있었다.

References

[1] E. van Denventer, E. van Rongen, and R. Saunders, "WHO research agenda for radiofrequency fields," *Bioelectromagnetics*, vol. 32, no. 5, pp. 417-421, 2011.

[2] C. H. Wei, H. Y. Kao, and Z. Lu, "PubTator: A web-based text mining tool for assisting biocuration," *Nucleic Acids Research*, vol. 41, no. W1, pp. 518-522, Jul. 2013.

[3] H. M. Müller, K. M. van Auken, Y. Li, and P. W. Sternberg, "Textpresso central: A customizable platform for searching, text mining, viewing, and curating biomedical literature," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1-16, 2018.

[4] C. Simon, K. Davidsen, C. Hansen, E. Seymour, M. B. Barnkob, and L. R. Olsen, "BioReader: A text mining tool for performing classification of biomedical literature," *BMC Bioinformatics,* vol. 19, no. S3, no. 57, pp. 165-170, 2019.

[5] Q. Le, T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the International Conference on Machine Learning*, 2014, pp. 1188-1196.

[6] R. Řehůřek, P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, 2010, pp. 45-50.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. Grisel et al., "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.