전자파 인체영향 관련 연구논문 분류를 위한 Doc2Vec 및 t-SNE 기반 군집분석 연구

이상우[1,o], 이솔비[1], 고하나[1], 이연희[1], 권정혁[2], 김의직[1,*]
[1]한림대학교 소프트웨어융합대학
[2]한림대학교 스마트컴퓨팅연구소
*ejkim32@hallym.ac.kr

I. 서론

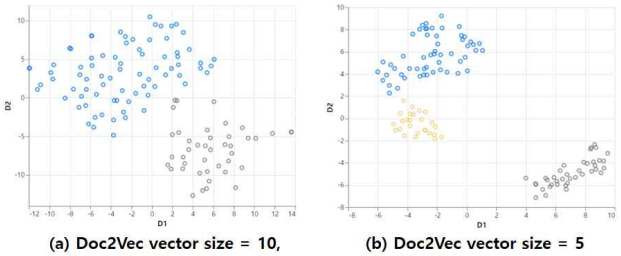
본 논문에서는 전자파 인체영향 관련 연구논문의 자동 분류를 위한 Doc2Vec과 t-Stochastic Neighbor Embedding (t-SNE) 기반의 텍스트 데이터 군집분석 기법을 제안한다[1]. 제안하는 기법은 문서임베딩 기법 Doc2Vec 및 차원축소 기법 t-SNE을 통해 각 연구논문에 포함된 텍스트 데이터를 2차원 데이터 그래프 형태로 시각화한다. 이를 통해 연구논문의 군집을 확인할 수 있다.

II. 본론

전자파 인체영향 관련 연구논문 데이터의 군집분석을 위한 시각화 과정은 총 3가지 과정을 거쳐서 이루어진다. 가장 먼저 각 연구논문을 대표하는 텍스트 파일을 문서 임베딩 기법인 Doc2Vec을 사용하여 여러 개 차원의 특성 값으로 표현한다. 그 다음으로 임베딩 된 각 연구논문의 특성 값 데이터를 t-SNE을 통해 2차원으로 축소한다. 마지막으로 2차원으로 표현된 각 연구논문의 특성 값 데이터는 그래프형태로 시각화 된다. 시각화된 그래프를 이용하여 전자파 인체영향 관련 연구논문의 군집분석을 수행한다.

III. 실험결과

제안기법의 실효성 확인을 위해, PubMed와 EMF-Portal에서 120개의 전자파 인체영향 관련 연구논문을 임의로 선정했으며, 이를 역학조사, 동물실험, 세포실험 논문으로 분류하기 위한 군집분석 실험을 수행하였다. 그림 1은 Doc2Vec의 각기 다른 특성 값의 개수를 정의하여 학습하고 그 결과를 시각화한 그래프를 보여준다. 그림의 각 그래프를 비교했을 때 Doc2Vec의 특성화 값의 개수를 적게 정의하는 것이 좀 더 명확한 군집화 현상을 보임을 확인할 수 있었다. 또한 Doc2Vec 특성 값의 개수를 10개로 정의했을 때 의 그래프에서는 2개의 군집이 관찰되며, 5개로 정의했을 때의 그래프에서는 3개의 군집이 관찰되었다.



(a) Doc2Vec vector size = 10, (b) Doc2Vec vector size = 5

그림 1. Doc2Vec과 t-SNE를 통한 군집분석 시각화 결과

IV. 결론

본 논문에서는 Doc2Vec과 t-SNE를 활용한 전자파 인체영향 관련 연구논문 데이터베이스의 군집분석 기법을 제안하였다. 실험 결과를 통해 제안기법이 전자파 인체영향 관련 연구논문을 세포실험, 동물실험, 역학조사 논문으로 분류할 수 있음을 확인할 수 있었다.

Acknowledgment

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신·방송 연구개발사업의 일환으로 수행하였음 [2019-0-00102, 복합 전파환경에서의 국민건강 보호기반 구축]. 이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2019R1I1A1A01059787).

참고문헌

[1] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in Proc. the 31st International Conference on Machine Learning, pp. 1188-1196, 2014.