Journal of Korea Multimedia Society Vol. 22, No. 2, February 2019(pp. 250-259)
https://doi.org/10.9717/kmms.2019.22.2.250

단어 연관성 가중치를 적용한 연관 문서 추천 방법

김선미[†], 나인섭[††], 신주현[†††]

A Method on Associated Document Recommendation with Word Correlation Weights

Seonmi Kim[†], InSeop Na[††], Juhyun Shin[†††]

ABSTRACT

Big data processing technology and artificial intelligence (AI) are increasingly attracting attention. Natural language processing is an important research area of artificial intelligence. In this paper, we use Korean news articles to extract topic distributions in documents and word distribution vectors in topics through LDA-based Topic Modeling. Then, we use Word2vec to vector words, and generate a weight matrix to derive the relevance SCORE considering the semantic relationship between the words. We propose a way to recommend documents in order of high score.

Key words: Big Data, Datamining, LDA, Word2vec, Topic Modeling, Information Retrieval, Document Recommendation

1. 서 론

4차 산업혁명 시대를 맞이하면서 인공지능(AI), 빅데이터, 사물인터넷(IoT), 로봇 등 다양한 기술들이 주목받고 있다. 빅데이터의 등장으로 인공지능이 본격적으로 시장에 확대되기 시작했고 구글의 인공지능 '알파고'의 바둑 대전을 통해 사람들의 관심도 높아졌다. 인공지능이란 기계가 사람과 유사한 지능을 가지도록 인간의 학습능력, 추론능력, 자연어 이해능력 등을 컴퓨터 프로그램으로 실현하는 기술이다. 인공지능 관련 기술 분야로는 패턴인식, 자연어 처리, 기계 학습(Machine Learning), 데이터마이닝, 시멘틱 웹, 지능 엔진 등이 있다. 데이터 분석 및 처리를 위한 핵심 기술들이 주로 해당되고 인공지능은 중요한 기반 기술로 자리 잡고 있으며 인공지능의 능력을 활용하여 더욱 가치 있는 분석 결과를 창출할 수 있다. 자연어 처리(Natural Language Processing, NLP)는 컴퓨터가 사람처럼 언어를 이해하고 처리할 수 있도록 해주는 인공 지능의 중요한 연구 분야이며 음성 인식, 정보 검색, 문서 자동 분류, 챗봇, 시스템 자동 번역 등 다양하게 응용되고 있다.

정보 검색 기술은 정보 사회를 대표하는 기술이며 검색 결과인 정보의 순위는 사람들에게 영향력을 미치는 것으로 검증되었다[1-2]. 대규모의 정보가 생성되고 있고 정보 과부하 문제로 인해 사람들은 필요한 정보를 찾아내는데 어려움을 겪고 있으며 연관 문서 추천 방법에 대한 다양한 연구가 진행되고 있다[3-5]. 기존의 용어 사전, 온톨로지와 같은 지식 리소스 기반의 연구는 사람의 개입과 구축비용, 유지보수가 필요하다. TF-IDF 같은 단순 빈도수 기반의 연구는

※ Corresponding Author : Juhyun shin, Address: (61452) Pilmun-daero 209, Dong-gu, Gwangju, Korea, TEL : +82-62-230-7162, FAX : +82-62-233-6896, E-mail : jhshinkr@chosun.ac.kr
Receipt date : Dec. 28, 2018, Approval date : Jan. 21, 2019
[†] Dept. of Software Convergence Engineering Chosun University (E-mail : smkim7472@naver.com)
[††] SW Convergence Education Institute, Chosun University (E-mail : ypencil@chosun.ac.kr)
[†††] Dept. of ICT Convergence, Chosun University
※ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1A1A1A05022526).

단어의 의미와 문장에서의 맥락을 고려하지 못하고 새로운 단어에 대한 해석이 불가능하므로 검색의 효율이 떨어진다.

한국 인터넷진흥원(Korea Internet & Security Agency, KISA)에 따르면 모바일 인터넷의 주요 이용 목적은 '커뮤니케이션' 다음으로 '자료/정보 습득'이 높았으며 자료/정보 습득 시 가장 많이 이용하는 방법 중 '뉴스'가 높은 순위를 차지하였다. 뉴스는 다양한 언론사를 통해 보도되고 넓고 방대한 정보 범위를 가지는 것을 특징으로 한다. 다양한 사건을 다루기 때문에 여러 주제를 내포하고 있으며 하나의 주제 속에 매우 다양한 키워드로 이루어져 있으므로 사용자 맞춤형 정보를 제공하는데 한계가 있다.

본 논문에서는 키워드와 연관성이 높은 문서를 자동으로 분류하기 위해 한국어 뉴스 기사를 이용하여 LDA 기반 토픽 모델링을 통해 문서 내 주제 분포와 주제 내 단어 분포를 추출하고 Word2vec을 이용해 단어를 벡터화한 후 가중치 행렬을 생성하여 단어 연관성 가중치를 적용해 연관성 SCORE를 도출한 다음 점수가 높은 순서대로 문서를 추천하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 본 논문의 기본 이론이 되는 LDA와 Word2vec에 대하여 설명하고, 3장에서는 본 논문에서 제안하는 단어 연관성 가중치를 적용한 연관 문서 추천 방법에 대하여 설명한다. 4장에서는 제안한 방법을 적용하여 성능을 평가하고 5장에서 결론에 대하여 기술한다.

2. 관련 연구

2.1 LDA

잠재 디리클레 할당(Latent Dirichlet Alocation, LDA)은 토픽 모델링 기법 중 가장 주목 받고 있으며 비 구조화된 대량의 문서 집합에서 잠재되어있는 주제(토픽)를 추출하여 숨겨진 의미 구조를 발견하기 위한 머신러닝 기법이다[6]. 특정 주제에 관련된 문서에서는 해당 주제에 대한 단어가 다른 단어들에 비해 더 자주 등장할 것이라는 개념을 바탕으로 하며 문서 내에 내포된 주제와 주제의 분포는 문서 내의 단어 통계를 수학적으로 분석하여 알아낸다. LDA는 확률 모델로 여러 주제가 혼합된 문서를 다룰 수 있는 것을 장점으로 한다.

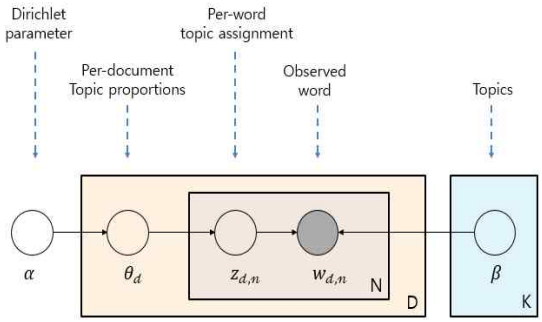

Fig. 1. Graphical model representation of LDA.

LDA 모델은 Fig. 1과 같이 표현되어지며 N은 단어의 개수이고 D는 문서의 개수이며 K는 주제의 개수를 나타낸다. 문서 집합(Corpus)에서 관측된 $W_{d,n}$을 이용하여 Hidden 상태의 θ_d와 β_k를 추론한다. 각 문서들이 갖는 주제 θ를 확률적으로 나타내며 각 토픽에 해당하는 단어들의 확률 분포 z도 나타낼 수 있다. 본 논문에서는 문서 내 주제 분포와 주제 내 단어 분포를 추출하기 위하여 LDA를 사용하였다.

2.2 Word2vec

Word2vec은 인공신경망 기반의 단어 임베딩(Word embedding) 알고리즘으로 빠른 학습 속도와 좋은 성능을 가지고 있다. 단어 임베딩은 딥러닝 분야에서 텍스트를 구성하는 각각의 단어를 수치화하는 방법이며 Word2vec은 문장을 구성하는 단어들의 전후 관계를 인공신경망에 학습시켜 단어의 의미를 내포하여 단어를 벡터 공간에 표현한다[7]. 인공신경망은 인간의 신경세포(neuron)의 구조에 많은 영향을 받았으며 각 입력 값을 받아들일 때 입력 값을 바로 출력하지 않고 일정한 가중치(weight)를 곱해 준다.

Fig. 2는 Word2vec의 두 가지 학습 모델인 Continuous Bag-of-Word(CBOW)모델과 Skip-gram 모델이다. CBOW 모델은 주변에 있는 단어들을 이용하여 대상 단어를 예측하는 방식이고, Skip-gram 모델은 대상 단어로 주변 단어를 예측하는 방식이다[8]. 이와 같이 Word2vec은 단어의 의미와 문장에서의 맥락을 고려하여 단어를 벡터로 표현하기 때문에 의미적으로 유사한 단어들끼리 근접한 벡터 공간에 위치하게 된다. 같은 단어라도 단어의 의미와 맥락에 따라 다른 벡터 공간에 학습될 수 있다는 것을 의미한다. 본 논문에서는 단어 벡터들 간의 거리를 코사

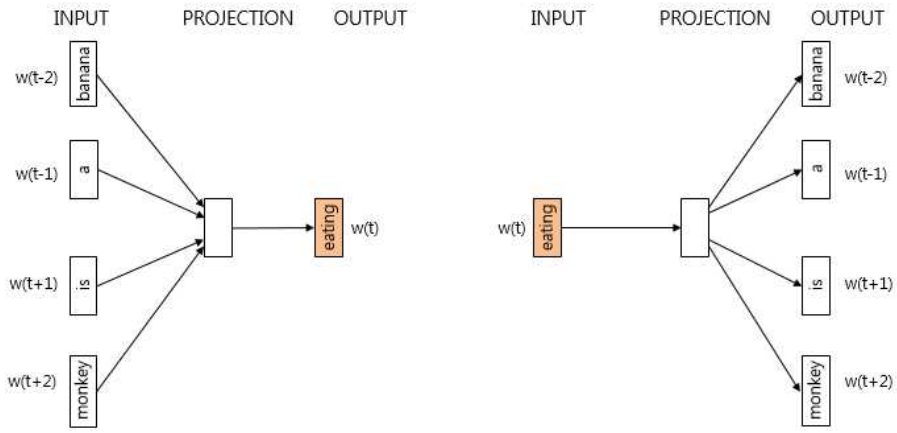

Fig. 2. Examples of (a) CBOW and (b) Skip-gram model configurations.

인 유사도를 통해 계산하여 단어 간 의미적 유사도를 구하기 위해 Word2vec을 사용하였다.

3. 본 론

3.1 시스템 구성도

본 논문에서는 키워드와 연관성이 높은 문서를 자동으로 분류하기 위해 LDA 기반 토픽모델링을 통해 문서 내 주제 분포와 주제 내 단어 분포를 추출하고 Word2vec을 사용하여 단어 간 유사도를 구한다. 두 결과 값을 이용해 가중치 행렬을 생성하고 연관성 SCORE를 도출하여 수치가 높은 순서대로 문서를 추천한다. Fig. 3은 본 논문에서 제안하는 의미 기반 연관 문서 추천 방법의 구성도이다.

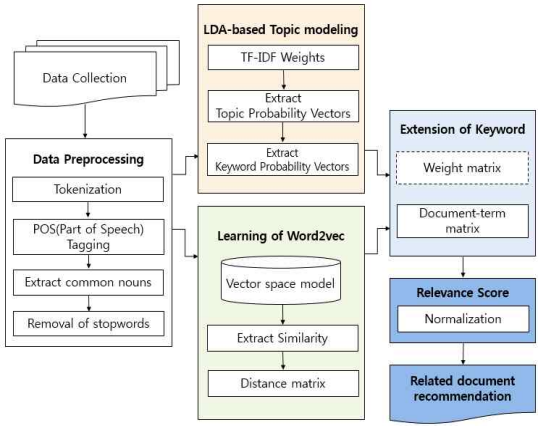

Fig. 3. System configuration diagram.

실험을 위한 데이터는 네이버(www.naver.com)에서 제공하는 정치 카테고리의 뉴스 기사로 선정해 Java 기반 환경에서 웹 크롤링하여 2018.07.01.부터 2018.07.31.까지 총 24,887개의 기사 내용을 수집해 .csv 파일로 저장하였다. R studio에서 한국어 자연어 처리를 위해 KoNLP 라이브러리를 사용하며 전처리 작업으로 문장을 어절 단위로 토큰화(Tokenizing) 시키고 형태소 분석을 통해 단어들의 품사를 판별하여 Pos Tagging 작업을 수행한 후 보통 명사를 추출하고 불용어를 제거한다.

TF-IDF 가중치를 부여하여 LDA 기반 토픽 모델링을 수행하고 문서 내 주제 분포와 주제 내 단어 분포를 추출한다. Word2vec 학습을 통해 VectorSpace model을 구축하고 단어를 벡터화한 후 단어 간 의미적 유사도를 구해 거리 행렬(Distance matrix)을 생성한다.

거리 행렬에서 주제 내 키워드에 해당하는 부분만 추출해 가중치 행렬(Weight matrix)을 생성하여 단어 연관성 가중치를 적용해 키워드를 확장하고 DTM (Document Term Matrix) 행렬과 가중합을 통해 연관성 SCORE를 도출한다. 연관성 SCORE의 범위를 0부터 1구간으로 일치시키기 위하여 정규화 과정을 거친 후 연관성 SCORE 수치가 높은 순서대로 문서를 추천하는 방법을 제안한다.

3.2 TF-IDF 가중치를 적용한 LDA 기반 토픽 모델링

LDA 분석을 위한 작업으로 문서에 나타나는 단어

를 행렬로 표현하는 TDM(Term Document Matrix)을 생성하여 단어가 문서에 몇 회 출현했는지 알 수 있다. TDM은 단어들의 단순 빈도수를 나타내기 때문에 빈도수가 적은 단어들은 중요도가 떨어지므로 어떤 단어가 특정 문서에서 얼마나 중요한지 나타내주는 TF-IDF 가중치를 부여하여 단어 별 TF-IDF 분포 값을 기준으로 TDM의 크기를 조절해 성능을 향상시켜 LDA 기반 토픽 모델링을 수행하였다. 그 결과 '오늘', '이번', '관련', '당시'와 같은 의미가 중요하지 않지만 자주 등장하는 단어를 제거할 수 있었다.

LDA기반 토픽 모델링 결과 총 24,887개의 각각의 뉴스 기사 문서에 대한 주제 분포와 주제 내 단어 분포 벡터를 추출할 수 있었다. 본 논문에서는 매개변수 K를 15으로 지정하였고 총 15개의 주제가 생성되었다. 15개의 주제에 따라 총 15개의 클러스터가 생성되었고 같은 클러스터 내에 있는 문서들은 서로 동일한 주제 범위를 갖는다.

Fig. 4는 각 문서들이 갖는 주제 번호 및 확률 값의 분포를 시각화하였고 Table 1은 문서 내 주제 분포 예시이다. X 좌표는 Topic 번호를 뜻하고 총 15개의 주제를 나타내며 Y 좌표는 15개 각 주제들에 대한 문서들의 확률 값을 의미한다. 하나의 문서가 갖는 각 토픽에 대한 최대 확률 값으로 한 문서 내에 여러 토픽이 내포되어있으며 다양한 확률 분포 값을 가지는 것을 알 수 있다. 사용자 관심문서가 가장 높은 확률 분포를 가지는 주제를 선택하고 해당 주제의 주제어를 키워드로 지정한다. Fig. 5는 Fig. 4에 나타난 15개 각 주제들에 해당하는 상위 단어들의 확률

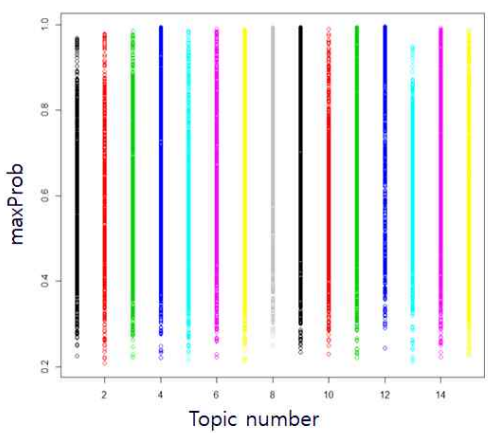

Fig. 4. Topic number and probability distribution.

Table 1. Topic distribution of documents

Document	Topic1	Topic2	...	Topic15
1	0.108623	0.005872	...	0.009183
2	0.001929	0.306251	...	0.086726
3	0.481501	0.003192	...	0.016178
4	0.286055	0.156239	...	0.001635
5	0.062923	0.000944	...	0.070899

분포를 보여준다. 주제 내 단어들을 통해 LDA의 결과인 각 토픽이 어떤 주제 범위를 갖는지 판단할 수 있으며 서로 동일한 주제 범위를 가지는 문서끼리 클러스터링된다. LDA 기반 토픽 모델링을 통해 문서를 구조화하여 잠재되어있는 문서와 문서 내 단어 간의 관계를 파악할 수 있었다.

3.3 단어 간 유사도 추출

전처리 작업을 거친 데이터를 Word2vec을 이용하여 200차원, 대용량 데이터에 성능이 좋은 Skip-gram 방식으로 학습하였다. 학습 결과인 단어 벡터 값들을 Vector Space Model로 구축했다. 그 결과 의미적으로 유사한 단어들끼리 근접한 벡터 공간에서 위치하는 것을 확인할 수 있었다. 서로 연관되어 있는 단어들이 군집을 형성하며 비슷한 공간에 위치하고 있으며 Word2vec 학습을 통해 단어를 벡터화 할 때 단어의 문맥적 의미를 보존하는 것을 알 수 있다. 단어 간의 유사도를 구하기 위해 단어를 벡터 값으로 표현한 수치를 cosine similarity를 이용해 계산하여 단어 벡터들 간의 거리를 측정하였다. 식 (1)은 벡터 A와 B의 cosine similarity를 구하는 계산식이다.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B)^2}} \quad (1)$$

cosine similarity는 내적 공간의 두 벡터 사이의 각도를 cosine 값을 이용해 측정하여 벡터 간의 유사한 정도를 구한다. 0에서 1사이의 값을 가지며 1에 가까울수록 두 단어가 유사하다. 다음과 같은 방식으로 문서 내 단어 벡터들 간의 거리를 계산하였고 Table 2와 같이 단어 간의 유사도를 나타내는 거리 행렬을 생성한다.

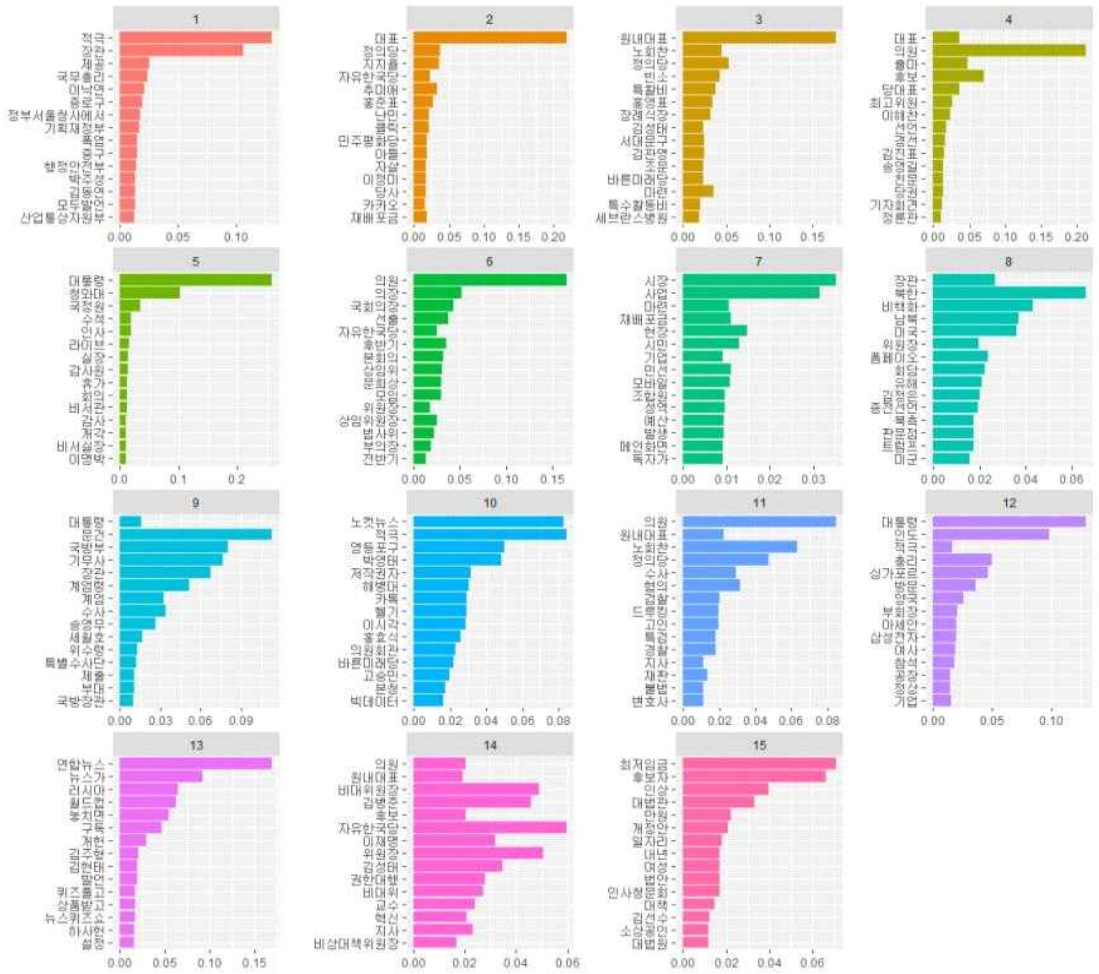

Fig. 5. Probability distributions of top 15 words by topic (Utilizing Korean news articles in political categories provided by Naver).

Table 2. Example of a distance matrix (Utilizing Korean news articles in political categories provided by Naver)

Word	일자리	창출	고용	...
일자리	1	0.834267	0.715028	...
창출	0.834267	1	0.765479	...
고용	0.715028	0.765479	1	...
안심공제	0.657072	0.580912	0.564536	...
소득	0.606912	0.532181	0.431789	...
...	...	...	...	...

3.4 연관성 Score 도출

3.2절에서는 LDA 기반 토픽모델링을 통해 문서 내 주제 분포와 주제 내 단어 분포를 추출했고 3.3절에서는 Word2vec을 사용하여 단어 간 유사도를 구해 거리 행렬을 생성하였다.

본 절에서는 문서의 주제 내 키워드에 해당하는 부분만 추출하여 가중치 행렬을 생성하고 단어 연관성 가중치를 적용해 연관성 SCORE를 도출한다. Fig. 6은 본 논문에서 제안한 연관 문서 추천 시스템의 프레임워크이며 사용자의 관심문서 또는 질의 문서의 확률 분포가 가장 높은 주제를 찾고 사용자가 원하는 주제 내의 키워드에 따라 가장 연관성이 높은 맞춤형 문서들을 추천해준다. 사용자 관심문서는 사용자의 조회 수가 높은 뉴스 기사로 정의한다. Fig. 7은 연관성 SCORE 도출 예시를 나타낸다.

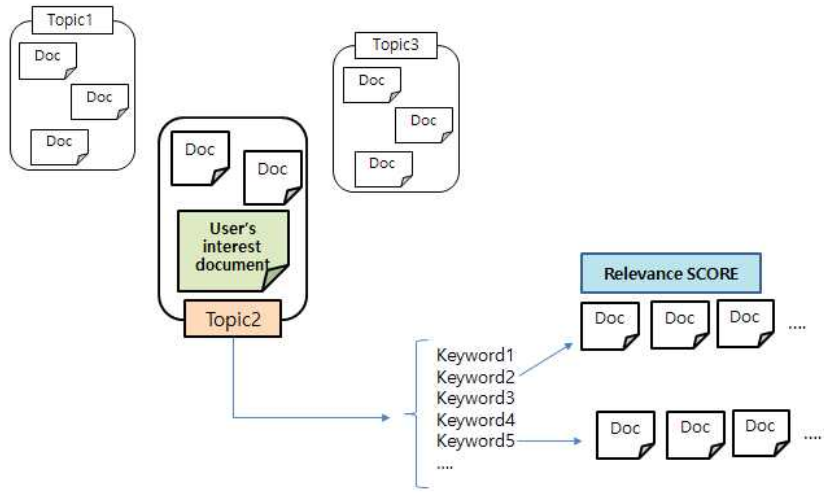

Fig. 6. Framework for Recommendation of Related Documents.

Fig. 7. Procedure of extracting relevance score (Utilizing Korean news articles in political categories provided by Naver).

사용자 관심문서의 키워드와 문서 집합 내 단어들의 의미 관계를 고려하기 위하여 Word2vec을 사용해 생성한 단어들 간의 유사도를 나타내는 거리 행렬에서 해당 주제의 키워드에 해당하는 부분만 추출하여 가중치 행렬로 사용하였고 문서와 단어의 관계를 빈도수로 나타내는 DTM을 생성하였다. 두 행렬의 가중합을 통해 연관성 SCORE를 산출하였고 키워드와 문서간의 연관성을 파악할 수 있다. 가중합은 각각의 수에 가중치 값을 곱한 후 이 곱셈 결과들을 다시 합하는 계산 방식을 의미한다. 연관성 SCORE 산출 과정은 식 (2)와 같이 나타낼 수 있다.

$$\text{Relevance SCORE} = \sum_{i=1}^{n}\sum_{j=1}^{q} w_j a_{i,j} \quad (2)$$

가중치 행렬에서 '비핵화'와 '핵물질', '일자리', '폭염'과 같은 문서 내 단어들의 유사도를 나타는 1행과 DTM에서 문서1 내에서 단어 등장 유무를 나타내는 1열을 가중합하면 '비핵화' 키워드와 문서 1의 연관성 SCORE를 도출할 수 있다. '비핵화' 키워드와 연관성이 높은 '핵물질'은 높은 가중치가 부여되고 연관성이 낮은 '일자리'와 '폭염'은 낮은 가중치를 갖게 된다. 단어의 의미와 문장에서의 맥락을 내포한 단어 간의 의미적 유사도를 가중치로 사용하였고 특정 키

Table 3 The relevance score normalization results of the 'denuclearization' keyword

Number of documents	Analysis	
	Relevance score for 'Denuclearization'	Normalization of Relevance score
388	224.46810	1.0000000
729	222.78679	0.8923682
635	197.73818	0.7192331
754	189.33109	0.6670238
760	172.44196	0.5863405

워드와 의미가 유사할수록 높은 가중치를 부여할 수 있다. 가중치가 적용된 키워드와 문서 내 단어들의 등장 유무를 통해 연관성 SCORE를 산출하였다. 연관성 SCORE를 통해 어떤 문서가 어떤 주제 내 특정 키워드와 얼마나 연관성이 있는지 수치화할 수 있게 되고 점수가 높은 순서대로 문서를 추천해준다.

제안하는 방법론을 통하여 문서를 검색할 때 여러 뜻을 가지고 있는 다의어와 모양이 달라도 의미는 같은 동음이의어를 처리하여 키워드를 확장할 수 있고 단어 간의 의미 관계를 고려한 의미 기반 문서 검색이 가능해진다.

Table 3은 '비핵화' 키워드에 대한 연관성 SCORE 결과와 정규화 과정을 거친 연관성 SCORE 값을 비교한 것이다. 키워드 결과 값마다 서로 다른 연관성 SCORE 범위를 가지므로 범위를 0에서 1 구간으로 일치시키기 위하여 연관성 SCORE를 정규화하였다. 1에 가까울수록 연관성이 높은 문서이며 정규화 할 때 사용한 수식은 식 (3)과 같다.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{3}$$

4. 실험 결과 및 고찰

4.1 실험 결과 및 성능 평가

실험에 사용한 키워드는 '비핵화'로 선정했으며 제안하는 방법의 상위 연관성 SCORE 범위와 하위 연관성 SCORE 범위에 있는 기사들은 Table 4와 같다.

상위 연관성 SCORE 범위에 있는 기사들은 핵, 핵물질, 핵탄두, 핵실험장, 핵국가, 핵무력 등 '비핵화' 키워드와 연관성이 높은 단어들로 구성되기 때문에 주로 직접적인 의견이 담긴 기사들이 나타났다. 하위 연관성 SCORE 범위에 있는 기사들은 종전선언, 유해, 송환, 휴가 등 '비핵화'와 낮은 연관성을 갖는 단어들을 포함하기 때문에 이와 같은 결과가 도출되었다.

본 논문에서 제안하는 방법에 대한 성능을 평가하기 위해 TextRank 알고리즘을 사용하여 문서 내 단어의 중요도를 측정하였다[9-12]. 문서 내의 단어들을 정점(Node)로 선택하였고 두 단어의 연관성을 확률적으로 계산하는 PMI(Pointwise Mutual Information) 값을 간선(Edge)으로 지정했다. PMI 값이 클수록 두 단어의 연관성이 높다는 것을 의미한다[13]. PMI의 계산식은 식 (4)와 같고 분모의 P(X,Y)는 X,Y가 동시에 출현할 확률을 나타내고 있으며 분자 P(X) P(Y)는 X와 Y가 각각 독립적으로 일어날 확률을 의미하고 있다. 성능 평가 또한 두 범위에서 진행하였으며 범위별 단어 중요도 결과는 Table 5와 같다.

$$PMI(X,Y) = \log \frac{P(X,Y)}{P(X)P(Y)} \tag{4}$$

성능 평가 결과 '비핵화' 키워드는 상위 연관성 SCORE에서 더 높은 중요도를 가졌으며 상위 연관성 SCORE 범위의 '트럼프', '김정은'과 하위 연관성

Table 4. Examples of articles by the range of relevance scores (Utilizing Korean news articles in political categories provided by Naver)

Range	Content of the article
High relevance score	'비핵화'에 대한 직접적인 의견이 담긴 기사 - 대통령 통일외교안보특보의 비핵화 협상에 대해 주목해야 할 3가지 - 감정적인 비핵화 비관론을 경계하는 내용
Low relevance score	'비핵화'에 관련된 사건이나 인물에 대한 기사 - 트럼프 대통령의 종전선언 제안 - 비핵화에 좋은 영향을 줄 동창리 발사장 해체 - 비핵화 협상에 긍정적인 영향이 기대되는 북한의 미군 유해 55구 송환 - 비핵화 협상의 중재자로서 강행군하던 문 대통령의 감기 몸살로 인한 휴가

Table 5. Importance of words by the range of relevance scores (Utilizing Korean news articles in political categories provided by Naver)

High relevance score		Low relevance score	
Word	Importance	Word	Importance
북한	0.100335	문, 대통령	0.126457
폼페이오	0.0905634	종전, 선언	0.0956396
트럼프, 대통령	0.0881026	미군, 유해	0.0562377
비핵화, 협상	0.0684315	휴가	0.0513322
미국	0.0676661	비핵화, 협상	0.0467537
핵, 문제	0.0653499	정상회담	0.0428777
김정은, 위원장	0.0641639	발사장	0.0027337
종전, 선언	0.0560652	송환	0.0193288
정상회담	0.0448876	트럼프	0.018463
미사일	0.0290696	김정은	0.0167175

SCORE 범위의 '종전', '선언'과 같이 각 범위에서 중요도가 높은 키워드는 서로 낮은 중요도를 가지는 것 또한 알 수 있었다. 본 논문에서 제안하는 단어 연관성 가중치를 적용한 연관성 SCORE에 따른 의미 기반 문서 추천의 적합성을 확인할 수 있었다.

4.2 비교 평가

키워드와 문서간의 연관성을 측정하여 문서의 랭킹을 매기기 위해 제안하는 방법과 기존 방법론을 통해 문서들에 대한 연관성을 도출하여 비교 실험을 진행한다. 검색 엔진에서 많이 사용되는 TF-IDF와 LDA를 사용하였고 본 논문에서 제안하는 방법의 결과인 상위 연관성 SCORE 범위와 하위 연관성 SCORE 범위에서 기존 방법론을 통해 키워드와 문서간의 연관성을 측정하였다. Table 6과 Table 7은 각 범위의 비교 실험에 대한 결과를 나타낸다.

비교 실험 결과 TF-IDF는 연관성이 불규칙적으로 일치하지 않게 나타났다. LDA의 결과 범위는 0부

Table 6. Comparison test results of the high relevance score range

Number of documents	Analysis		
	TF-IDF	LDA	Suggested method
388	0.000698015753626	0.94474836688648	1.0000000
729	0.001206196817954	0.916006799364045	0.8923682
635	0.000967950498577	0.910056267509865	0.7192331
754	0.001714655168908	0.928782147566483	0.6670238
760	0.0009430603428994	0.935486125847749	0.5863405

Table 7. Comparison test results of the low relevance score range

Number of documents	Analysis		
	TF-IDF	LDA	Suggested method
872	0.000974235891425	0.931601280947023	0.004741640
1085	0.001391765559178	0.979940217495704	0.004741640
379	0.001004680763032	0.194610244156912	0.004741640
616	0.001339574350709	0.732551890720113	0.002848416
972	0.000428663792227	0.082377963835496	0.000000000

터 1까지이며 제안하는 방법의 상위 연관성 SCORE 범위에서는 LDA 또한 범위 기준으로 높은 수치 값이 나타났지만 하위 연관성 SCORE 범위에서는 상위 연관성 SCORE와 유사한 값의 수치 결과가 나타난 것을 확인할 수 있었다.

문장에서의 맥락이나 문맥상의 의미를 통해 단어 간의 관계를 고려하지 않고 TF-IDF는 단어 간의 관계를 단순 빈도수를 기반으로 계산하며 확률 모델인 LDA는 확률적으로 계산하기 때문에 위와 같은 결과가 도출된 것으로 판단할 수 있다. 제안하는 방법은 단어의 문맥적 의미를 보존하여 키워드와 문서의 연관성 SCORE를 도출할 수 있으며 기존 방법론보다 더 효과적인 의미 기반 문서 추천이 가능하다는 것을 알 수 있었다.

5. 결 론

본 논문에서는 키워드와 연관성이 높은 문서를 자동으로 분류하기 위해 한국어 뉴스 기사를 이용하여 LDA 기반 토픽 모델링을 통해 문서 내 주제 분포와 주제 내 단어 분포를 추출하고 Word2vec을 이용해 단어를 벡터화한 후 가중치 행렬을 생성하여 단어 연관성 가중치를 적용해 연관성 SCORE를 도출한 다음 점수가 높은 순서대로 문서를 추천하는 방법을 제안하였다.

가중치 행렬을 통해 사용자가 원하는 키워드와 문서집합 내 단어들의 의미적 연관성을 가중치로 부여하였고 키워드와 문서의 연관성을 SCORE로 수치화하였다. LDA는 확률 모델로 단어 간의 관계를 확률적으로 계산하지만 Word2vec을 이용해 가중치 행렬을 생성하여 단어 연관성 가중치를 적용해 의미적 검색을 가능하게 하였고 확률에 의존한 일반화의 한계를 극복할 수 있도록 했다. 의미적 모호성을 해소하여 문서 검색의 성능이 향상될 수 있고 사용자가 원하는 키워드와 가장 연관성이 높은 문서를 추천해주므로 사용자 맞춤형 정보를 제공할 수 있으며 같은 주제에서 각 키워드와 관련된 사건들을 파악하기 쉬워진다.

연관성 SCORE를 통한 문서 추천 성능 평가 결과로 질의 키워드는 상위 연관성 SCORE 수치 값이 클수록 더 높은 중요도를 가졌다. 상위 연관성 SCORE와 하위 연관성 SCORE 범위에서 중요도가 높은 키워드는 서로 낮은 중요도를 가지는 것 또한 알 수 있었으며 제안하는 방법의 적합성을 확인할 수 있었다. 비교실험을 통하여 기존 문서 랭킹 방법론인 TF-IDF와 LDA보다 더 효과적인 의미 기반 문서 추천이 가능하다는 것을 알 수 있었다.

REFERENCE

[1] J.Y. Kim, "Internet Search Engine : Technological Mode that Draws User's Attention to Make Its Expertise Reinforce," *Journal of Science and Technology Studies*, Vol. 13, No. 1, pp. 181-216, 2013.

[2] J.Y. Oh and S.G. Park, "The Effects of Search Engine Credibility and Information Ranking on Search Behavior," *Journal of Korean Society for Journalism and Communication Studies*, Vol. 53, No. 6, pp. 26-49, 2009.

[3] G.J. Ham, "Semantic-based Document Retrieval Technology Trend," *Journal of Korean Society of Mechanical Engineers*, Vol. 55, No. 5, pp. 38-42, 2015.

[4] R. Kwak, S. Kim, S. Lee, and B. Suh, "Intelligent Issues Tracking System : Exploring Relationship between Stock-specific Keywords and Stock Price," *Proceedings of HCI KOREA*, pp. 351-356, 2018.

[5] M.S. Kim and G.Y. Hae, "XML Information Retrieval by Document Filtering and Query Expansion Based on Ontology," *Journal of Korea Multimedia Society*, Vol. 8, No. 5, pp. 596-605, 2005.

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

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint*, arXiv:1301.3781, 2013.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Proceeding of International Conference on Neural Information Processing Systems,*

pp. 3111-3119, 2013.
[9] L. Page, S. Brin, R. Motwani, and T. Winograd, *ThePageRank Citation Ranking: Bringing Order to the Web*, Stanford Digital Libraries Working Paper, 1998.
[10] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," *Journal of Computer Networks and ISDN Systems*, Vol. 33, pp. 107-117, 1988.
[11] R. Mihalcea and P. Tarau, "TextRank: Brigning Order into Texts," *Proceeding of EMNLP-04 and the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404-411, 2004.
[12] J.Y. Son and Y.T. Shin, "Music Lyrics Summarization Method Using TextRank Algorithm," *Journal of Korea Multimedia Society,* Vol. 21, No. 1, pp. 45-50, 2015.
[13] Turney and M. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417-424, 2002.
[14] S.M. Kim, *Method of Related Document Recommendation Considering Semantic Relation between Words*, Master's Thesis of Chosun University, 2019.

김 선 미

2016년 조선대학교 컴퓨터공학부 학사 졸업
2016년~현재 조선대학교 소프트웨어 융합공학과 석사 과정
관심분야 : 빅데이터 처리, 데이터 마이닝, 머신러닝

나 인 섭

1997년 전남대학교 전산학과 졸업
1999년 전남대학교 전산통계학과 석사 졸업
2008년 전남대학교 전산학과 박사 졸업

2011년~2018년 전남대학교 학술연구교수
2018년~현재 조선대학교 SW융합교육원 조교수
관심분야 : 시각지능(인공지능),영상처리, 객체 검출/인식/추적/이해, 패턴인식, 자율주행, 문자인식 등

신 주 현

1986년~2011년 ㈜청전정보 팀장, ㈜투루텍 기술이사
2007년 조선대학교 전자계산학과 이학박사
2018년 조선대학교 미래사회융합대학 ICT융합학부 부교수
관심분야 : 멀티미디어 데이터베이스, 빅 데이터 분석, 텍스트마이닝, 감성정보 처리 등