Journal of Korea Multimedia Society Vol. 22, No. 11, November 2019(pp. 1313 -1323)
https://doi.org/10.9717/kmms.2019.22.11.1313

키워드의 유사도와 가중치를 적용한 연관 문서 추천 방법

임명진†, 김재현††, 신주현†††

Method of Related Document Recommendation with Similarity and Weight of Keyword

Myung Jin Lim†, Jae Hyun Kim††, Ju Hyun Shin†††

ABSTRACT

With the development of the Internet and the increase of smart phones, various services considering user convenience are increasing, so that users can check news in real time anytime and anywhere. However, online news is categorized by media and category, and it provides only a few related search terms, making it difficult to find related news related to keywords. In order to solve this problem, we propose a method to recommend related documents more accurately by applying Doc2Vec similarity to the specific keywords of news articles and weighting the title and contents of news articles. We collect news articles from Naver politics category by web crawling in Java environment, preprocess them, extract topics using LDA modeling, and find similarities using Doc2Vec. To supplement Doc2Vec, we apply TF-IDF to obtain TC(Title Contents) weights for the title and contents of news articles. Then we combine Doc2Vec similarity and TC weight to generate TC weight-similarity and evaluate the similarity between words using PMI technique to confirm the keyword association.

Key words: Keyword Similarity, Keyword Weight, Similarity-Weight, Related Document, Document Recommendation

1. 서 론

인터넷의 발달과 스마트 폰의 증가로 사용자들은 언제 어디서나 정보를 공유하고 확인할 수 있다. 인터넷 환경이 좋아지면서 사이트에 새로 올라온 글 중에서 사용자가 필요한 정보만 볼 수 있는 RSS[1] 또는 사용자가 작성한 단어가 다른 사용자에게 유용한 정보로 검색될 수 있도록 해주는 Tag[2] 등 사용자들의 편의성을 고려한 다양한 서비스가 늘어나고 있다. 또한 사용자들은 인터넷을 통해 실시간으로 뉴스 기사를 확인할 수 있다. 하지만 하루에도 수많은 뉴스 기사들이 업데이트가 되기 때문에 필요한 뉴스를 찾아내는데 어려움을 겪고 있으며 대부분의 뉴스 기사들은 날짜별, 언론사별 또는 정치, 경제, 사회 등의 카테고리로 구분되어 있어서 각각에 해당하는 뉴스만 볼 수 있기 때문에 사용자가 찾고자 하는 키워

※ Corresponding Author : Ju Hyun 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 : Aug. 30, 2019, Revision date : Oct. 29, 2019
Approval date : Nov. 25, 2019

† Dept. of Computer Engineering, Graduate School, Chosun University
†† Dept. of Development Devision, Bichgalam Information Co.
††† Dept. of Advanced Industry Convergence, Chosun University
※ This study was supported by research fund from Chosun University, 2019

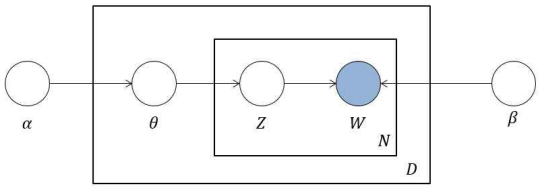

Fig. 1. LDA Model.

드/와 연관되어있는 뉴스를 한꺼번에 검색하기에는 어려운 실정이다. 기존의 연관 문서 추천에 대한 연구는 용어 사전이나 온톨로지와 같은 지식 리소스를 기반으로 하여 사람의 개입과 구축비용, 유지보수가 필요하고[3], TF-IDF와 같은 단순 빈도수 기반으로 문장에서의 단어의 의미를 고려하지 못하고 새로운 단어를 해석하지 못해 검색의 효율성이 떨어진다[4]. 따라서 본 논문에서는 네이버 뉴스 기사의 특정 키워드에 Doc2Vec 유사도와 TF-IDF를 변형한 가중치를 적용하여 보다 정확한 연관 문서를 추천하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로 LDA와 Doc2Vec, TF-IDF에 대해 서술하고, 3장에서는 본 논문에서 제안하는 유사도와 가중치를 적용한 연관 문서 추천 방법에 대해 서술한다. 4장에서는 제안한 방법의 실험 및 결과를 서술하고 5장에서는 결론 및 향후 연구에 대해 서술하고 마무리한다.

2. 관련연구

2.1 LDA

LDA(Latent Dirichlet Alocation)는 2003년 Michaeall Jordan, David Blei, Andrew Ng이 제안한 모델로 주어진 문서들에 대하여 각각의 문서에 어떠한 주제들이 있는지 추출하기 위한 확률모형이다. LDA는 문서별 주제의 분포와 주제별 단어의 분포 모두를 추정해 낸다[5]. LDA 모델은 Fig. 1과 같이 표현되며 N은 단어의 개수이고, D는 문서의 개수를 나타낸다. 문서 집합에서 관측된 W를 이용하여 히든 상태의 θ와 β를 추론한다. 문서들이 갖는 주제 θ를 확률적으로 나타내고 주제에 해당하는 단어들의 확률 분포 Z를 나타낼 수 있다. 본 논문에서는 문서 내 주제 분포를 추출하기 위해 사용하였다.

조태민의 연구에서는 LDA 모델을 기반으로 해당 문서와 비슷한 문서들을 구성하고, 문서들의 키워드를 후보 키워드로 설정한 후 후보 키워드를 구성하는 각각의 단어들을 이용하여 그들의 중요성을 평가하였다[6]. 문서에 잠재되어있는 키워드를 추출하는데 활용할 수 있는 방법이지만 정확률을 더 높일 수 있는 방법이 필요하다.

2.2 Doc2Vec

Doc2Vec은 Mikolov가 제안한 모델로 문장이나 문서에 대하여 vector로 나타내는 비지도 학습 방법이다[7]. Doc2Vec 모델은 두 가지 알고리즘으로 나뉘는데 Fig. 2와 같이 DM(Distribted Memory)과 DBOW(Distributed bag of words) 모델이 있다[8]. DM 모델은 단어를 학습시킬 때 학습 단계마다 벡터에 기억하게 하여 최종 학습된 벡터를 정의하는 방식이고, DBOW 모델은 어떤 문서가 주어지면 그 문서에 포함되어 있는 단어를 예측하는 방식이다[9]. 본 논문에서는 문서에서 단어 간의 거리를 Cosine 유사도로 계산하여 의미적 유사도를 구하기 위해 사용하였다.

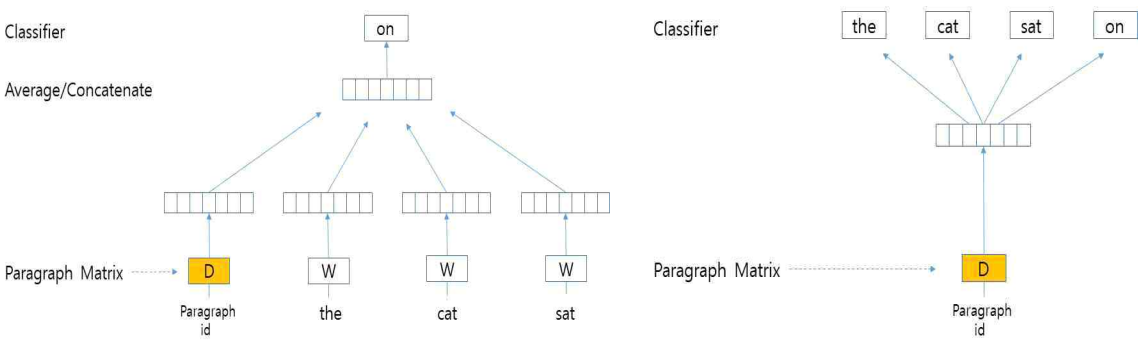

Fig. 2. (a)DM Model and (b)DBOW Model.

김선미의 연구에서는 LDA와 Word2Vec를 활용하여 가중치 행렬을 생성하여 단어 연관성 가중치를 적용한 문서 추천 방법을 제안했다[10]. 하지만 Word2Vec는 한 문서 내 단어의 의미를 파악할 수는 있지만 문서간의 의미 파악은 어렵기 때문에 본 논문에서는 Doc2Vec를 사용하였다.

2.3 TF-IDF

TF-IDF는 TF(Term Frequency)와 IDF(Inverse Document Frequency)의 합성어로 텍스트 마이닝에서 활용하는 가중치로서 전체 문서 중 해당 단어가 어느 정도 중요한지를 나타내는 수치이다[11]. TF는 한 문서에서 해당하는 단어가 얼마나 등장하는지 나타내고, DF는 전체 문서에서 해당하는 단어가 등장한 문서의 수이다. IDF는 전체 문서에서 해당하는 단어가 등장한 문서의 역수인 역문서 빈도이다. 식(1)은 TF-IDF를 수식으로 표현한 것이다. tf(t,d)는 특정 문서(d)에서 단어(t)의 총 빈도수이다. idf(t,D)는 전체 문서(D)에서 단어(t)가 포함된 문서의 수로 나누어 log를 취해 얻는다.

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \quad (1)$$

이성직의 연구에서는 중요 키워드를 TF-IDF를 변형하여 전체 문서집합에서 산출하였다[12]. 하지만 전체 문서집합에서 키워드를 추출했기 때문에 단일 문서를 요약하거나 탐색하는 방법에는 활용할 수 없다. 따라서 문서별로 구분하여 키워드를 추출하고 활용할 수 있는 기술이 필요하다.

3. 키워드의 유사도와 가중치를 적용한 연관 문서 추천

본 논문에서는 사용자가 찾고자 하는 키워드와 연관된 문서 추천을 위해 LDA를 사용하여 토픽을 추출하고 Doc2Vec을 통해 유사도를 구한다. 뉴스 기사의 제목과 내용에 각각 TF-IDF를 진행해 TC(Title Contents) 가중치를 만든 다음 Doc2Vec의 결과와 비교하여 동일한 키워드에 TC 가중치-유사도를 적용시켜 추출한 키워드들을 통해 연관 문서를 추천하는 방법을 제안한다.

3.1 시스템 구성도

Fig. 3은 본 논문에서 제안하는 시스템 구성도를 나타내고 있으며, LDA 토픽 추출 단계와 Doc2vec 유사도 추출단계 그리고 TF-IDF 적용 단계로 구성된다. 본 논문에서 사용한 데이터는 네이버 정치 카테고리에 해당하는 뉴스 기사를 Java기반 환경에서 수집하여 실험을 진행하였다. 수집된 데이터는 R 환경에서 KoNLP 패키지를 사용하여 한국어 처리를 하고 Corpus를 생성한 후 영어, 특수문자 등 불용어를 제거하였다. 다음 단계로는 LDA를 사용하여 주제별 토픽을 추출하고, Doc2Vec을 사용하여 유사도를 추출한다.

Doc2Vec 유사도를 보완하기 위한 방법으로 뉴스 기사의 제목과 내용에 각각 단어별 TF-IDF를 추출하였고 내용에서 나온 키워드는 제목과 일치하는 키워드만 추출해 빈도수 기반 평균을 구하여 TC 가중치로 정의하였다. 이후 Doc2Vec의 결과와 비교하여 일치하는 키워드에 TC 가중치를 적용하고 의미와 빈도수를 결합하기 위해 Doc2Vec 유사도와 TC 가중치를 결합해 TC 가중치-유사도가 높은 키워드들을 추출하여 보다 정확한 연관 문서를 추천하는 방법을 제안한다.

3.2 LDA 토픽 추출

본 절에서는 네이버 정치 카테고리의 뉴스 기사를 웹 크롤링으로 수집하고 전처리한 다음 LDA 모델을 사용하여 토픽을 추출하는 과정에 대해 서술한다. 본 논문에 사용된 데이터는 네이버 정치 카테고리의 뉴스 기사 중 2018년 7월 19일부터 8월 19일까지 1개월 동안 Java환경에서 웹 크롤링하여 총 22,779개를 수집하였다. 수집된 뉴스 기사는 R 환경에서 KoNLP

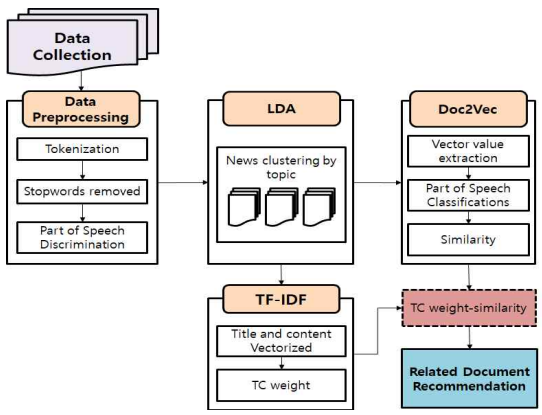

Fig. 3. System Configuration Diagram.

	A	B	C	D	E	F	G	H	I	J	K	L
2	전기요금	정부 여당이 이르면 이번 주 말이나 다음주 초쯤 전기요금 인하 방안을 확정해 발표할 방침인 것으로 확인됐습니다										
3	'北 이송'	북한에서 이송된 미군 전사자 유해 55구를 미국으로 보내는 유해 송환식이 오늘 오산 미군기지에서 엄수됐습니다										
4	다녀오기는	오는 20일부터 열릴 남북 이산가족 상봉 행사를 앞두고 오늘 통일부 차관이 금강산에 다녀왔습니다. 또 모레에는										
5	여군도 최	여군도 최전방 철책을 지키는 GOP, 즉 일반전초 대대의 중대장과 소대장을 맡을 수 있게 됐습니다. 국방부가 국방										
6	허성무 창	각 분야 전문가와 허심탄회한 대화 나눠 [김종성 기자(=경남)] 경남 창원시는 '창원 미래 30년 먹거리										
7	제주도의회	입법고문에양영철교수.법률고문에강기탁·고성호변호사 제주도의회는 1일 오전 11시 의장실에서 양영철 제주대 교										
8	함양신용협	취약계층 등에 보급할 소화기67대, 감지기67대 기증 [김상우] 경남 함양소방서는 함양신용협동조합으										
9	이산상봉	이산가족 상봉행사를 위한 시설 개보수 상황 점검차 1일 금강산을 방문한 천해성 통일부 차관(왼쪽 두번째)이 금강										
10	라오스 외	강경화 장관, 1일 라오스 등 아세안 6개국과 양자회담미얀마 장관 "한류로 청년들 '대디' 대신 '아버지'라 한다"강경										

Fig. 4. Data Set.

패키지로 한국어 처리를 진행하고 품사를 판별하여 두 글자 이상의 보통 명사들만 추출한 다음 Corpus를 생성하여 영어, 특수문자, 완전한 글자가 아닌 단어 등을 불용어로 처리하였다. Fig. 4는 본 논문에서 사용한 데이터 셋 이다.

전처리 과정을 거친 데이터 셋은 LDA를 통해 토픽을 추출한다. LDA의 클러스터 수는 하이퍼 파라미터로 반복적인 실험을 통해 최적의 클러스터를 구한다. 본 논문에서는 클러스터의 수를 12개로 지정했을 때 가장 성능이 좋았으며 Fig. 5는 LDA 토픽 추출 결과이다.

정치 카테고리의 뉴스 기사 22,779개를 LDA 클러스터링 한 결과 12개의 클러스터를 생성했고 각 클러스터에는 '이산가족', '일자리', '특검', '비핵화' 등 서로 연관된 토픽으로 묶여있는 것을 볼 수 있다. 첫 번째 토픽은 상봉, 이산가족, 금강산 등 이산가족에 관련된 단어들이 분포되어 있다. 따라서 본 논문에서는 '이산가족'을 키워드로 선정하였다.

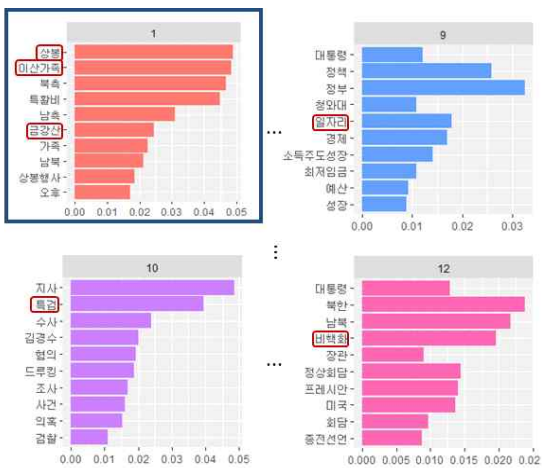

Fig. 5. LDA Topic Extraction.

3.3 Doc2Vec 유사도 추출

본 절에서는 Python Gensim패키지를 이용하여 Doc2Vec으로 토픽의 벡터를 추출한 후 키워드 유사도를 추출하는 과정에 대해 서술한다. Doc2Vec은 띄어쓰기가 하나의 단어로 학습되기 때문에 같은 단어라도 여러 가지 뜻을 나타낸다. 22,779개의 데이터를 랜덤으로 20,000개의 학습 데이터와 10,000개의 테스트 데이터로 구분하여 실험을 진행하였다. Fig. 6은 Doc2vec을 실험하기 위한 사전 구축 방법이다.

Table 1은 사전 구축된 Doc2Vec을 토대로 '이산가족' 키워드와 관련된 상위 20개 연관 키워드와 유사도를 출력한 결과이다. '이산가족' 키워드와 연관 키워드 간의 유사도를 확인할 수 있고, 유사도는 0부터 1로 나타내며 1에 가까울수록 단어 간 유사도가 더 높다.

Table 2는 Table 1의 연관 키워드와 유사도의 정확성을 비교하기 위해 관심 뉴스 기사와 관련된 연관 뉴스 기사A와 연관 뉴스 기사B를 보여주고 있다. 관심 뉴스 기사와 연관 뉴스 기사A를 비교해보면 공통적으로 '이산가족 상봉', '금강산', '개별'이라는 연관 키워드들이 있으며, 공통적이진 않지만 '오붓하게', '작별', '점심' 이라는 키워드를 확인할 수 있다. 따라서 공통적이지 않은 키워드가 있어도 연관된 기사임을 알 수 있다.

```
tagged_train_docs=[TaggedDocument(d,[c])for d,c in train_docs]
tagged_test_docs=[TaggedDocument(d,[c])for d,c in test_docs]

from gensim.models import doc2vec
doc_vectorizer=doc2vec.Doc2Vec(vector_size=300,window=10, min_count=5, workers=11,
                               alpha=0.025, min_alpha=0.025, epochs=20)
doc_vectorizer.build_vocab(tagged_train_docs)
doc_vectorizer.train(tagged_train_docs, epochs=doc_vectorizer.epochs,
                     total_examples=doc_vectorizer.corpus_count)

pprint(doc_vectorizer.most_similar('이산가족/Noun'))
```

Fig. 6. Doc2Vec Dictionary Construction.

Table 1. Similarity of keywords related to separated families

No.	Related Keywords	Doc2Vec Similarity	No.	Related Keywords	Doc2Vec Similarity
1	이산	0.643995	11	눈물	0.545792
2	작별	0.621491	12	만남	0.534576
3	금강산	0.589559	13	점심	0.527799
4	개별	0.584738	14	어머니	0.527620
5	이산가족상봉	0.578467	15	설레다	0.516723
6	이별	0.559387	16	오붓하다	0.506205
7	가족	0.558603	17	식사	0.504146
8	65년	0.556476	18	형제	0.502905
9	68년	0.555606	19	상봉	0.470939
10	오열	0.548075	20	재회	0.466823

Table 2. Related news article Example

Interested News articles	이번 **이산가족 상봉** 행사는 오는 20일부터 26일까지 금강산관광지구에서 진행됩니다. 첫날에는 금강산호텔에서 단체상봉과 북측 주최 환영만찬이 진행됩니다. 둘째 날 오전에 객실에서 개별상봉을 진행한 다음 객실에서 오붓하게 오찬까지 이어갑니다. 셋째 날의 경우에도 예전에는 작별상봉을 마친 다음 남북이 오찬을 따로 진행했으나, 이번 행사에서는 작별상봉과 공동오찬을 묶어서 진행하게 됩니다[13].
Related news articles A	2년 10개월 만에 재개되는 **이산가족 상봉**행사를 하루 앞두고 남측 가족들이 오늘 강원도 속초에 모였습니다. 강원도 고성의 동해선 남북출입사무소를 거쳐 북측 통행검사소에서 심사를 받고, 낮 12시 30분쯤 금강산 온정각에 도착할 것으로 보입니다. 특히 둘째 날인 21일에는 2시간의 개별상봉에 이어 1시간 동안 객실에서 함께 점심을 먹게 됩니다[14].
Related news articles B	문 대통령은 다음 달로 합의한 김정은 북한 국무위원장과의 평양 남북 정상회담을 통해 분단의 원인이 된 65년간의 전쟁 체제를 종식하는 종전선언을 연내 이루겠다는 뜻을 분명히 밝혔다. 오는 평양 남북 정상회담에서 기존의 4·27판문점 남북 정상회담에서 합의한 **비핵화**와 종전선언을 구체화시키겠다는 뜻을 공개적으로 밝힌 것이다[15].

하지만 관심 뉴스 기사와 연관 뉴스 기사B를 비교해보면 '남북'이라는 키워드는 공통적으로 있지만 관심 뉴스 기사는 '이산가족'에 관련된 뉴스 기사이고 연관 뉴스 기사B는 '비핵화'에 관련된 뉴스 기사이다. 따라서 Doc2Vec에서는 연관이 없는 기사인 경우에도 단어가 아닌 문장과 문서간의 유사도를 측정하여 공통된 단어가 포함되어 있다는 이유로 높은 유사도가 추출된 것을 볼 수 있다.

3.4 TC(Title Contents) 가중치 추출

본 절에서는 Doc2Vec 결과로 관련성이 낮은 키워드가 추출되는 문제점을 보완하기 위하여 전체 기사에 대한 TF-IDF 가중치와 본 논문에서 제안하는 뉴스 기사의 제목과 내용에 대한 TF-IDF 가중치를 적용한 실험을 진행한다. 먼저, TF-IDF 가중치는 Fig. 7과 같이 Count- Vectorizer를 사용하여 DTM을 구한 후 TfidfVectorizer를 사용하여 계산한다.

TF-IDF 알고리즘을 통해 중요 단어를 추출하였고, 상위 100개의 키워드를 고차원 데이터의 유사성을 저차원으로 임베딩 하는 방법인 t-SNE로 시각화하였다[16]. Fig. 8은 TF-IDF를 적용한 연관 키워드 추출 결과이다. 100차원에 있던 22,779개의 뉴스 기사를 2차원으로 표현하였고 그 중 상위 100개의 키워

```
tf = CountVectorizer()

tf.fit_transform(documents)
tf.fit_transform(documents)[0:1].toarray()

tfidf = TfidfVectorizer(max_features = 100, max_df=0.95, min_df=0)

A_tfidf_sp = tfidf.fit_transform(documents)

tfidf_dict = tfidf.get_feature_names()
```

Fig. 7. TF-IDF Calculation Algorithm.

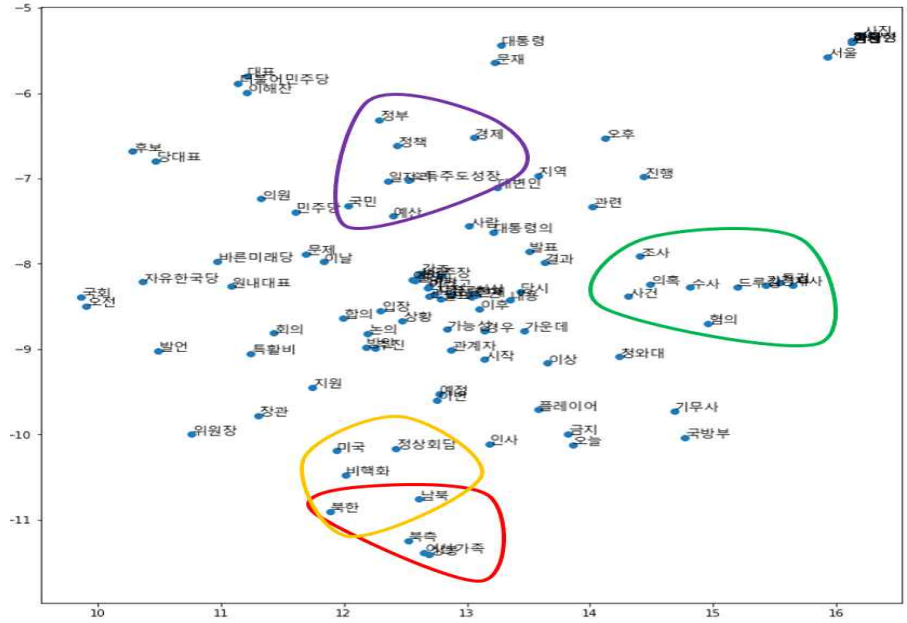

Fig. 8. Related keyword extraction using TF-IDF.

드만 추출하였다. 추출된 키워드들은 크게 4개의 그룹인 '일자리', '특검', '비핵화', '이산가족'과 연관되어 있는 것을 알 수 있다. 그리고 '비핵화'와 '이산가족' 그룹에는 '남북'과 '북한'이라는 공통된 키워드가 있음을 확인할 수 있다.

본 논문에서는 TF-IDF 결과인 4개의 그룹 중에서 '이산가족'에 관련된 뉴스 기사만 분류하여 실험을 진행하였다. 이산가족에 관련된 뉴스 기사 총 1,210개를 분류하여 '제목과 내용'에 TF-IDF 가중치를 구하고 이를 결합해 키워드별 TC 가중치를 추출한다[17]. 식 (3)은 TC 가중치를 적용하는 식을 나타낸다.

$$V(t,d,D) = tf(t_1,s_1) \times idf(t_1,S_1) + tf(t_2,c_2) \times idf(t_2,C_2) \quad (3)$$

- $V(t,d,D)$: TC(제목과 내용) 가중치
- $tf(t_1,s_1) \times idf(t_1,S_1)$: 특정제목(s_1)내 단어(t_1)의 총 빈도수 × 전체제목(S_1) 중 단어(t_1)가 들어간 빈도수
- $tf(t_2,c_2) \times idf(t_2,C_2)$: 특정내용(c_2)내 단어(t_2)의 총 빈도수 × 전체내용(C_2) 중 단어(t_2)가 들어간 빈도수

이산가족에 관련된 뉴스 기사 1,210개에 대하여 제목과 내용의 TF-IDF를 계산한 결과 제목의 TF-IDF가 더 높게 나타났다. 뉴스 기사의 제목은 내용의 핵심적인 단어이고 문서를 대표하는 키워드이기 때문이다. Fig. 9는 제목과 내용의 키워드에 가중치를 적용하는 방법이다.

제목의 키워드 A, C와 내용의 키워드 A~F가 있다고 가정 했을 경우 일치하는 A, C 키워드에 대해서만 TF-IDF 가중치를 적용하였다. Table 3은 제목과 내용 상위 10개의 키워드이다. 내용의 키워드들 중 제목과 동시에 출현하는 키워드에 대하여 가중치를 구하였다. TC 가중치는 Doc2Vec의 연관 키워드와 비교하여 일치하는 키워드에 가중치를 적용하였다.

Table 4는 제목과 내용을 결합시킨 결과의 키워드별 가중치를 Doc2Vec의 연관 키워드를 기준으로 일치하는 키워드만 추출하여 적용한 결과이다.

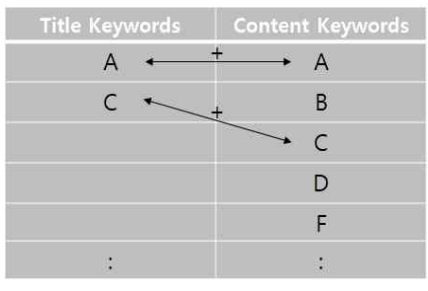

Fig. 9. Keyword weighting method for title and content.

Table 3. Title and content keywords

No.	Title Keywords	TF-IDF Weight	No.	Content Keywords	TF-IDF Weight
1	선물	0.899776	1	아버지	0.26336
2	아들	0.843453	2	할머니	0.218488
3	기다림	0.839531	3	가족	0.216677
4	정상회담	0.781644	4	눈물	0.208386
5	이산가족상봉	0.765738	5	개별상봉	0.201079
6	눈물	0.751263	6	단체상봉	0.17964
7	작별	0.686951	7	이산가족상봉	0.177541
8	남북	0.677432	8	북한	0.172477
9	작별상봉	0.673981	9	어머니	0.157506
10	가족	0.654491	10	금강산	0.143367

Table 4. TF-IDF Results

No.	Related Keywords	TC Weight	No.	Related Keywords	TC Weight
1	가족	0.216677	11	이별	0.193808
2	만남	0.213922	12	이산	0.177541
3	눈물	0.208386	13	이산가족상봉	0.177541
4	상봉	0.201575	14	작별	0.176505
5	개별	0.201079	15	식사	0.174134
6	오열	0.198652	16	어머니	0.157506
7	설레다	0.198652	17	금강산	0.143367
8	오붓하다	0.198652	18	형제	0.094733
9	점심	0.195899	19	68년	0.028064
10	재희	0.195899	20	65년	0.011805

Doc2Vec에서 '이산가족' 키워드와 연관성이 높았던 '65년'과 '68'년은 일치하는 키워드는 있지만 가중치가 낮아진 것을 볼 수 있다.

3.5 TC 가중치-유사도 추출

앞에서 추출된 Doc2Vec의 유사도와 TC 가중치를 적용하여 의미와 빈도수를 결합하기 위해 TC 가중치-유사도를 만든 결과는 Table 5와 같다.

의미기반의 Doc2Vec의 유사도와 본 논문에서 제안하는 TC 가중치를 합하여 TC 가중치-유사도를 구한 결과 Table 1의 '이산가족' 연관 키워드 중에서 '65년'과 '68년'의 키워드는 유사도가 낮아져 연관성이 떨어짐을 증명하였다. 그리고 연관성이 높지만 Doc2Vec 유사도에서 연관성이 낮게 나왔던 키워드들이 TC 가중치-유사도를 통해 높은 유사도를 보여 추출된 연관 키워드들을 통해 보다 정확한 연관 문서를 추천할 수 있게 된다. Fig. 10은 Table 5를 그래프로 나타낸 것이다. 본 논문에서 제안한 TC 가중치-유사도를 적용한 결과 '이산가족'에 연관된 키워드들의 유사도가 더 높아진 것을 볼 수 있다. 추출된 연관 키워드들을 통해 사용자가 찾고자 하는 키워드와 보다 관련 있는 연관 문서 추천이 가능해 진다.

4. 실험 및 결과

본 장에서는 앞에서 제안했던 TC 가중치-유사도를 통해 추출한 키워드에 대해 성능 평가를 수행한다.

본 논문에서는 연관성 평가를 위해 두 단어 A와 B의 연관성을 계산하는 방법으로 확률론을 적용하여 수치화하는 PMI(Pointwise Mutal Information) 기법을 이용한다. 식 (4)는 각각 독립적으로 발생할 확률 $P(A)$와 $P(B)$를 갖는 두 단어 A와 B에 대한 $PMI(A,B)$ 식이다.

Table 5. Similarity and Weight by Keyword in relation to separated families

Related Keywords	Doc2Vec Similarity	TC Weight	TC Weight-Similarity
이산	0.643995	0.177541	0.821536
작별	0.621491	0.176505	0.797996
개별	0.584738	0.201079	0.785817
이산가족상봉	0.578467	0.177541	0.756008
눈물	0.545792	0.208386	0.754178
이별	0.559387	0.193808	0.753195
만남	0.534576	0.213922	0.748498
금강산	0.589559	0.143367	0.732926
점심	0.527799	0.195899	0.723698
:	:	:	:
68년	0.555606	0.028064	0.58367
65년	0.556476	0.011805	0.568281

$$PMI(A,B) = \log\frac{P(A,B)}{P(A)P(B)} \quad (4)$$

PMI 기법은 두 단어가 같이 등장하는 정도를 각각 독립적으로 나올 확률로 나눈 것으로 A와 B단어가 어느 정도 연관성을 나타내는지 확률적으로 계산하는 방법이다[18]. PMI 기법을 사용하여 '이산가족'과 연관 키워드들의 연관성 관계를 계산한 결과 Table 6과 같다.

연관성 평가 결과 '이산'부터 '점심'까지 키워드들을 보면 백분율로 환산할 경우 80%가 넘는 유사도를 보였다. 하지만 Doc2Vec의 결과에서 연관성이 높았던 '65년'과 '68년'은 본 논문에서 제안한 TC 가중치

Fig. 10. Keywords related to separated families.

를 적용한 TC 가중치-유사도와 비슷하게 PMI 결과가 현저히 떨어지는 것을 확인했으며 TC 가중치-유사도로 추출된 연관 키워드들을 통해 사용자가 찾고자 하는 키워드와 보다 관련 있는 연관 문서 추천이 가능해 진다. Fig. 11은 이산가족 키워드의 연관성 평가 그래프이다.

PMI 기법으로 성능을 평가 한 결과 본 논문에서 제안한 TC 가중치를 적용한 TC 가중치-유사도와 동일하게 '이산', '작별', '개별'이라는 키워드는 연관성이 높고, '68년', '65년'이라는 키워드는 연관성이 낮은 것을 볼 수 있다. 이렇게 추출된 연관 키워드들을 통해 사용자가 찾고자 하는 키워드와 보다 관련 있는 연관 문서 추천이 가능해 진다.

5. 결 론

본 논문에서는 연관 문서 추천을 위해 TC 가중치를 적용한 TC 가중치-유사도를 생성하여 추출된 키워드를 통해 연관 문서를 추천하는 방법을 제안하였다. 1개월 동안 네이버 정치 카테고리에 해당하는 뉴스 기사 총 22,779개를 Java환경에서 웹 크롤링으로 수집하여 실험 데이터로 사용하였다.

실험 데이터는 R 환경에서 KoNLP 패키지를 사용해 한국어 처리를 한 후 전처리를 진행하였다. 전처리를 거친 데이터는 LDA 토픽 모델링을 이용하여 토픽을 추출하였다.

추출된 토픽은 Doc2Vec을 통해 유사도를 구하였다. 추출된 유사도를 TF-IDF를 적용해 전체기사에

Table 6. Separated families keyword relevance assessment

Related Keywords	Doc2Vec Similarity	TC Weight-Similarity	ㄴPMI Verification
이산	0.643995	0.821536	0.908
작별	0.621491	0.797996	0.881
개별	0.584738	0.785817	0.865
이산가족상봉	0.578467	0.756008	0.926
눈물	0.545792	0.754178	0.882
이별	0.559387	0.753195	0.875
만남	0.534576	0.748498	0.811
금강산	0.589559	0.732926	0.876
점심	0.527799	0.723698	0.862
:	:	:	:
68년	0.555606	0.58367	0.462
65년	0.556476	0.568281	0.434

대해 시각화 한 후 '이산가족'에 대한 연관 키워드 중에서 '65년'과 '68년'의 연관성이 없음을 증명하기 위해 이산가족에 관한 기사 1210개를 별도로 분류하여 제목과 내용에 각각 단어별 TF-IDF를 추출하였고 내용에서 나온 키워드들은 제목과 동시에 출현하는 키워드를 추출해 TC 가중치를 구하였다. 이후 Doc2Vec에서 추출된 연관 키워드와 일치하는 키워드에 TC 가중치를 적용하여 연관성이 없는 키워드를 증명해 보였다.

단순 빈도수가 아닌 의미와 빈도수를 결합하기 위해 TC 가중치와 의미기반의 Doc2Vec 유사도를 결합해 TC 가중치-유사도를 생성하여 연관성이 높지만 낮게 나왔던 키워드를 추출해 보다 정확한 연관 문서 추천이 가능해 진다.

키워드의 연관성을 평가하기 위해 PMI기법을 이

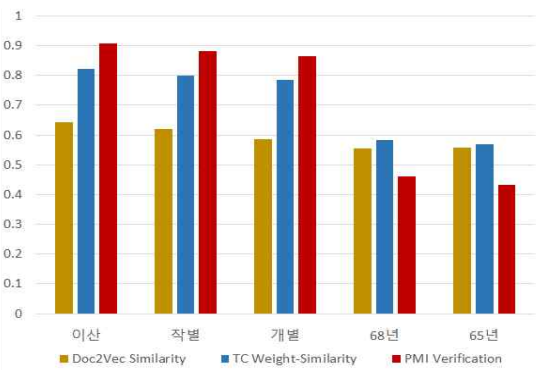

Fig. 11. Relevance evaluation graph.

용하여 단어 간 유사성을 평가한 결과 Doc2Vec 유사도와 달리 본 논문에서 제안한 TC 가중치를 적용한 TC 가중치-유사도의 키워드 결과와 같이 키워드 '65년'과 '68년'이 연관성이 현저히 떨어진 것을 확인하였다. 따라서 TC 가중치-유사도를 적용하여 추출된 연관 키워드들을 통해 사용자가 찾고자 하는 키워드와 보다 정확한 연관 문서를 추천할 수 있게 된다.

REFERENCE

[1] D. Ayers, A. Watt, *Beginning Rss And Atom Programming*, John Wiley & Sons Inc., 2005.

[2] K.P. Lee, D.N. Kim, H.J. Kim, "A Survey on Tagging in the Web 2.0 Environment", *Communications of the Korean Institute of Information Scientists and Engineers*, Vol. 25, No. 10, pp. 36-42, 2007.

[3] M.S. Kim, 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.

[4] E.S. You, G.H. Choi, S.H. Kim, "Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels", *Journal of the Korea Society of Computer and Information*, Vol. 20, No. 2, pp. 121-129, 2015.

[5] D. Blei, A.Y. Ng, M. Jordan, "Latent Dirichlet

allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.

[6] T.M. Cho, J.H. Lee, "Latent Keyphrase Extraction using LDA Model", *Korean Institute of Intelligent Systems*, Vol. 24, No. 2, pp. 125-126, 2014.

[7] An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation(2016). https://arxiv.org/abs/1607.05368 (accessed September 1, 2018).

[8] Q. Le, T. Milokov, "Distributed Representations of Sentences and Documents", *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[9] K. Cheng, J. Li, J. Tang, H. Liu, "Unsupervised Sentiment Analysis with Signed Social Networks", *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 777-786, 2017.

[10] S.M. Kim, I.S. Na, J.H. Shin, "A Method on Associated Document Recommendation with Word Correlation Weights", *Journal of Korea Multimedia Society*, Vol. 22, No. 2, pp. 250-259, 2019.

[11] S. Robertson, "Understanding Inverse Document Frequency: On theroretical arguments for IDF", *Journal of Documentation*, Vol. 60, No. 5, pp. 503-520, 2004.

[12] S.J. Lee, H.J. Kim, "Keyword Extraction from News Corpus using Modified TF-IDF", *The Jounal of Society for e-Business Studies*, Vol. 14, No. 4, pp. 59-73, 2009.

[13] Brothers and children reunion Family members' five days ahead 'Thank you alive' (2018). http://www.newsis.com/view/?id=NISX20180814_00003910928&cID=103018&pID=10300 (accessed September 1, 2018).

[14] Discrete reunion day ahead ... Southern family gathering Sokcho(2018). http://news.kbs.co.kr/news/view.do?ncd=40182418 &ref=A (accessed September 1, 2018).

[15] North America's Denuclearization Will End Stall ... Wen Ji-hye re-enacts 'Korean peninsula driver'(2018). http://www.newsis.com/view/?id=NISX20180815_0000391295&cID=10301&pID=10300 (accessed September 1, 2018).

[16] A. Muller, S. Guido, *Introduction to Machine Learning with Python*, O'REILLY, 2017.

[17] J.H. Kim, *Method of Keyword Recommendation Considering Importance and Correlation of words*, Master's Thesis of Chosun University, 2018.

[18] P.D. Turney, M.L. 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, July 2002.

임 명 진

2000년 군산대학교 컴퓨터과학과 학사 졸업
2018년 조선대학교 소프트웨어융합공학과 석사 졸업
2018년~현재 조선대학교 컴퓨터공학과 박사 과정

관심분야 : 빅데이터 처리, 데이터마이닝, 자연어처리, 머신러닝

신 주 현

1986년~2011년 ㈜청전정보 팀장, ㈜투루텍 기술이사
2007년 조선대학교 전자계산학과 이학박사
2018년~현재 조선대학교 미래사회융합대학 신산업융합학부 부교수

관심분야 : 데이터베이스, 데이터마이닝, 자연어처리, 인공지능

김 재 현

2017년 조선대학교 제어계측로봇공학과 학사 졸업
2019년 조선대학교 소프트웨어융합공학과 석사 졸업
2019년~현재 빛가람정보주식회사 개발사업부 사원

관심분야 : 자연어처리, 데이터마이닝