한국지반신소재학회논문집 제22권 2호 2023년 6월 pp. 47 ~ 54
J. Korean Geosynthetics Society Vol.22 No.2 June 2023 pp. 47 ~ 54
DOI: https://doi.org/10.12814/jkgss.2023.22.2.047
ISSN: 2508-2876(Print) ISSN: 2287-9528(Online)

워드 임베딩(Word Embedding)을 활용한 최적의 키워드 추출 및 검색 방법 연구

A Study on the Optimal Search Keyword Extraction and Retrieval Technique Generation Using Word Embedding

이정인[1], 안진희[2], 고경택[3], 김영석[4]*
Jeong-In Lee[1], Jin-Hee Ahn[2], Kyung-Taek Koh[3], YoungSeok Kim[4]*

[1]*Member, Researcher, Korean Peninsula Infrastructure Special Committee, Korea Institute of Civil Engineering and Building Technology, 283 Goyangdae-Ro, Ilsanseo-Gu, Gyeonggi-Do 10223, Republic of Korea*
[2]*Non-Member, Researcher, Korean Peninsula Infrastructure Special Committee, Korea Institute of Civil Engineering and Building Technology, 283 Goyangdae-Ro, Ilsanseo-Gu, Gyeonggi-Do 10223, Republic of Korea*
[3]*Non-Member, Senior Research Fellow, Korean Peninsula Infrastructure Special Committee, Korea Institute of Civil Engineering and Building Technology, 283 Goyangdae-Ro, Ilsanseo-Gu, Gyeonggi-Do 10223, Republic of Korea*
[4]*Member, Senior Research Fellow, Northern Infrastructure Specialized Team, Korea Institute of Civil Engineering and Building Technology, 283 Goyangdae-Ro, Ilsanseo-Gu, Gyeonggi-Do 10223, Republic of Korea*

ABSTRACT

In this paper, we propose the technique of optimal search keyword extraction and retrieval for news article classification. The proposed technique was verified as an example of identifying trends related to North Korean construction. A representative Korean media platform, BigKinds, was used to select sample articles and extract keywords. The extracted keywords were vectorized using word embedding and based on this, the similarity between the extracted keywords was examined through cosine similarity. In addition, words with a similarity of 0.5 or higher were clustered based on the top 10 frequencies. Each cluster was formed as 'OR' between keywords inside the cluster and 'AND' between clusters according to the search form of the BigKinds. As a result of the in-depth analysis, it was confirmed that meaningful articles appropriate for the original purpose were extracted. This paper is significant in that it is possible to classify news articles suitable for the user's specific purpose without modifying the existing classification system and search form.

요 지

본 논문에서는 자료 조사를 위한 최적의 키워드 추출 및 검색 방법을 제안하였으며, 북한 건설 관련 동향 파악을 예시로 제안 방법을 검증하였다. 대표적인 국내 언론 플랫폼인 빅카인즈(BigKinds)를 활용하여 표본 기사를 선정하고 키워드를 추출하였다. 추출된 키워드는 워드 임베딩(Word Embedding)을 활용하여 벡터화하였으며, 이를 토대로 코사인 유사도(Cosine Similarity)를 통해 추출된 키워드 간의 유사도를 검사하였다. 또한 상위 빈도수 10개에 대한 키워드를 기준으로 유사도 0.5 이상인 키워드들을 군집화하였다. 각 군집들은 빅카인즈 검색 양식에 맞추어 군집 내부 키워드 간에는 'OR', 군집 간에는 'AND'로 형성하였다. 심층 분석 결과, 본래 목적에 맞는 유의미한 기사들이 추출되었음을 확인할 수 있었다. 기존의 분류체계 및 검색 양식을 변형시키지 않은 상태에서 사용자의 세부 목적을 충족시키는 자료 조사·분류가 가능하게 되었다는 점에서 의의를 갖는다.

Keywords **:** Keyword extraction, Retrieval technique, Word embedding, Cosine similarity, BigKinds

Received *5 Jun. 2023,* ***Revised*** *13 Jun. 2023,* ***Accepted*** *19 Jun. 2023*

****Corresponding author***
Tel: +82-31-910-0371; Fax: +82-31-910-0561
E-mail address: kimys@kict.re.kr (Y. Kim)

1. 서 론

4차 산업혁명이 대두됨에 따라 클라우드 서비스와 빅데이터를 기반으로 한 온라인 플랫폼 및 아카이브 서비스가 다양한 분야에 적용·확대되고 있다. 특히 언론 분야의 경우 한국언론진흥재단이 제공하는 '빅카인즈(BigKinds)'가 대표적인 사례라고 할 수 있다(BigKinds, 2023). 이는 종합일간지, 지역 일간지, 방송사 등 54개의 언론사로부터 텍스트 기사, 사진, PDF의 DB를 수집하여 뉴스 카테고리에 맞추어 자동 분류하고 핵심 키워드를 추출한다. 또한 'AND'와 'OR' 및 형태소 단위의 검색 기능을 제공하여 원하는 기사를 세부적으로 분류할 수 있다. 'AND'의 경우 앞뒤의 단어가 모두 포함된 기사를 분류하고, 'OR'은 앞뒤의 단어 중 하나라도 포함된 기사를 분류하게 된다. 추가적으로 제외 키워드를 설정해 해당 기사를 제외하는 것도 가능하다. 또한, 사용자에 의해 분류된 기사를 통해 관계도 분석, 키워드 트렌드, 연관어 분석이 가능하며 시각화 자료로는 그래프, 워드 클라우드 등이 있다. 그러나 매일 최신화되어 축적되는 정보의 양에 비해 이를 분류하는 체계 및 검색 기법은 미약한 것이 현 실정이다. 예를 들어, 빅카인즈의 통합 분류 체계는 정치, 경제, 사회, 문화, 국제, 지역, 스포츠, IT_과학 등 총 8개의 카테고리로 이루어져 있으며, 여기에 추가적으로 결과 내 재검색을 통해 세부 분류가 가능하다. 그러나 실제로 사용자가 원하는 세부 목적에 맞는 기사를 추출하기 위해서는 어떠한 키워드 및 검색 식을 사용해야 하는지가 난제이다. 한 예로 '북한 건설 관련 동향 파악'을 위해 기사를 추출한다고 가정하면, 빅카인즈의 카테고리 및 검색 기능을 활용하여 기사 분류를 하였을 때 검색에 사용할 최적의 키워드에 대한 기준이 불명확하다. 또한, 이를 분류하는 기존의 8개의 카테고리 또한 광범위하다. 이와 더불어 최적의 키워드를 추출했다고 할지라도 키워드를 검색 양식에 맞추어 어떠한 방식으로 입력해야 되는지도 알 수 없다.

이러한 문제점을 해결하고자 본 논문은 사용자의 세부 목적에 따른 최적의 검색 키워드 추출과 함께 검색 기법을 생성하는 방법론을 제안하고 있으며, '국내외 북한 건설 관련 동향 파악'을 예시로 적용하여 검증하였다. 우선 목적에 맞는 표본 기사를 선정하여 빈도수를 기준으로 키워드를 선정하였고, 이에 워드 임베딩(Word Embedding) 벡터를 구축하였다. 구축된 벡터들을 활용하여 키워드 간의 코사인 유사도(Cosine Similarity) 검사를 진행하였으며, 이를 통해 군집을 형성하였다. 형성된 군집은 군집 내의 키워드 간에는 'AND'를, 군집 간에는 'OR'을 사용하였으며 시각화 자료 및 정석적인 평가를 통해 최적의 키워드 추출 및 검색 기법 생성에 대한 작업을 수행하였다.

2. 선행 연구 및 이론적 배경

2.1 선행 연구

자연어 처리 분야는 같은 모델이라 할지라도 언어에 따라 다른 성능을 보인다. 특히 한국어는 교착어라는 특성을 갖고 있어 자연어 처리에 비교적 한계가 많다(An and Kim, 2015). 따라서 한국어를 대상으로 한 워드 임베딩 및 유사도 검사 활용 연구를 이공계열, 사회과학계열, 인문계열에 대해 조사·분석하였다. 이공계열에서는 건축물 설계 품질 자동 검토의 고도화를 위한 기초 연구(Song and Lee, 2018), 한국 법령정보를 워드 임베딩에 적용하여 연관 정보 검색 방법 연구(Kim and Kim, 2017), 교량 점검 보고서에서 손상 및 손상 인자를 자동으로 식별하는 방법 연구(Chung et al., 2018) 등이 수행되었다. 사회과학계열에서는 북한 관련 뉴스들에 대한 매체별 의제 설정 효과 측정 연구(Kim et al., 2020), 기업가 정신 관련 연구논문을 대상으로 한 동향 분석 연구(Yoo and Sung, 2021), 텍스트 마이닝을 활용한 소비자학 동향 분석 연구(Kim, 2020) 등이 있다. 인문계열에서는 노동신문의 이념적 어휘 연구를 통한 북한 사회문화 변화 양상 분석 연구(Cheong et al., 2020), 기술용어 분산 표현을 활용한 특허문헌 분류에 관한 연구(Choi and Choi, 2019), 온라인 뉴스 기사에서 추출된 키워드를 활용한 세부 주제별 토픽 추출 연구(Choi and Choi, 2018)가 진행되었다. 또한 워드 임베딩 벡터 구축 이후 코사인 유사도를 활용한 사례로는 게임 리뷰를 보다 명확하고 운영에 유용한 주제들로 자동 분류하는 시스템을 개발하는 연구(Yang et al., 2019) 등이 있다.

선행연구를 조사하고 분석한 결과, 워드 임베딩 및 코사인 유사도의 활용 방안은 대체적으로 토픽 모델링(Topic Modeling) 및 문헌 분류에 주로 사용되었다. 그러나 이를 기반으로 한 검색 기법 생성 연구는 미비하였다. 이에 따라 본 논문에서는 워드 임베딩 기법 및 코사인 유사도를 활용하여 최적의 검색 기법 생성 방안을 제안하고 검토하였다.

2.2 워드 임베딩 벡터

자연어 처리를 위해 기존에는 원-핫 인코딩(One-Hot Encoding)의 방식을 사용해왔다. 원-핫 인코딩이란 표현하고자 하는 단어의 인덱스(Index)에 1을 부여하고 다른 인덱스에는 0을 대입하는 것이다. 이로부터 얻어진 벡터를 원-핫 벡터(One-Hot Vector)라고 한다. 예를 들면 개, 고양이, 사자라는 3개의 단어가 있고 각각의 인덱스가 순차적으로 1, 2, 3이라고 할 때, 개에 대한 원-핫 벡터는 [1, 0, 0]이며 사자에 대한 원-핫 벡터는 [0, 0, 1]이 된다. 이러한 원-핫 벡터는 단어가 추가되는 개수에 따라 크기가 증가하며 해당하는 단어에 대한 인덱스를 제외하고는 모두 0의 값을 가져 저장 공간 측면에서 비효율적이다(Yoo and An, 2023). 또한 원-핫 인코딩으로 표현 시에는 의미적으로 유사한 값들을 분별해내지 못한다는 단점이 있다. 예를 들면 위에서 제시한 원-핫 벡터의 경우 사자와 유사성이 높은 단어에 대한 판단이 불가하다. 이러한 한계를 해결하기 위해 0, 1로만 표현했던 벡터를 실수화하였고 이를 밀집 벡터(Dense Vector)라고 한다. 예를 들면 10000개의 단어를 원-핫 벡터로 표현하면 [1 0 0 … 0]으로 표현되고 벡터의 차원은 10000이 된다. 이를 밀집 벡터로 표현하면 사용자가 지정한 차원으로 크기가 결정되고 각 인덱스에 실수값이 들어가게 된다. 예시로 개를 밀집 벡터로 표현하면 [0.2 0.5 0.1 0.6 −1.1 ...] 이 된다. 가령 차원을 128로 설정한다면 숫자의 개수는 128개가 된다. 이렇게 단어를 밀집 벡터 형식으로 표현하는 것을 워드 임베딩이라 하며 이로부터 얻어진 벡터를 워드 임베딩 벡터라 한다. 워드 임베딩 기법은 분산 표현(Distributed Representation)이라고도 하는데 이는 의미가 비슷한 단어 간의 벡터들이 비슷한 값을 가진다는 분포 가설(Distributional Hypothesis)(Harris, 1954)을 배경으로 하여 벡터 간 유사도 검사가 가능하다. 워드 임베딩 기법의 대표적인 모델로는 LSA, Word2Vec, FastText, Glove 등이 있다. Word2Vec의 대표적인 모델로는 CBOW(Continuous Bag of Words)와 Skip-gram이 있다. CBOW는 전체 문맥을 고려하여 중심 단어를 예측하는 방법이며 이와 반대로 Skip-gram은 중심 단어를 고려하여 전체 문맥을 예측하는 방법이다. 본 논문에서는 Word2Vec 모델 중 CBOW를 활용하였으며 구조는 Fig. 1과 같다.

CBOW는 총 3개의 층으로 이루어져 있고 각각 Input layer, Hidden layer, Output layer가 있다. Input layer는

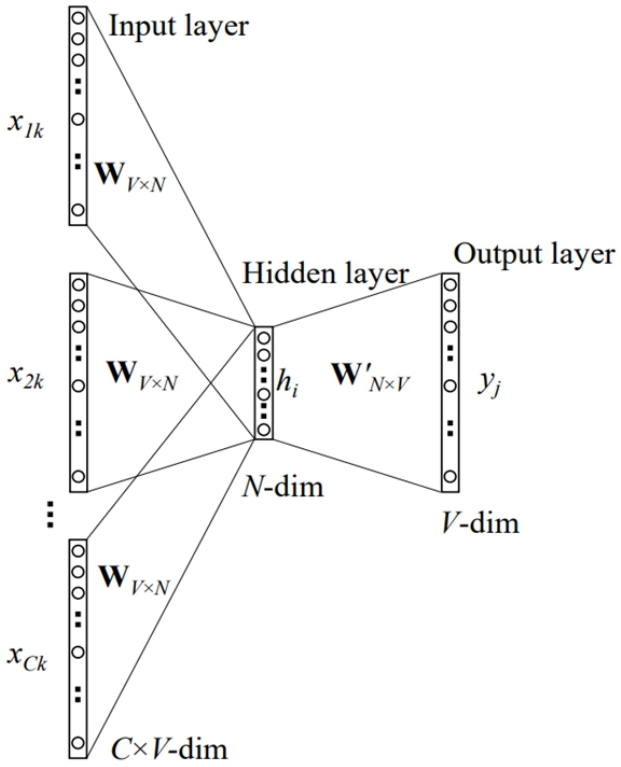

Fig. 1. CBOW (Continuous Bag of Words) Model Structure Presented by Rong (Rong, 2014)

사용자가 정한 중심 단어의 주변 단어에 대한 원-핫 벡터가 입력으로 들어간다. 입력으로 들어간 벡터는 가중치 W와 곱해진다. 이때 입력 벡터가 i번째 인덱스에 대한 원-핫 벡터이므로 W의 i 번째 행이 결과로 나오게 된다. Input layer의 가중치를 통과한 벡터들은 Hidden layer에서 평균값을 구하게 된다. 구해진 평균 벡터는 가중치 행렬 W'와의 연산을 통해 V 차원을 가진 벡터로 변환되어 Output layer으로 보내진다. Output layer로 보내진 벡터는 Soft-max 함수를 적용하여 0과 1사이의 실수값으로 표현된다. 이를 기반으로 역전파를 통해 가중치가 재설정이 되고 반복 학습을 하며 최종적으로 워드 임베딩 벡터가 형성된다.

2.3 코사인 유사도

벡터의 유사도를 검사하는 방법에는 코사인 유사도, 유클리드 거리(Euclidean Distance), 자카드 유사도(Jaccard Similarity) 등이 있다. 본 논문에서는 코사인 유사도를 사용하였다. 코사인 유사도는 두 벡터 A, B 간의 각을 기준으로 유사도를 측정하는 방법으로 계산식은 식 (1)과 같다. 같은 방향일 경우는 1, 90°일 경우는 0, 반대 방향의 경우

는 -1을 갖게 된다. 분모에는 각각의 벡터의 크기가 곱의 형태로 들어가고 분모는 두 벡터의 내적이 들어간다. 이로 인해 두 벡터의 사이각을 코사인 함수에 넣은 값이 유사도로 표현된다. 일반적으로 정보 검색 및 텍스트 마이닝 분야에서 활용되며 차원이 클수록 유사도를 명확히 구분할 수 있으며 벡터의 크기 값은 결과에 영향을 끼치지 않는다.

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

3. 키워드 추출 및 검색 기법 생성

검색 키워드 추출 및 기법 생성 과정은 Fig. 2와 같다. 빅카인즈의 기사를 분류하기 위한 최적의 검색 키워드 추출 및 기법 생성을 목적으로, '국내외 북한 건설 관련 동향 파악'을 예시로 검증하였다. 첫째로, 기사를 직접 추출하여 표본 기사를 선정하여 중요 키워드를 생성하였다. 키워드 생성은 빅카인즈의 키워드 추출 알고리즘인 토픽랭크 알고리즘을 사용하였다. 둘째로, 추출된 키워드에 대한 워드 임베딩 벡터를 생성하였다. 셋째로, 생성된 워드 임베딩 벡터를 활용하여 키워드 간의 유사도 검사를 실시하였다. 유사도 검사에서는 코사인 유사도를 활용하였다. 넷째로 빈도수 상위 10개에 대한 키워드를 기준으로 유사도가 0.5 이상인 키워드들을 포함시켜 10개의 군집을 생성하였다. 마지막으로, 생성된 군집들을 활용하여 빅카인즈의 검색 양식에 맞추어 최적의 검색 기법을 생성하였다.

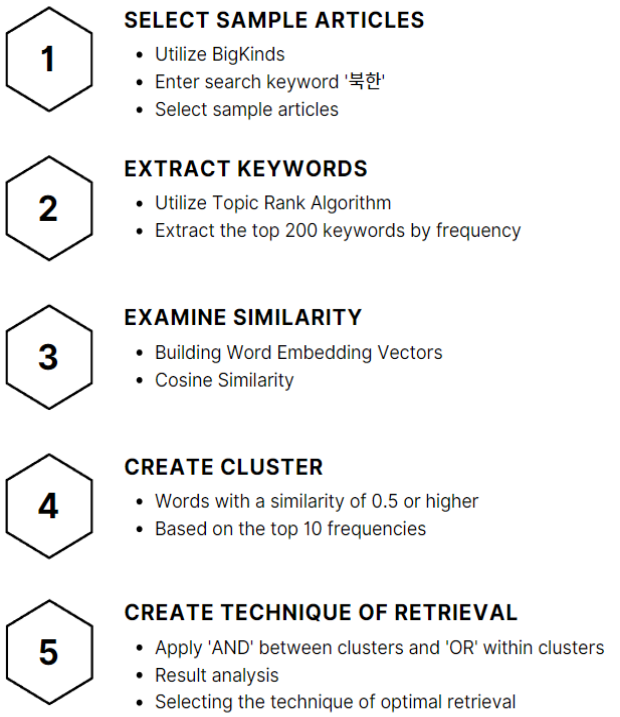

Fig. 2. Flowchart for Keyword Extration and Retrieval Technique

3.1 표본 기사 선정 및 키워드 추출

키워드 추출을 위한 표본 기사를 선정함에 있어 기간을 2022.09.01.~2022.11.30.으로 설정하고 기본 검색 키워드를 '북한'으로 지정하였다. 이후 '국내외 북한 건설 관련 동향 파악'에 부합하는 기사를 직접 추출하여 120개의 기사를 표본 기사로 선정하였다. 선정된 표본 기사에 빅카인즈의 토픽랭크 알고리즘을 적용하여 키워드 추출을 진행하였다. 추출된 키워드는 총 6787개이나 빈도수가 낮은 키워드는 성능 저하를 일으킬 수 있다고 판단하였다. 따라서 빈도수 기준 상위 200개의 키워드를 우선적으로 선정하였다. 상위 키워드 200개는 Table 1에서 확인할 수 있다.

3.2 워드 임베딩 벡터 구축

3.1에서 생성된 키워드를 기반으로 한국어 기반 사전학습(Pre-trained) 모델을 활용해 워드 임베딩 벡터를 구축하였으며, Fig. 3을 통해 관련 내용을 확인할 수 있다. 워드 임베딩 벡터 구축에 사용한 모델은 CBOW 모델로 vector size는 200, corpus size는 339M, Vocabulary size는 30185이다(Park, 2023). Fig. 4는 키워드 '북한'의 워드 임베딩 벡터 예시로 200개의 숫자로 표현되었다.

3.3 유사도 검사 및 군집 생성

키워드 간의 유사도를 검사하기 위해 코사인 유사도를 활용하였다. Fig. 5와 같이 Correlation matrix를 활용하여 빈도수 상위 10개의 키워드와 유사도가 0.5 이상인 키워드를 같은 군집으로 포함하였다. 결과적으로 Table 2와 같이 총 10개의 군집을 형성하였으며, 모든 키워드에 대해 군집을 형성하는 것은 기사 분류에 성능 저하를 일으킬 수 있으므로 제외하였다.

Table 1. List of top 200 keywords by frequency

Number	Keyword	Number	Keyword	Number	Keyword	Number	Keyword	Number	Keyword
1	북한	41	민방위	81	방안	121	참석	161	투자
2	시설	42	구축	82	회의	122	지시	162	방식
3	건설	43	이동	83	활용	123	개성	163	추가
4	사업	44	미국	84	로켓	124	미래	164	산업
5	지역	45	전쟁	85	확대	125	지적	165	작업
6	위성	46	설치	86	발전	126	가능성	166	관측
7	정부	47	안전	87	평가	127	활성	167	운용
8	평화	48	강화	88	설명	128	중심	168	비상
9	미사일	49	협력	89	도로	129	확인	169	재개
10	추진	50	조성	90	의원	130	예정	170	갱도
11	지원	51	공사	91	준비	131	국회	171	기간
12	주민	52	포착	92	부산	132	위원	172	공개
13	발사	53	보도	93	공격	133	관심	173	울산시
14	상황	54	해양	94	열차	134	관계자	174	엔진
15	위원장	55	노동자	95	대운하	135	노력	175	공원
16	대피	56	국가	96	개최	136	전략	176	가동
17	서해	57	이날	97	국제	137	유엔	177	이용
18	관광	58	안보	98	만큼	138	정보	178	전문가
19	러시아	59	대비	99	발생	139	인프라	179	체험
20	연결	60	운행	100	무기	140	지사	180	의미
21	철도	61	관리	101	정책	141	역할	181	대상
22	계획	62	유치	102	참여	142	상태	182	발사대
23	울산	63	생태	103	당국	143	투입	183	결과
24	건물	64	서울	104	대피소	144	징후	184	확장
25	발사장	65	평양	105	인천	145	원전	185	접경
26	도시	66	올림픽	106	규모	146	공간	186	연평도
27	한반도	67	활동	107	경기도	147	인천시	187	해체
28	남북	68	점검	108	위협	148	통합	188	성공
29	체계	69	모습	109	정상	149	글로벌	189	현장
30	가능	70	시장	110	물류	150	최대	190	발사체
31	한국	71	구상	111	대북	151	주장	191	전력
32	사진	72	시작	112	위치	152	이야기	192	주목
33	대통령	73	마련	113	확보	153	특별	193	군사
34	통일	74	대응	114	예산	154	자리	194	온실
35	도발	75	회장	115	개선	155	두만강	195	동해
36	진행	76	구간	116	내년	156	포함	196	사실
37	경제	77	변화	117	나라	157	화물	197	연장
38	중국	78	교수	118	주변	158	방문	198	협의
39	강조	79	주민	119	유지	159	관광객	199	수준
40	운영	80	세계	120	공동	160	민간	200	제공

3.4 검색 기법 추출

빈도수 상위 10개의 키워드에 대한 유사 키워드를 검토한 결과, 알맞게 배열되어 있음을 확인하였다. 이후 Table 3과 같이 빅카인즈 검색 양식을 적용하여 군집 내에는 'OR', 군집 간에는 'AND'를 대입하였다. 이와 더불어 키워드의 개수와 분류 성능은 상위 순위의 군집부터 시작하여 아래 순위의 군집을 포함하는 형식으로 진행하였다. '국내외 북한 건설 관련 동향 파악'에 부합하는 기사 추출을 목적으로 한 군집 적용 1차 방법으로 군집 1, (2, 3)과

	순번	용어	총합점수	keyword_vector
0	1	북한	34.894603	[0.2486978, -0.8847707, -0.9347374, -0.2326398...
1	2	시설	13.280148	[-0.84067506, 1.8924757, -1.3579122, 1.0384867...
2	3	건설	12.403362	[0.55848515, 1.4872274, 0.4933927, -2.7018063,...
3	4	사업	12.294296	[-0.37335056, 1.6146672, 0.09771919, 0.1945375...
4	5	지역	11.959652	[-0.2855183, 0.31378302, -0.8135753, 2.2986414...
...	...	...	...	...
195	196	사실	1.507692	[0.31787598, -2.1725864, -1.9555322, 1.4490211...
196	197	연장	1.507692	[0.41875264, 1.7283463, 0.41108057, -0.8767218...
197	198	협의	1.507692	[1.1261988, 1.4469838, -0.9143139, -1.1417241,...
198	199	수준	1.480801	[-0.4710171, -1.6423651, 0.21667932, 0.5343417...
199	200	제공	1.467355	[0.49686816, 0.50206107, -2.29371, 0.18946783,...

200 rows x 4 columns

Fig. 3. 200 Keyword Word Embedding

```
[ 0.2486978  -0.8847707  -0.9347374  -0.23263983  1.8970758  -1.16173
 -0.57918835 -0.8924936  -0.3768157  -0.18518075 -0.4161335   1.9047316
 -0.7778831   0.51254696  0.70542616 -0.10539569  0.9548685   0.26961955
 -0.01389012  1.3998398   1.1097412  -0.07920853 -0.67564535  0.16280118
  0.68326914  0.04360166  0.76472205 -1.7556192  -0.4840832   0.64922655
 -0.89105743 -1.4001704  -0.7126922  -1.0179977  -0.6261426  -2.4080012
  1.5713788   0.62829226  1.8153019  -0.26413062  0.9542449  -0.6308111
 -0.386876    0.23282605 -0.22317344  0.24222448  0.38980097 -0.95080215
  1.6283474   0.06524281  0.67488873  1.9117275  -0.35176325 -0.1923995
  0.5810412   0.0200636   2.1886697   0.97313935  1.1144085   0.00828805
  0.8394228   0.22108872 -0.37353724 -0.10869028 -1.3056719   1.4115642
  0.09018154 -0.78046596 -0.86056715 -0.4097849  -0.8243281   0.86761695
  0.7113783   2.1144798   0.7269868   1.042686   -0.3526522  -1.1194823
 -1.3219104   0.4824192  -0.8180067  -1.3364824  -0.147851   -1.0661035
 -0.551675   -0.32109493 -0.79381317  1.389041    0.48134613 -0.4557207
 -0.72605044  0.35707206  1.7158687  -0.2980089   1.0337183   0.28398904
  0.24527438  0.48223302 -0.29285863 -0.23719272  0.45478013  0.15277591
  1.5705265   1.0302799  -1.0721366   0.30702174 -0.2607731   0.94718593
 -1.9965591   1.8313867   0.508777   -1.9768238   0.8954022   1.7410827
 -0.4392656  -0.707059    0.78626627 -1.0776435  -2.085822   -0.42009744
  2.8167045   1.2474786   1.1213086   1.7439896   0.70457304  0.6426826
  0.8155653   0.28786576  0.11209501  0.9553492  -1.9253454  -2.283984
  0.33379343 -1.9074464  -0.8279475   1.712637    0.5976519  -0.40815428
  0.23974136  0.78237176 -1.9169147  -0.1617542   0.44883016  0.7886139
 -0.92149657  0.41785353 -1.1265228  -1.5381031  -0.08039241  1.0714904
 -1.6725781   0.25975868 -1.9738      1.0627644  -0.8867672  -0.37413135
  1.3220898   1.4098681   1.4801114   0.14969066 -1.6172514  -1.1845008
 -1.3230474   0.04082477 -0.20767744  0.78037554 -1.9586127  -0.11154819
  1.2932822   0.12073455  0.3018544   0.74926746  1.1970576   0.473617
  0.9895413   2.1636744  -1.8789783   0.41857845  1.1762631  -0.6436434
  2.8998144   2.2531104   0.3085538  -3.698696   -0.65370774  0.51402706
  0.21777803  0.92564183  0.02370971  1.0619551  -0.46658248  0.50489354
  1.3797712  -0.6858585  -0.27837005  2.0740707  -1.5506817  -0.2505416
 -0.12480228  0.6524813 ]
```

Fig. 4. Keyword ‘North Korea’ word embedding vector

군집 1, 2, 3을 비교하였다. 여기서 괄호는 해당 군집을 하나의 군집으로 가정함을 의미하며 이로 인해 군집 (2, 3)은 AND가 아닌 OR로 연결이 된다. 관련 내용은 Table 2, 3을 통해 확인할 수 있다. 비교 결과 군집 1, 2, 3의 기사가 군집 1, (2, 3)에 포함되며, 중복 기사를 삭제해 본 결과 6124개의 기사가 추출되었다. 추출 기사 내용 중 건설과 무관한 정치적 내용이 대부분을 이루어 군집 1, 2, 3을 선정하였다.

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.19	0.14	0.21	0.20	0.16	0.37	0.31	0.30	0.30
2	0.19	1.00	0.33	0.49	0.34	0.30	0.11	0.07	0.19	0.27
3	0.14	0.33	1.00	0.44	0.18	0.14	0.23	0.18	0.15	0.50
4	0.21	0.49	0.44	1.00	0.22	0.21	0.21	0.11	0.08	0.47
5	0.20	0.34	0.18	0.22	1.00	0.24	0.10	0.11	0.04	0.12
6	0.16	0.30	0.14	0.21	0.24	1.00	0.04	0.00	0.40	0.15
7	0.37	0.11	0.23	0.21	0.10	0.04	1.00	0.29	0.08	0.37
8	0.31	0.07	0.18	0.11	0.11	0.00	0.29	1.00	0.10	0.23
9	0.30	0.19	0.15	0.08	0.04	0.40	0.08	0.10	1.00	0.17
10	0.30	0.27	0.50	0.47	0.12	0.15	0.37	0.23	0.17	1.00
11	0.23	0.31	0.29	0.39	0.14	0.13	0.28	0.16	0.14	0.52
12	0.22	0.24	0.15	0.15	0.41	0.13	0.24	0.17	-0.02	0.12
13	0.21	0.03	0.27	-0.01	-0.08	0.47	0.07	0.00	0.55	0.27
14	0.21	0.14	-0.07	0.08	0.17	0.03	0.10	0.12	0.11	0.04
15	0.25	0.10	0.13	0.24	0.10	0.00	0.21	0.16	-0.08	0.32

Fig. 5. Part of the 200 keywords correlation matrix

Table 2. Top 10 Keywords Cluster For Frequency

Ranking	Keyword	Similar word
1	북한	한반도
2	시설	물류, 인프라, 건물
3	건설	설치, 조성, 공사, 구축, 추진
4	사업	산업, 물류, 방안, 투자, 공사, 작업
5	지역	도시
6	위성	발사체, 로켓
7	정부	국가, 당국, 국회
8	평화	안보
9	미사일	로켓, 발사장, 발사체, 발사대, 발사
10	평화	구상, 유치, 준비, 지원, 운영

군집 적용 2차 방법으로 군집 1, 2, 3과 군집 1, 2, 3, 4를 비교하였다. 비교 결과 군집 1, 2, 3, 4의 기사가 군집 1, 2, 3에 포함되었다. 포함된 기사를 삭제해 본 결과 379개의 기사가 추출되었다. 추출 기사 내용 중 DMZ, 북한 관련 관광지에 대한 내용이 다수 포함되었다. 불필요한 기사도 존재하였으나 건설과 관련한 유의미한 기사가 다수를 차지하여 군집 1, 2, 3을 선정하였다. 군집 적용 3차 방법으로 군집 1, 2, 3과 군집 1, 2, (3, 4)를 비교하였다. 비교 결과 군집 1, 2, 3의 기사가 군집 1, 2, (3, 4)에 포함되었다. 포함된 기사를 삭제해 본 결과 588개의 기사가 추출되었다. 추출 기사 내용 중 미사일 관련 정치적인 요소가 대부분이며, 필요 부분은 군집 1, 2, 3에 포함되는 내용임을 확인하였다. 군집 적용 4차 방법으로 군집 1에 키워드 ‘한반도’를 제외한 경우와 포함한 경우를 비교하였다. 비교 결과 ‘한반도’를 제외한 경우가 ‘한반도’를 포함한 경우에 모두 포함되었다. 포함된 기사를 삭제해 본 결과 453개의 기사가 추출되었다. 기사 내용 중 태풍 피해는 한반도 중심으로 서술되어 태풍 기사가 다수를 차지하였으며, 외교 관련

Table 3. Search Technique for Clusters Using Bigkinds search form

Including Cluster Number	BigKinds Search Form	Number of articles
Cluster 1, (2, 3)	(북한 OR 한반도) AND (시설 OR 물류 OR 인프라 OR 건물 OR 건설 OR 설치 OR 조성 OR 공사 OR 구축)	7348
Cluster 1, 2, 3	(북한 OR 한반도) AND (시설 OR 물류 OR 인프라 OR 건물) AND (건설 OR 설치 OR 조성 OR 공사 OR 구축)	1224
Cluster 1, 2, 3, 4	(북한 OR 한반도) AND (시설 OR 물류 OR 인프라 OR 건물) AND (건설 OR 설치 OR 조성 OR 공사 OR 구축) AND (사업 OR 산업 OR 물류 OR 방안 OR 투자 OR 공사 OR 작업)	845
Cluster 1, 2, (3, 4)	(북한 OR 한반도) AND (시설 OR 물류 OR 인프라 OR 건물) AND (건설 OR 설치 OR 조성 OR 공사 OR 구축 OR 사업 OR 산업 OR 물류 OR 방안 OR 투자 OR 공사 OR 작업)	1812
Cluster 1, 2, 3 (Except for '한반도')	(북한) AND (시설 OR 물류 OR 인프라 OR 건물) AND (건설 OR 설치 OR 조성 OR 공사 OR 구축)	771

기사에서 대한민국을 지칭할 때 한반도를 사용하여 건설과 관련이 없는 내용이 다수 포함되었다. 따라서 한반도 키워드를 제외한 군집 1, 2, 3을 최종 군집으로 선정하였다.

3.5 최적의 키워드 및 군집 선정 결과

'국내외 북한 건설 관련 동향 파악'에 부합하는 최적의 군집은 1, 2, 3('한반도' 제외)이 선정되었다. 이에 따른 실제 검색 기법은 (북한) AND (시설 OR 물류 OR 인프라 OR 건물) AND (건설 OR 설치 OR 조성 OR 공사 OR 구축)으로 형성되었다. 최적의 검색 기법을 빅카인즈에 입력해 본 결과 2022.09.01.~2022.11.30. 기간의 북한 관련 기사 28272개 중 771개의 기사로 압축되었다.

분류된 기사를 분석한 결과 771개 중 271개의 기사가 불필요한 기사로 분류되었으며 주요 특징은 다음과 같다. ① 하나의 기사 내에 다양한 주제가 있는 브리핑의 형식이다. 이러한 형식은 해당 날짜의 주요 이슈를 전반적으로 다루게 된다. 주요 키워드가 하나의 기사 내에 있다고 할지라도 다른 주제에 분포하고 있어 유의미한 기사로 보기에 한계가 있다. 이 경우 기사의 제목이 "브리핑"으로 시작하므로 일괄적인 삭제가 가능하다. ② 본인의 의견을 서술하는 '칼럼'의 형식이다. 이러한 경우 북한이라는 주제를 건설 이외에도 다양하게 접근하게 된다. 특히 건설과 관련한 글이라 할지라도 사실이 아닌 의견을 서술하였으므로 동향 파악에 적합한 형식이라고 하기에는 신뢰성이 부족하다. 이 경우 기사 제목이 '칼럼'으로 시작하므로 일괄적인 삭제가 가능하다. ③ 기사의 끝 단락에 북한의 비핵화에 대한 언급이 있다. 특히 정치 분야의 기사에서 북한과 무관한 주제이지만 마지막 단락에 북한의 비핵화를 언급하며 '북한'이라는 주요 키워드를 충족하게 된다. 이 경우 일괄적인 삭제는 불가하나 제목을 통해 불필요한 기사임을 확인할 수 있다.

4. 결론 및 고찰

본 논문에서는 다양한 목적에 맞춰 자료 조사·분류가 가능한 검색 기법을 제안하고 검증하였다. 우선 목적에 맞는 기사 일부를 선택하여 표본 기사를 생성하였으며, 토픽랭크 알고리즘을 통해 표본 기사의 키워드를 추출하였고 빈도수를 기준으로 상위 200개의 키워드를 선정하였다. 이를 기반으로 워드 임베딩 벡터를 구축하였고 코사인 유사도를 활용해 키워드 간 유사도를 검사하였다. 빈도수 상위 10개의 키워드를 기준으로 유사도가 0.5 이상인 키워드를 선별해 총 10개의 군집을 생성하였다. 생성된 군집에 빅카인즈의 검색 양식을 사용하여 최적의 검색 기법을 도출하였다. 제안한 방법은 '국내외 북한 건설 관련 동향 파악'을 중심으로 검증하였으며, 다음과 같은 결론을 도출하였다.

(1) 키워드 추출 시 토픽랭크 알고리즘과 빈도수를 기준으로 사용하였다. 그러나 이로 인해 사용자의 목적에 부합하지 않는 키워드가 추출될 가능성이 존재한다. 이에 따라 필수 포함 키워드를 사전에 지정하거나 중요 키워드에 가중치를 높인다면 키워드 순위에 신뢰성이 확보될 것이다.

(2) 워드 임베딩은 한국어 기반 사전학습 모델을 활용하였다. 기사로부터 추출된 키워드들에 워드 임베딩 벡터를 구축하여 유사도 검사를 실시한 결과, 실제로 비슷한 의미를 갖는 값들이 군집으로 형성되었다. 그러나 워드 임베딩 벡터는 트레이닝 셋에 따라 다르게 형

성될 수 있다. 따라서 본 논문에서 제시한 방법론을 통해 얻은 기사를 기반으로 전이 학습(Transfer learning)을 진행한다면 세부 목적에 맞는 워드 임베딩 벡터가 구축될 것이다.

(3) 군집화 생성과정에서 코사인 유사도 및 유사도 0.5 이상을 군집화의 기준으로 하였다. 코사인 유사도는 크기 값이 결과에 영향을 끼치지 않는다. 이로 인해 필요한 정보가 누락될 가능성이 있다. 또한 0.5라는 기준은 기사 분류 결과에 적합한 기준으로 보였으나 최적의 기준임을 보장하기에 한계가 있다. 따라서 다양한 유사도 검사 및 군집화 기법들을 추가 검토한다면 보다 좋은 성능을 낼 수 있는 군집화가 가능할 것이다.

(4) 기사 분류의 세부 목적은 '국내외 북한 건설 관련 동향 파악'이지만, 북한과의 교류가 직접적이지 않은 현 상태에서 충분한 표본 기사를 확보하는 것에는 한계가 있다. 따라서 본 논문에서 제안한 방법을 활용하여 1차적으로 기사를 추출한 이후 불필요한 기사를 삭제하고 이를 표본 기사로 활용하여 제안한 방법을 반복적으로 수행한다면 비교적 좋은 분류 성능을 내는 검색 기법을 생성할 수 있을 것이다. 특정 전공뿐만 아니라 모든 분야에서 방대하고, 다양한 자료들(논문, 기준 등)을 조사하고 분류하기 위한 검색 방법으로 활용 가능할 것으로 판단된다.

향후 다양한 주제에 대한 검증, 다른 분석 기법과의 비교 연구가 필요하며, 표본 기사 추출, 유사도 검사 기준과 같이 사용자가 직접 수행하거나 선택해야 하는 부분에 대한 자동화 연구가 수행된다면 사용자가 세부 목적을 입력하는 것만으로도 기사 분류가 가능해질 것으로 기대된다.

Acknowledgement

Research for this paper was carried out under the KICT Research Program(20230068-001, Research on the establishment of integrated and linked infrastructure for the co-prosperity of South and North Korea) funded by the Ministry of Science and ICT.

References

1. An, J. and Kim, H. W. (2015), "Building a Korean Sentiment Lexicon Using Collective Intelligence", *Journal of Intelligence and Information Systems*, Vol.21, No.2, pp.49-67.
2. Bigkinds (2023), http://www.bigkinds.or.kr.
3. Cheong, Y., Wang, G. and Song, S. (2020), "A Deep Learning-based Analysis of Ideological Words in Rodong Sinmun", *Korean Linguistics*, Vol.88, pp.213-245.
4. Choi, G. and Choi, S. P. (2018), "A Study on the Deduction of Social Issues Applying Word Embedding: With an Empasis on News Articles related to the Disables", *Journal of the Korean Society for Information Management*, Vol.35, No.1, pp.231-250.
5. Choi, Y. and Choi, S. P. (2019), "A Study on Patent Literature Classification Using Distributed Representation of Technical Terms", *Journal of the Korean Society for Library and Information Science*, Vol.53, No.2, pp.179-199.
6. Chung, S., Moon, S. and Choi, S. (2018), "Bridge Damage Factor Recognition from Inspection Reports Usin Deep Learning", *Journal of the Korean Society of Civil Engineers*, Vol.38, No.4, pp.621-625.
7. Harris, Z. S. (1954), "Distributional Structure", *WORD*, Vol.10, No.2-3, pp.146-162.
8. Kim, K., Kang, K., Son, M., Lee, C., Hong, S. and Kim, S. (2020), "A Big-Data Analysis of Issues on North Korea and Media Agenda Setting Functions: Applying Topic Modeling and Word-embedding Methods", *Peace and Democracy Institute*, Vol.28, No.1, pp.287-33.
9. Kim, K. O. (2020), "Analysis of Research Trends in Consumer Science through Text Mining", *Journal of Consumer Studies*, Vol.31, No.5, pp.19-47.
10. Kim, N. and Kim, H. J. (2017), "A Study on the Law2Vec Model for Searching Related Law", *Journal of Digital Contents Society*, Vol.18, No.7, pp.1419-1425.
11. Park, K. (2023), Pre-trained word vectors of 30+ languages, https://github.com/Kyubyong/wordvectors.
12. Rong, X. (2014), Word2vec parameter learning explained, Computation and Language(cs.CL).
13. Song, J. and Lee, J. K. (2018), "Approach to Word Embedding-based Semantic Analysis of Building Rule Checking-related Sentences for the Automated Rule Checking", *Korean Journal of Computational Design and Engineering*, Vol.23, No.4, pp.384-393.
14. Yang, Y. J., Lee, B. H., Kim, J. S., and Lee, K. Y. (2019), "Development of An Automatic Classification System for Game Reviews Based on Word Embedding and Vector Similarity", *The Journal of Society for e-Business Studies*, Vol.24, No.2, pp.1-14.
15. Yoo, S. H. and Sung, S. (2021), "Methodology for Semantic R&D Knowledge Clustering Analysis through Data Similarity Analysis: Entrepreneurship Research Field Study", *Journal of Business Research*, Vol.36, No.3, pp.167-180.
16. Yoo, W. and An, S. (2023), WikiDocs, https://wikidocs.net/book/2155.