creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조선대학교
CHOSUN UNIVERSITY

[UCI]I804:24011-200000267148

2019년 2월
석사학위 논문

단어 의미 관계를 고려한 연관 문서 추천 방법

조선대학교 산업기술융합대학원

소프트웨어융합공학과

김 선 미

조선대학교
CHOSUN UNIVERSITY

2019년 2월 석사학위논문

단어 의미 관계를 고려한 연관 문서 추천 방법

김 선 미

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

단어 의미 관계를 고려한 연관 문서 추천 방법

Method of Related Document Recommendation considering Semantic Relation between Words

2019년 2월 25일

조선대학교 산업기술융합대학원

소프트웨어융합공학과

김 선 미

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

단어 의미 관계를 고려한 연관 문서 추천 방법

지도교수 신 주 현

이 논문을 공학석사학위신청 논문으로 제출함

2018년 11월

조선대학교 산업기술융합대학원

소프트웨어융합공학과

김 선 미

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

김선미의 석사학위논문을 인준함

위원장 조선대학교 교수 김 판 구 (인)

위 원 조선대학교 교수 곽 근 창 (인)

위 원 조선대학교 교수 신 주 현 (인)

2018년 11월

조선대학교 산업기술융합대학원

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

목 차

ABSTRACT

I. 서론 ····· 1
A. 연구 배경 및 목적 ····· 1
B. 연구 내용 및 구성 ····· 3

II. 관련연구 ····· 4
A. TF-IDF ····· 4
B. 토픽 모델링(Topic modeling) ····· 5
1. LDA ····· 5
C. Word2vec ····· 7
D. PageRank와 TextRank ····· 8
1. PageRank ····· 8
2. TextRank ····· 9

III. 단어 의미 관계를 고려한 연관 문서 추천 ····· 11
A. 시스템 구성도 ····· 11
B. TF-IDF 가중치를 적용한 LDA기반 토픽 모델링 ····· 13
1. 데이터 수집 및 전처리 과정 ····· 13
2. TF-IDF 가중치 적용 ····· 16
3. LDA기반 토픽 모델링 ····· 17
C. 단어 간 유사도 추출 ····· 20
1. Word2vec ····· 20
D. 연관성 SCORE 도출 ····· 23
IV. 실험 및 평가 ····· 27
A. 실험 결과 및 성능 평가 ····· 27
B. 연관성 비교 실험 ····· 31

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

V. 결론 및 제언 ····· 33

참고문헌 ····· 34

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

표 목 차

[표 3-1] KAIST 품사 태그셋 ····· 14
[표 3-2] 명사의 종류 ····· 15
[표 3-3] 문서 내 주제 분포 예시 ····· 18
[표 3-4] '일자리'와 유사한 단어들의 유사도 ····· 22
[표 3-5] Distance Matrix 예시 ····· 22
[표 3-6] '비핵화' 키워드의 연관성 SCORE 정규화 결과 비교 ····· 26
[표 4-1] 실험 환경 ····· 27
[표 4-2] 연관성 SCORE 범위의 기사 내용 예시 ····· 28
[표 4-3] 상위 연관성 SCORE 범위 내 단어 중요도 결과 ····· 30
[표 4-4] 하위 연관성 SCORE 범위 내 단어 중요도 결과 ····· 30
[표 4-5] 상위 연관성 SCORE 범위 비교 실험 결과 ····· 31
[표 4-6] 하위 연관성 SCORE 범위 비교 실험 결과 ····· 31
[표 4-7] 방법론 비교 설명 ····· 32

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

그 림 목 차

[그림 1-1] 미디어 이용률 추이(1993년~2014년) ····· 2
[그림 1-2] 포털 뉴스 이용 빈도 추이(2011년~2014년) ····· 2
[그림 2-1] LDA 모델 구조 ····· 6
[그림 2-2] CBOW 모델과 Skip-gram 모델 구조 ····· 7
[그림 2-3] PageRank 알고리즘 개념도 ····· 8
[그림 2-4] PageRank 값 계산 과정 예시 ····· 9
[그림 2-5] PageRank 기반 TextRank ····· 10
[그림 3-1] 연관 문서 추천을 위한 시스템 구성도 ····· 11
[그림 3-2] 실험 데이터의 일부 ····· 13
[그림 3-3] 보통 명사 추출 알고리즘 ····· 15
[그림 3-4] 보통 명사 추출 예시 ····· 15
[그림 3-5] TF-IDF 가중치를 적용한 LDA기반 토픽 모델링 절차 ····· 16
[그림 3-6] TDM 크기 조절 예시 ····· 17
[그림 3-7] 주제 번호 및 확률 분포 ····· 18
[그림 3-8] 주제별 상위 단어 확률 분포 ····· 19
[그림 3-9] Word2vec 학습 ····· 20
[그림 3-10] Word embedding 벡터 값 예시 ····· 20
[그림 3-11] '일자리'와 유사한 단어군 ····· 21
[그림 3-12] 연관 문서 추천 프레임워크 ····· 23
[그림 3-13] 연관성 SCORE 도출 과정 ····· 24
[그림 3-14] 연관성 SCORE 도출 예시 ····· 25
[그림 4-1] TextRank 알고리즘을 이용한 문서의 핵심어 추출 ····· 29

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

ABSTRACT

Method of Related Document Recommendation considering Semantic Relation between Words

SeonMi Kim
Advisor : Prof. JuHyun Shin, Ph.D.
Department of Software Convergence Engineering
Graduate School of Industry Technology Convergence, Chosun University

The information retrieval technology is a representative technology of the information society and the ranking of information as a result of the search has been proven to influence people. Various researches have been conducted on the recommendation method of related documents in the field of information retrieval technology. Knowledge resource-based research, such as a dictionary of terms and ontology, requires human intervention and deployment costs and maintenance. Simple frequency-based research, such as TF-IDF, does not take into account the context of words and sentences, and it is impossible to interpret new words. Therefore, the efficiency of search is reduced.

News is characterized by having a wide range of information reported by various media companies. There are limitations to providing customized information because it deals with various incidents and it is composed of various keywords in one topic.

Therefore, in this paper, we use Korean news articles to extract topic distributions in documents and word distribution vectors in topics through

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

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.

As a result of document recommendation by relevance SCORE, it was confirmed that documents in SCORE range of higher association consist of words that are highly related to the query keyword. As a result of the document recommendation performance evaluation by the relevance SCORE using the TextRank algorithm which can recognize the importance of the words in the document, the query keyword has a higher importance as the higher relevance SCORE numerical value is larger. And the suitability of the proposed method was verified. The comparative experiment shows that semantic-based document recommendation is more effective than the existing document ranking methodologies TF-IDF and LDA.

The methods proposed in this paper can improve the performance of document searches by eliminating semantic ambiguity, and provide customized information by recommending documents that are most relevant to users' keywords, and identify the events associated with each keyword in the same topic.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅰ. 서론

A. 연구 배경 및 목적

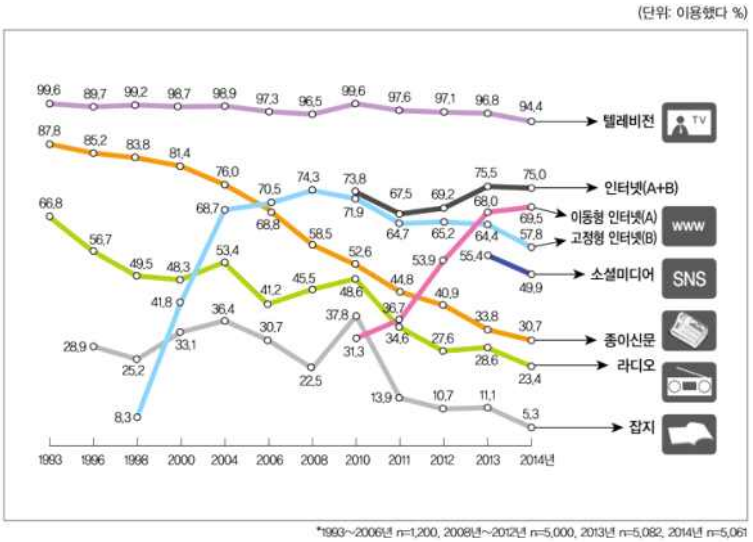
인터넷 기술의 발전과 스마트폰의 보급으로 사람들은 공간과 시간에 제약받지 않고 실시간으로 신속하고 편리하게 정보를 습득할 수 있다. 또한 SNS(Social Networking Service), 블로그 등 소셜 미디어(Social Media)가 활성화되면서 다양한 경로에서 정보를 공유할 수 있다. 하지만 대규모의 정보가 생성되고 있고 정보 과부하 문제로 인해 사람들은 필요한 정보를 찾아내는데 어려움을 겪고 있으며 그로 인해 빅데이터 처리 기술에 대한 관심도 급증하고 있다. 정보 검색 기술은 정보 사회를 대표하는 기술이며 검색 결과인 정보의 순위는 사람들에게 영향력을 미치는 것으로 검증되었다[1][2]. 정보 검색 기술 분야의 연관 문서 추천 방법에 대한 다양한 연구가 진행되고 있으며 기존의 용어 사전, 온톨로지와 같은 지식 리소스 기반의 연구는 사람의 개입과 구축비용, 유지 보수가 필요하다[3]. TF-IDF 같은 단순 빈도수 기반의 연구는 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 이용목적은 '커뮤니케이션' 다음으로 '자료/정보 습득'이 높았으며 자료/정보 습득 시 가장 많이 이용하는 방법 중 '뉴스'가 높은 순위를 차지하였다. [그림 1-1]은 1993년부터 2014년까지 미디어 이용률 추이를 나타내며, [그림 1-2]는 2011년부터 2014년까지 포털 뉴스 이용 빈도 추이를 나타내는 그래프이다[4].

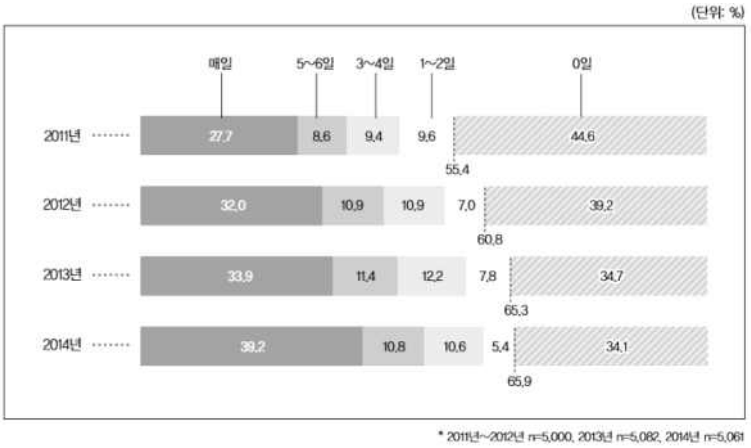
뉴스는 다양한 언론사를 통해 보도되고 넓고 방대한 정보 범위를 가지는 것을 특징으로 한다. 다양한 사건을 다루기 때문에 여러 주제들을 내포하고 있으며 하나의 주제 속에 매우 다양한 키워드로 이루어져 있으므로 사용자 맞춤형 정보를 제공하는 데 한계가 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 1-1] 미디어 이용률 추이(1993년~2014년)



[그림 1-2] 포털 뉴스 이용 빈도 추이(2011년~2014년)

따라서 본 논문에서는 단어의 의미적 모호성을 해소하고 효율적인 정보 검색을 위해 한국어 뉴스 기사를 이용하여 LDA 기반 토픽 모델링을 통해 문서 내 주제 분포와 주제 내 단어 분포 벡터를 추출하고 Word2vec을 이용해 단어를 벡터화한 후 가중치 행렬을 생성하여 단어 간의 의미 관계를 고려해 연관성 SCORE를 도출한 다음 점수가 높은 순서대로 문서를 추천하는 방법을 제안한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 연구 내용 및 구성

본 논문에서는 NAVER에서 제공하는 정치 카테고리의 뉴스 기사를 수집하여 전처리 작업을 수행하고 LDA 기반 토픽 모델링을 통해 문서 내 주제 분포와 주제 내 단어 분포 벡터를 추출하며 Word2vec을 사용하여 단어를 벡터화한 후 단어 간 유사도를 구한다. 가중치 행렬을 생성하여 단어 간 의미적 연관성을 고려하고 연관성 SCORE를 도출하여 수치가 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 의미 관계를 고려한 연관 문서 추천 방법에 대하여 설명한다. 4장에서는 제안한 방법을 적용하여 성능 평가 및 비교 실험을 진행하고 5장에서 결론과 향후 연구 방향에 대하여 기술한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅱ. 관련연구

A. TF-IDF

TF-IDF는 단어 빈도와 역문서 빈도의 곱으로 식 (1)과 같이 표현되며 각각의 문서 내에 있는 단어들에 수치 값을 주는 가중치 알고리즘이다[5].

$$tf-idf_{t,d,D} = tf_{t,d} \times idf_{t,D} \tag{1}$$

TF(Term Frequency)는 문서 d 내 특정 단어 t의 빈도를 의미하고 IDF(Inverse Document Frequency)는 특정 단어 t가 나타난 문서 수의 역수를 취한 값으로 식 (2)와 같이 나타낸다.

$$idf_{t,D} = \log(\frac{D}{df_t}) \tag{2}$$

- D : 전체 문서 수
- d : 특정 문서
- t : 특정 단어
- df_t : 특정 단어 t가 포함된 문서의 수

따라서 특정 단어의 중요도는 단어의 출현 횟수에 비례하고 그 단어가 언급된 모든 문서의 총 수에 반비례하며 TF-IDF 값이 큰 단어일수록 다른 문서에는 많지 않고 해당 문서에서 자주 등장하는 단어를 의미한다.

TF-IDF는 문서 내 단어의 중요도나 문서 간의 유사도를 추출하거나 검색 엔진에서 검색 결과의 순위를 결정하는 등 다양한 용도로 사용되고 있다[6][7][8][9][10].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 토픽모델링(Topic modeling)

토픽 모델링은 데이터마이닝에서 비 구조화된 대량의 문서 집합에서 잠재되어있는 주제(토픽)를 추출하여 숨겨진 의미 구조를 발견하기 위해 사용하는 방법론이다. 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.

대표적인 알고리즘으로는 잠재 의미 분석(Latent Semantic Analysis, LSA)과 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)가 있으며 HDP(Hierarchical Dirichlet Process), ATM(Author Topic Model), DMR(Dirichlet Multinomial Regression)과 같이 여러 확장된 기법들이 있다[11][12][13].

1. LDA

LDA는 토픽 모델링 기법 중 가장 주목 받고 있는 기법으로 자연어 처리에서 주어진 대량의 문서들에 대해 각 문서에 무슨 토픽들이 존재하고 있는지에 대한 확률 모형이며 여러 주제가 혼합된 문서를 다룰 수 있게 하였다[14].

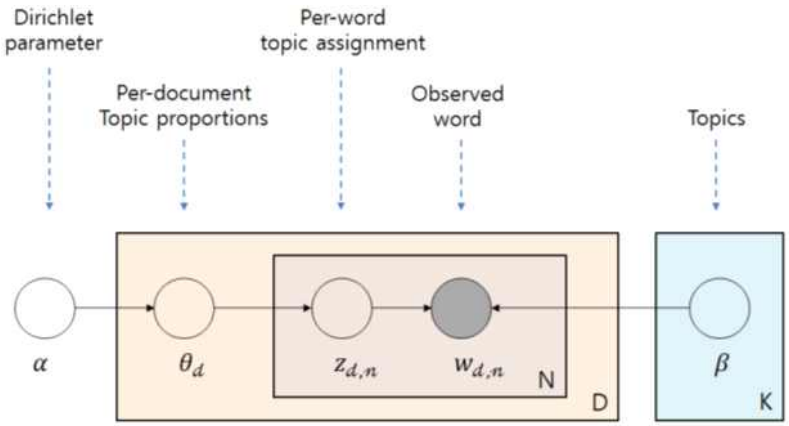
LDA 모델은 [그림 2-1]과 같은 구조이며 기호의 뜻은 다음과 같다.

- Corpus 레벨의 매개변수
  - K : 주제의 개수
  - α : 주제들의 Dirichlet 분포
    주제들이 어떤 패턴인지 나타나게 된다.
  - β : 단어들의 Dirichlet 분포
    단어들이 어떤 패턴인지 나타나게 된다.
- 문서 레벨의 변수
  - D : 문서의 개수
  - θ : 하나의 문서에 대한 주제어 분포
- 단어 레벨의 변수

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

- N : 단어의 개수
- z : 주어진 주제에 대한 단어 확률
- w : 주어진 주제에 대한 실제 관찰된 단어



[그림 2-1] LDA 모델 구조

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 주제에 해당하는 단어들의 확률 분포 z도 나타낼 수 있다. 따라서 본 논문에서는 문서 내 주제 분포와 주제 내 단어 분포 벡터를 추출하기 위하여 LDA를 사용하였다.

심준식의 연구에서는 LDA 기반 토픽 모델링을 이용하여 주제어 태그를 부여하고, 주제어 태그의 단어 그룹과 코사인 유사도를 계산하여 유사한 판례 그룹을 분류하는 시스템을 제안하였다[15].

곽창욱의 연구에서는 대본을 기반으로 학습된 무한 사전 온라인 LDA 토픽 모델에 외부 블로그 데이터를 추가 학습하여 Word2Vec를 이용해 의미적 연관 관계를 반영하여 토픽을 확장했다[16].

조태민의 연구에서는 LDA 기반 토픽 모델링을 이용해 유사한 문서의 키워드를 후보 키워드를 선택하였다. 후보 키워드를 구성하는 단어들의 등장 확률을 이용해 후보 키워드의 중요도를 평가하여 문서의 잠재 키워드를 추출하였다[17].

김성호의 연구에서는 LDA 기반 토픽 모델링을 이용해 악성 트래픽의 자동 분석과 악성 시그니처 추출을 통하여 탐지규칙 생성 기법을 제안하였다[18].

강범일의 연구에서는 LDA 기반 토픽 모델링을 통해 신문기사의 주제를 추출하여

Collection @ chosun

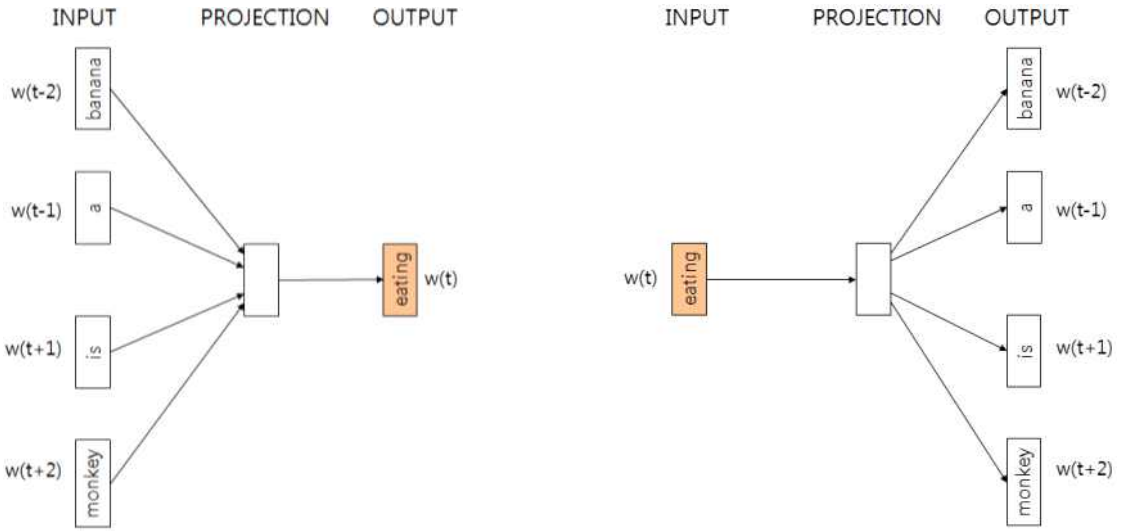
조선대학교
CHOSUN UNIVERSITY

오피니언 마이닝을 수행하였다[19].

다음과 같이 LDA는 텍스트 마이닝, 정보 보안, 오피니언 마이닝 등 여러 분야의 연구에서 다방면으로 이용되고 있다.

C. Word2vec

Word2vec은 단어 임베딩(Word embedding)의 대표적인 알고리즘으로 빠른 학습 속도와 좋은 성능을 가지고 있다. 단어 임베딩(Word embedding)은 딥러닝(Deep Learning) 분야에서 텍스트를 구성하는 각각의 단어를 수치화하는 방법이며 Word2vec은 단어를 벡터 공간에 표현한다. Word2vec은 문장을 구성하는 단어들의 전후 관계를 인공신경망에 학습시켜 단어의 의미를 내포한다[20].



[그림 2-2] (좌) CBOW 모델과 (우) Skip-gram 모델 구조

[그림 2-2]는 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[21].

이와 같이 Word2vec은 단어의 의미와 문장에서의 맥락을 고려하여 단어를 벡터로

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

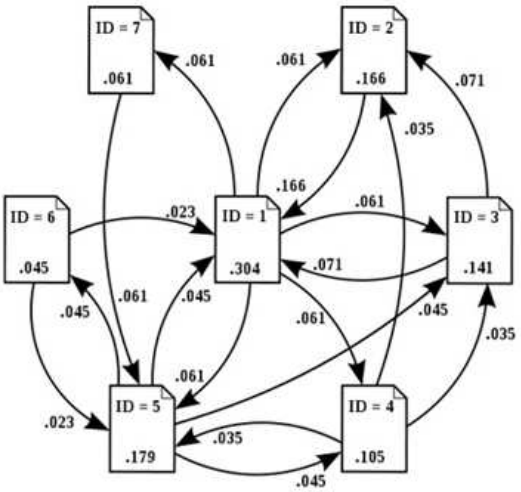
표현하기 때문에 의미적으로 유사한 단어들끼리 근접한 벡터 공간에 위치하게 된다. 같은 단어라도 단어의 의미와 맥락에 따라 다른 벡터 공간에 학습될 수 있다는 것을 의미한다. 따라서 본 논문에서는 단어 벡터들 간의 거리를 코사인 유사도를 통해 계산하여 단어 간 의미적 유사도를 구하기 위해 Word2vec을 사용하였다.

이앞길의 연구에서는 Word2vec을 이용해 발명 명칭에 대하여 특허 문서 간 유사도를 계산하였고 LDA를 통해 초록과 청구항에 대하여 특허 문서 간 유사도를 구해 특허 문서를 추천하였다[22].

D. PageRank와 TextRank

1. PageRank

PageRank는 Google 검색 엔진의 핵심 기술로 하이퍼링크 구조를 가지는 웹 문서에 상대적 중요도에 따라 가중치를 부여하는 방법이다[23]. [그림 2-3]과 같이 각각의 웹 페이지를 정점(node)으로, 페이지와 페이지를 연결하는 링크를 간선(edge)으로 간주하는 그래프를 대상으로 하는 알고리즘이다.

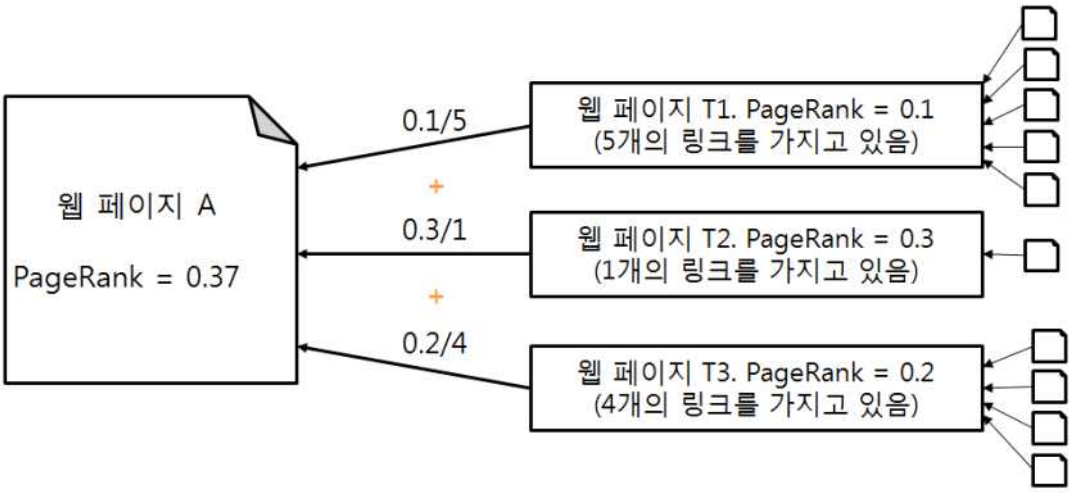


[그림 2-3] PageRank 알고리즘 개념도

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

더 중요한 페이지는 더 많은 다른 사이트로부터 링크를 받는다는 관찰에 기초하고 있으며 PageRank 값이 클수록 다른 사이트가 많이 참조되었다고 해석할 수 있다[24]. 페이지 간 PageRank 값을 주고받는 것을 반복하다보면, 가중치가 특정한 값에 수렴하는 것을 통해 각 페이지의 최종 PageRank를 계산하여 중요한 핵심 페이지를 산출할 수 있다[25]. PageRank 계산 과정은 [그림 2-4]와 같이 표현할 수 있다.



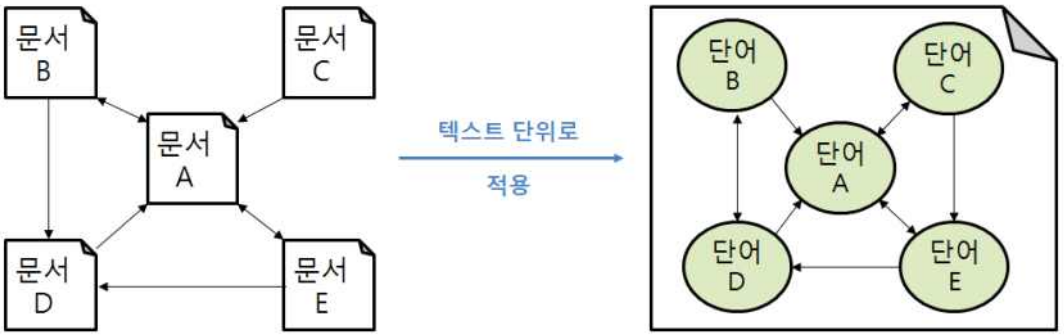
[그림 2-4] PageRank 값 계산 과정 예시

2. TextRank

TextRank는 [그림 2-5]와 같이 PageRank 알고리즘을 기반으로 중요도가 높은 웹 사이트는 다른 많은 사이트로부터 링크를 받는다는 점을 착안하여 텍스트를 단위로 적용시킨 알고리즘이다[26].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 2-5] PageRank 기반 TextRank

문서 내 문장이나 단어를 이용하여 기존 Pagerank의 원리와 같이 핵심어구나 문장을 추출하며 TextRank 값은 식 (3)과 같이 계산할 수 있다.

$$TR(V_i) = (1-d) + d * \sum_{V_j \in (V_i)} \frac{w_{ji}}{\sum_{V_{k \in Out(V_j)}} w_{jk}} TR(V_j) \qquad (3)$$

- $TR(V_i)$: 문장 또는 단어 V_i에 대한 TextRank 값
- w_{ij} : 문장 또는 단어 i와 j 사이의 가중치
- d : 제동 계수(Damping factor)

안인석의 연구에서는 Twitter의 타임라인을 요약하기 위해 TextRank 알고리즘에 비슷한 키워드들을 하나로 묶어주는 산봉우리 개념을 적용한 알고리즘을 제안하였고 기존 빈도수 기반의 요약 방법보다 효과적인 것을 확인하였다[27].

이원형의 연구에서는 PageRank 알고리즘 기반의 가중치를 제안하여 Twitter에서 화제성이 높은 단어를 추출하였다[28].

Collection @ chosun

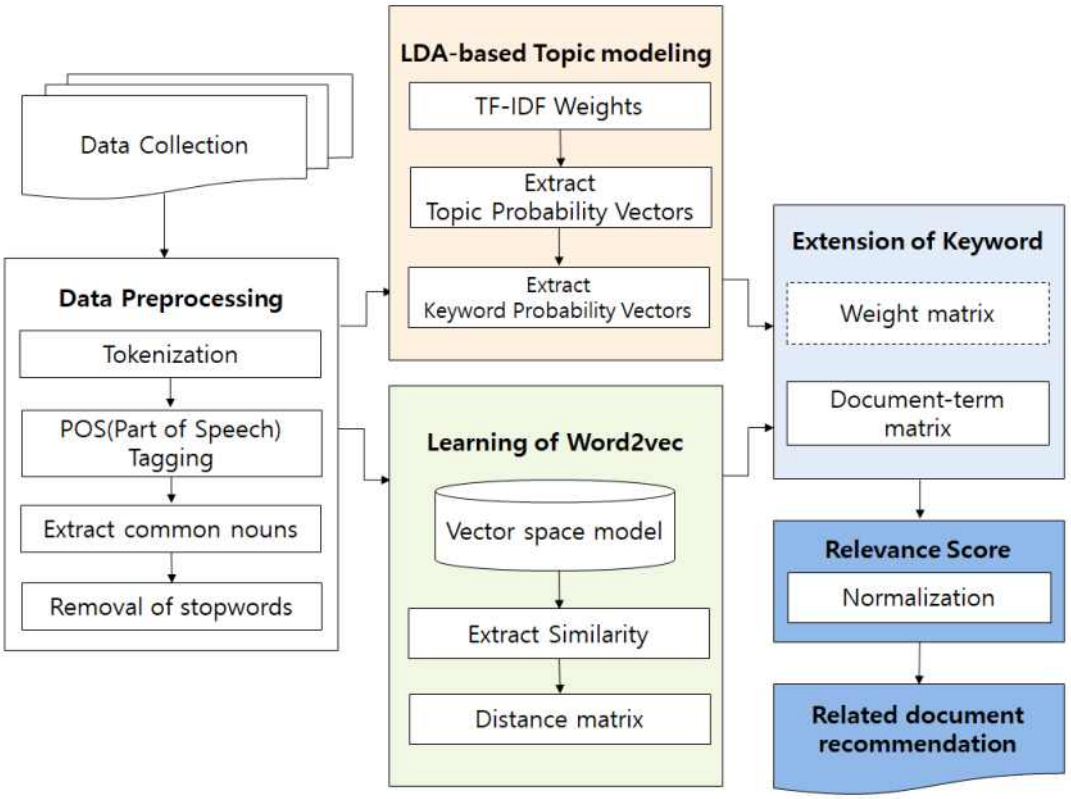
조선대학교
CHOSUN UNIVERSITY

Ⅲ. 단어 의미 관계를 고려한 연관 문서 추천

본 논문에서는 단어의 의미적 모호성을 해소하고 효율적인 정보 검색을 위해 한국어 뉴스 기사를 이용하여 LDA 기반 토픽 모델링을 통해 문서 내 주제 분포와 주제 내 단어 분포 벡터를 추출하고 Word2vec을 이용해 단어를 벡터화한 후 가중치 행렬을 생성하여 단어 간의 의미 관계를 고려해 연관성 SCORE를 도출한 다음 점수가 높은 순서대로 문서를 추천하는 방법을 제안한다.

A. 시스템 구성도

[그림 3-1]은 본 논문에서 제안하는 의미 기반 문서 추천 방법의 시스템 구성도이다.



[그림 3-1] 연관 문서 추천을 위한 시스템 구성도

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

실험을 위한 데이터는 네이버에서 제공하는 정치 카테고리의 뉴스 기사로 선정하여 2018.07.01.부터 2018.07.31.까지 총 24,887개의 기사 내용을 수집했다.

보통 명사를 추출하고 불용어를 제거하는 전처리 작업을 수행한 후 먼저, TF-IDF 가중치를 부여하여 LDA 기반 토픽 모델링을 수행하고 문서 내 주제 분포와 주제 내 단어 분포를 추출한다. 두 번째로는 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 간의 의미 관계를 고려해 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.

Collection @ chosun

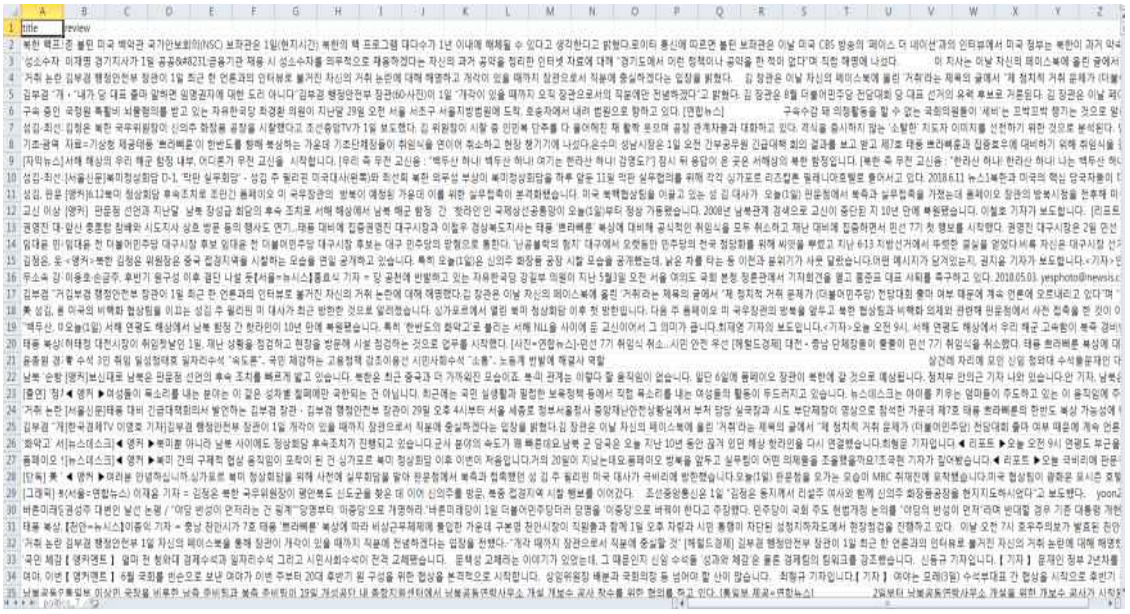
조선대학교
CHOSUN UNIVERSITY

B. TF-IDF 가중치를 적용한 LDA기반 토픽 모델링

본 절에서는 한국어 뉴스 기사 데이터로 전처리 작업을 수행한 후 TF-IDF 가중치를 적용한 LDA기반 토픽모델링을 통해 문서 내 주제 분포와 주제 내 단어 분포를 추출하는 방법에 대해 기술한다.

1. 데이터 수집 및 전처리 과정

본 연구는 네이버(www.naver.com)에서 제공하는 정치 카테고리의 뉴스 기사를 Java 기반 환경에서 웹 크롤링하여 2018.07.01.부터 2018.07.31.까지 총 24,887개의 기사를 수집하여 [그림 3-2]와 같이 .csv 파일로 저장했다.



[그림 3-2] 실험 데이터의 일부

R studio에서 한국어 자연어 처리를 위해 KoNLP 라이브러리를 사용하며 수집한 데이터가 저장된 .csv 파일을 가져온다. 문장을 어절 단위로 토큰화(Tokenizing) 시키고 한나눔 한국어 형태소 분석기에서 지정한 69개의 품사 중 22개의 품사를 기준으로 분

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

류해주는 SimplePos22() 함수를 이용한 형태소 분석을 통해 단어들의 품사를 판별하여 POS(Part of Speech) Tagging 작업을 수행한다. 다음 [표 3-1]은 품사 판별에 사용한 KAIST 품사 태그 집합을 나타낸다.

[표 3-1] KAIST 품사 태그셋

POS Tag	품사	POS Tag	품사
S	기호	MA	부사
F	외국어	II	감탄사
NC	보통명사	JC	격조사
NQ	고유명사	JX	보조사
NB	의존명사	JP	서술격조사
NP	대명사	EP	선어말어미
NN	수사	EC	연결어미
PV	동사	ET	전성어미
PA	형용사	EF	종결어미
PX	보조용언	XP	접두사
MM	관형사	XS	접미사

효과적인 LDA 학습을 위해 [표 3-2]와 같은 명사의 종류 중 두 글자 이상의 보통명사(NC)를 추출하고 Corpus를 생성한다. [그림 3-3]은 이에 대한 과정을 나타내는 알고리즘이다. 숫자, 공백, 기호, 반복되는 의미 없는 단어 등을 불용어로 지정하여 제거하는 단계를 마지막으로 텍스트 전처리 작업을 수행했다. [그림 3-4]는 POS Tagging 된 데이터에서 보통 명사에 해당하는 단어만 추출한 데이터의 예시이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-2] 명사의 종류

대분류	소분류	세분류
체언(N)	보통명사(NC)	서술성 명사(ncp) 비서술성 명사(ncn)
	고유명사(NQ)	성(nqpa) 성+이름(nqpc) 이름(nqpb) 기타-일반(nqq)
	의존명사(NB)	단위성 의존명사(nbu) 비단위성 의존명사(nbs) 비단위성 의존명사(nbn)
	대명사(NP)	인칭대명사(npp) 지시대명사(npd)
	수사(NN)	양수사(nnc) 서수사(nno)

```
#Extract common nouns

for (i in 1:length(doc$review)) {
  a1 <- tryCatch({
    q <- as.character(doc$review[i])
    pos <- unlist(SimplePos22(q))

    extracted <- str_match(pos, '([가-힣]+)/NC')

    keyword <- extracted[, 2]

  }, error = function(err){
    print(paste("line=", as.character(i), err))
    return (NULL)
  })
  if (!is.null(a1)){
    pos <- c(keyword[!is.na(keyword)])
    z = "kkk"
    z <- paste(pos, collapse = " ")

    doc$review[i] <- z
  }
}

corp=VCorpus(VectorSource(parsedDataRe$pContent))
```

[그림 3-3] 보통 명사 추출 알고리즘

지방규제혁신/NC	지방규제혁신
우수사례/NC	우수사례
경진대회/NC	경진대회
우수상/NC	우수상
국무총리상과/NC	국무총리상과
NA	NA
NA	NA
상금/NC	상금
수상/NC	수상
NA	NA
NA	NA
NA	NA
이번/NC	이번
대회/NC	대회
행정안전부/NC	행정안전부
주관/NC	주관
국민생활/NC	국민생활
불편해소/NC	불편해소
투자/NC	투자
유치/NC	유치
일자리/NC	일자리
창출/NC	창출

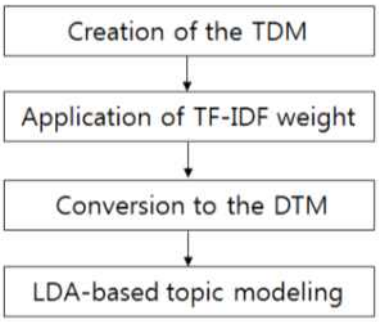
[그림 3-4] 보통 명사 추출 예시

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. TF-IDF 가중치 적용

LDA 기반 토픽 모델링은 [그림 3-5]와같이 진행하였다.



[그림 3-5] TF-IDF 가중치를 적용한 LDA 기반 토픽 모델링 절차

Corpus를 텍스트 문서 형식으로 변환하고 단어가 문서에 몇 회 출현하였는지 나타내는 TDM(Term Document Matrix)을 생성한다.

TDM은 단어들의 단순 빈도수를 나타내기 때문에 빈도수가 적은 단어들은 중요도가 떨어지게 된다. 따라서 [그림 3-5]와같이 어떤 단어가 특정 문서에서 얼마나 중요한지 나타내주는 TF-IDF 가중치를 부여하여 단어별 TF-IDF 분포 값을 기준으로 TDM의 크기를 조절해 성능을 향상시켰다. 그 결과 '오늘', '이번', '관련', '당시'와 같은 의미가 중요하지 않지만 자주 등장하는 단어를 제거할 수 있었다. TDM을 각 문서에 대한 단어의 출현을 행렬로 표현하는 DTM(Document Term Matrix)으로 변환하여 LDA 분석을 수행하였다. [그림 3-6]은 TDM 크기 조절 예시를 시각화한 것이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

TDM(Term Document Matrix)

	Doc1		Doc2		Doc3		Doc4		Doc5	
대통령	1	7.8476	3	6.0484	2	8.3747	2	7.5648	3	7.5123
최저임금	0	5.3177	5	1.4778	0	3.1722	1	8.1474	0	3.1587
이산가족	0	3.1447	0	3.1578	0	9.2417	2	3.4862	0	9.5755
비핵화	1	1.2458	0	9.4782	2	7.5789	0	2.6872	5	6.4532
통일	1	1.8789	0	6.4872	0	7.1248	3	2.8897	2	7.9521
관련	6	0.0121	15	0.0235	17	0.0187	8	0.0014	11	0.0077
이번	17	0.0126	9	0.0056	12	0.0258	10	0.0127	9	0.0268
오늘	10	0.0124	7	0.0175	8	0.0024	7	0.0248	13	0.0123
당시	15	0.0034	11	0.0178	17	0.0548	13	0.0156	16	0.0045

TF-IDF

[그림 3-6] TDM 크기 조절 예시

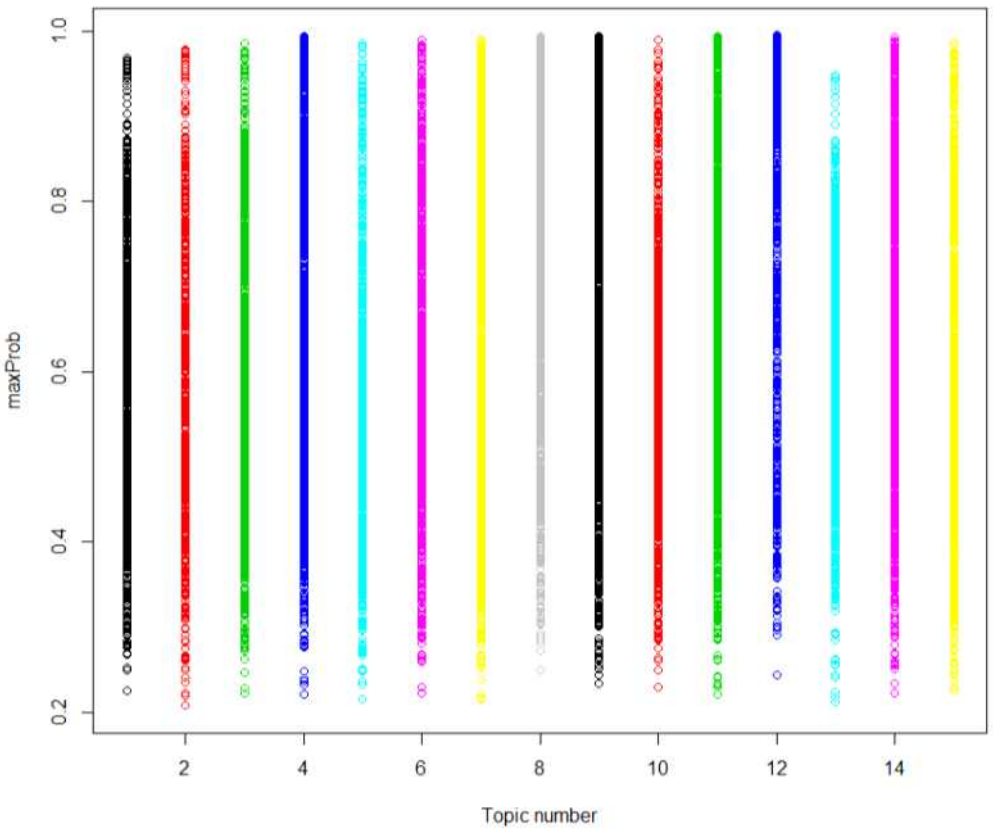
3. LDA기반 토픽 모델링

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.

[그림 3-7]은 각 문서들이 갖는 주제 번호 및 확률 값의 분포를 시각화한 것이다. X 좌표는 Topic 번호를 뜻하고 총 15개의 주제를 나타내며 Y 좌표는 15개 각 주제들에 대한 문서들의 확률 값을 의미한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 3-7] 주제 번호 및 확률 분포

[표 3-3]은 문서 내 주제 분포의 예시를 나타낸다. 사용자 관심문서가 가장 높은 확률 분포를 가지는 주제를 선택하고 해당 주제의 주제어를 키워드로 지정한다.

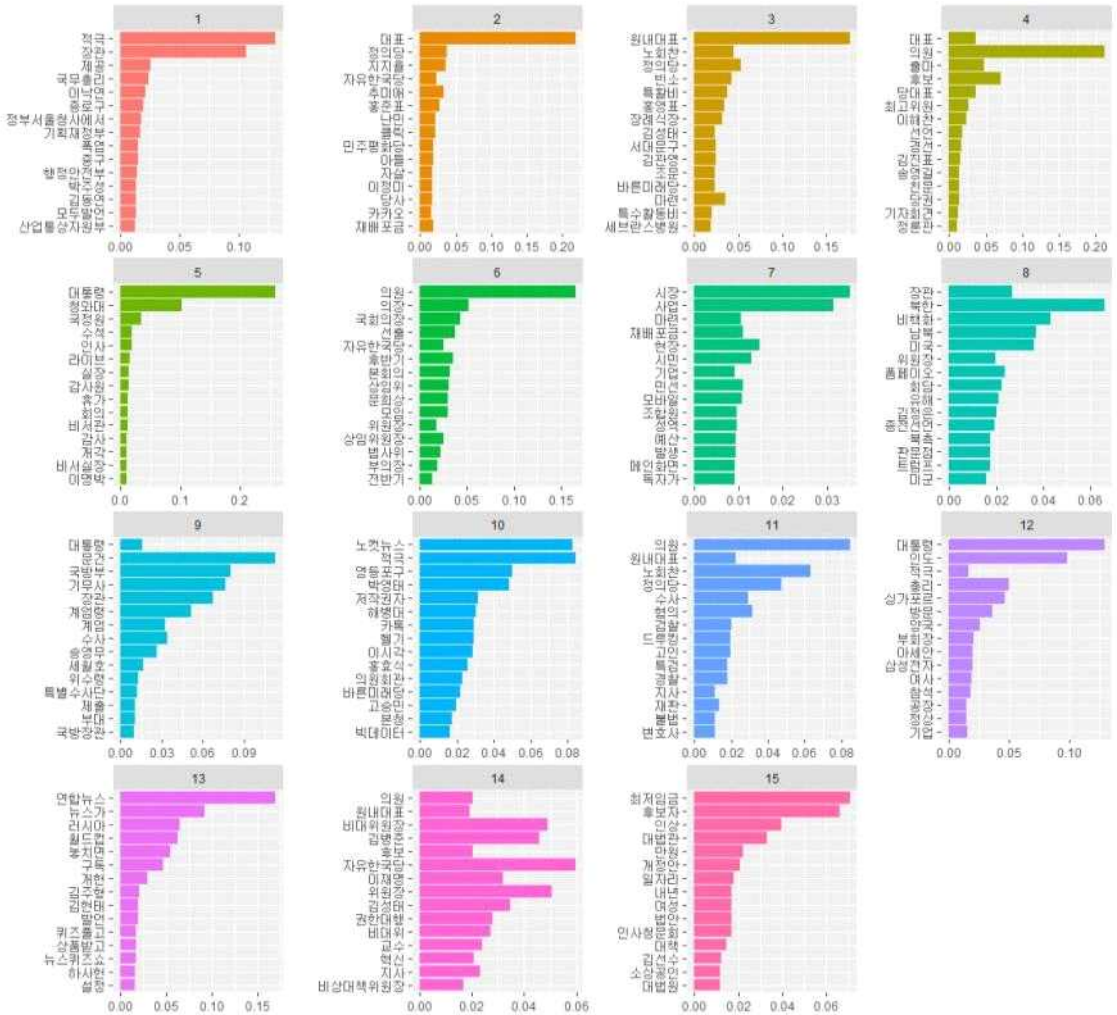
[표 3-3] 문서 내 주제 분포 예시

Document	Topic 1	Topic 2	...	Topic 15
1	0.001929	0.001929	...	0.001929
2	0.001929	0.001929	...	0.001929
3	0.481501	0.003192	...	0.003192
4	0.286055	0.156239	...	0.001635
5	0.062923	0.000944	...	0.070899

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-8]은 [그림 3-7]에 나타난 15개 각 주제들에 해당하는 상위 단어들의 확률 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.



[그림 3-8] 주제별 상위 단어 확률 분포

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

C. 단어 간 유사도 추출

1. Word2vec

전처리 작업을 거친 데이터를 [그림 3-9]와 같이 Word2vec을 이용하여 200차원, 대용량 데이터에 성능이 좋은 Skip-gram 방식으로 학습하였다. 학습 결과인 단어 벡터 값들을 Vector Space Model로 구축했다.

```
> writeLines(doc$review, 'question_tag.txt')
>
> w2v_model <- train_word2vec('question_tag.txt', 'question_vectors_tag.bin',
+                             vectors=200, threads=4, window=10, iter=10, negative_samples=4, force=TRUE)
Starting training using file C:/Users/chosun/Documents/question_tag.txt
Vocab size: 27391
Words in train file: 2560252
```

[그림 3-9] Word2vec 학습

단어들은 [그림 3-10]과 같이 벡터 공간에 단어 임베딩(Word embedding)되었다.

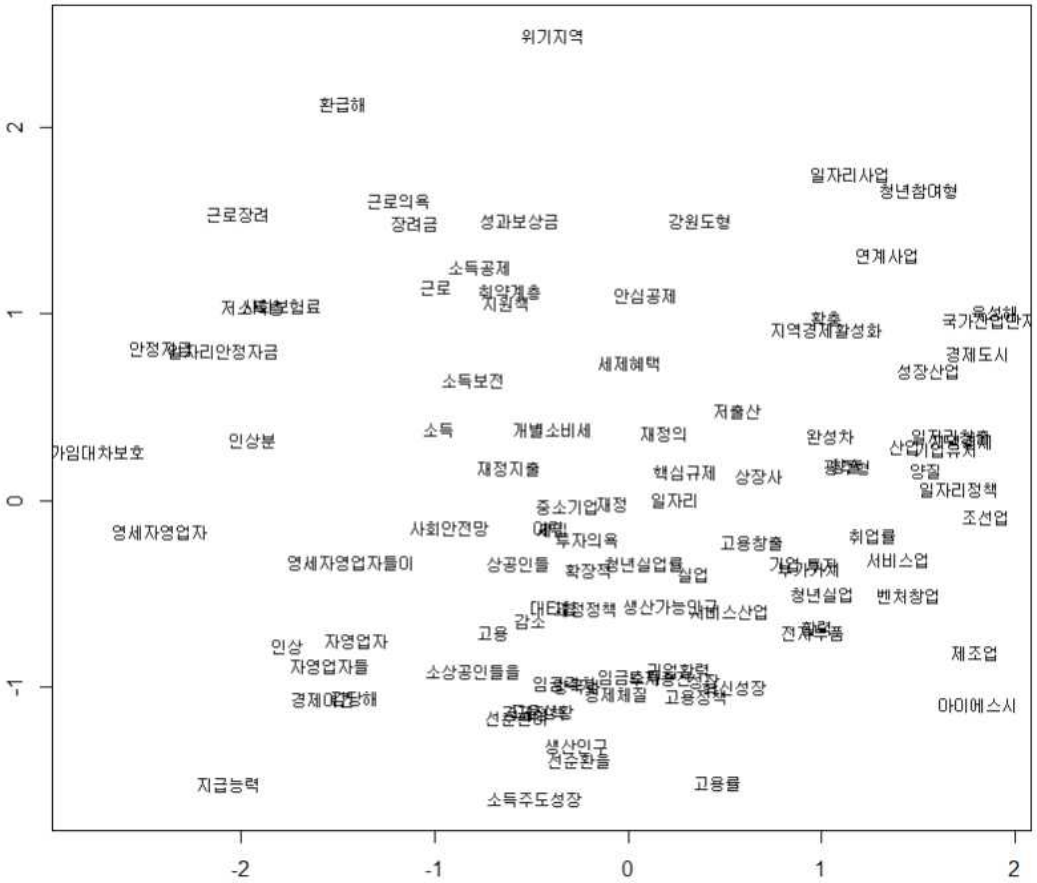
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
판단	-0.2812839	-0.24666503	-0.37673825	0.02232044	0.23798600	-0.03172816
가치	-0.2173035	-0.43035743	-0.11608051	0.02100480	-0.13927707	-0.17299138
적극	0.2577493	0.06759233	-0.08510575	0.01801304	0.22361235	-0.54705405
미날	-0.3521540	-0.19777407	0.12857533	-0.05782565	0.19361474	-0.16532086
오후	-0.5473653	-0.09973244	0.20388427	-0.38178411	0.05463321	0.04299617
반영	0.1135120	0.25482354	0.03027296	-0.08447479	0.18421446	0.17912033
문제	-0.2096169	0.30165121	-0.19521038	-0.08318360	0.27041018	-0.02944713
문건	0.2525446	0.13923781	0.05095774	0.28952730	0.51138753	0.08092804
소중한	-0.4253899	-0.13236308	0.13793221	-0.19408594	-0.25801182	-0.53936559
독자	-0.1730357	0.07234088	0.13397107	-0.47429976	0.05602616	0.05408442

[그림 3-10] Word embedding 벡터 값 예시

예를 들어, [그림 3-11]은 15번 주제 내 키워드에 해당하는 '일자리'와 근접한 단어들에 대한 벡터 공간을 시각화한 것이다. 그 결과 의미적으로 유사한 단어들끼리 근접한 벡터 공간에서 위치하는 것을 확인할 수 있었다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 3-11] ‘일자리’와 유사한 단어군

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 (4)는 벡터 A와 B의 cosine similarity를 구하는 계산식이다.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B)^2}} \tag{4}$$

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

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. [표 3-4]는 '일자리'와 연관된 단어들의 유사도이다.

[표 3-4] '일자리'와 유사한 단어들의 유사도

	일자리
일자리	1.0000000
창출	0.8342675
고용	0.7150288
안심공제	0.6577072
안정자금	0.6309876
소득	0.6069126
핵심규제	0.6028961
사회안전망	0.5933583
장려금	0.5912703

다음과 같은 방식으로 문서 내 단어 벡터들 간의 거리를 계산하였고 단어 간의 유사도를 나타내는 거리 행렬을 [표 3-5]와 같이 생성한다.

[표 3-5] Distance Matrix 예시

	일자리	창출	고용	...
일자리	1.0000000	0.8342675	0.7150288	...
창출	0.8342675	1.0000000	0.7536542	...
고용	0.7150288	0.7654796	1.0000000	...
안심공제	0.6577072	0.5809124	0.564536	...
안정자금	0.6309876	0.5321813	0.511789	...
...	...	...	...	...

Collection @ chosun

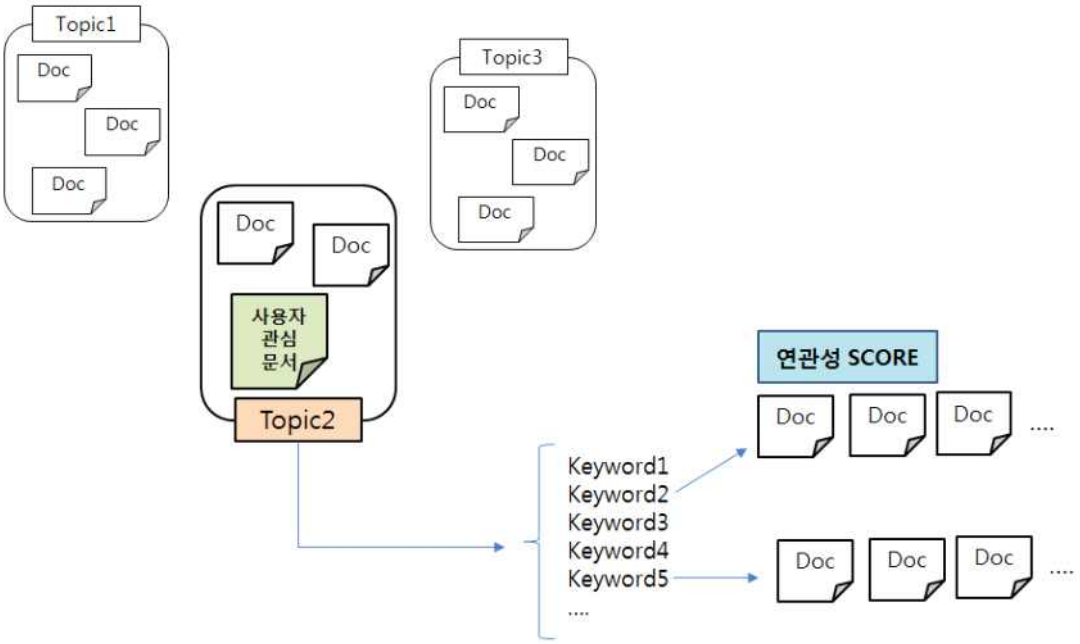
조선대학교
CHOSUN UNIVERSITY

D. 연관성 SCORE 도출

A절에서는 LDA 기반 토픽 모델링을 통해 문서 내 주제 분포와 주제 내 단어 분포를 추출했고 B절에서는 Word2vec을 사용하여 단어 간 유사도를 구하여 거리 행렬을 생성하였다.

본 절에서는 문서의 주제 내 키워드에 해당하는 부분만 추출하여 가중치 행렬을 생성하고 단어의 의미 관계를 고려해 연관성 SCORE를 도출한다.

[그림 3-12]는 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.



[그림 3-12] 연관 문서 추천 프레임워크

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-13] 은 연관성 Score를 도출하는 과정을 시각화한 것이다.

단어 w

Similarity	word1	word2	word3	···
핵심어 Keyword1	9.23	0.002	0.001	
Keyword2	0.001	8.79	0.002	
···				

X

x

DTM	word1	word2	word3	···
Doc1	5	0	1	
Doc2	1	2	3	
···				

=

wx^T

Relevance SCORE	Keyword1	Keyword2	···
Doc1	38.127	1.958	
Doc2	52.698	2.347	
···			

[그림 3-13] 연관성 SCORE 도출 과정

w는 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 x는 문서와 단어의 관계를 빈도수로 나타내는 DTM이다.

w와 x 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 (5)과 같이 나타낼 수 있다.

$$\text{Relevance SCORE} = wx^T = \sum_{i=1}^{n}\sum_{j=1}^{q} w_j a_{i,j} \tag{5}$$

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-14]는 연관성 SCORE 도출 예시를 나타낸다.

단어

Similarity	핵물질	일자리	폭염	···
핵심어 비핵화	9.23	0.002	0.001	
최저임금	0.001	8.79	0.002	
···				

X

DTM	Doc1	Doc2	Doc3	···
핵물질	5	0	1	
일자리	1	2	3	
폭염	0	3	0	
···				

=

Relevance SCORE	비핵화	최저임금	···
Doc1	38.127	1.958	
Doc2	52.698	2.347	
···			

[그림 3-14] 연관성 SCORE 도출 예시

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 키워드와 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 처리할 수 있으며 단어 간의 의미 관계를 고려한 의미 기반 문서 검색이 가능해진다.

[표 3-6]은 '비핵화' 키워드에 대한 연관성 SCORE 결과와 정규화 과정을 거친 연관

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

성 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 (6)과 같다.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{6}$$

[표 3-6] '비핵화' 키워드의 연관성 SCORE 정규화 결과 비교

문서 번호	'비핵화' 연관성 SCORE	연관성 SCORE 정규화
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
557	159.48510	0.5609640
1044	159.13191	0.5099847
595	157.13786	0.4708749
373	151.66184	0.4698088
568	150.80822	0.4637898
604	149.01069	0.4472607
561	148.59182	0.4446841
837	144.53352	0.4392583
996	142.49887	0.4379939
133	138.13977	0.4257441

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅳ. 실험 및 평가

본 장에서는 제안하는 단어 의미 관계를 고려한 연관 문서 추천 방법에 대한 실험을 위해 총 24,887개의 NAVER에서 제공하는 정치 카테고리의 한국어 뉴스 기사를 이용하며 제안하는 방법의 연관성 SCORE 결과 및 성능을 평가하고 비교 실험을 진행한다.

본 연구는 [표 4-1]과 같은 환경에서 실험을 진행하였다.

[표 4-1] 실험 환경

Division		Contents
HW	CPU	Intel(R) Core(TM) i5-6500 CPU @ 3.20GHz
	GPU	NVIDIA Geforce GTX 1060 6G
	RAM	16.00GB
	HDD	237GB
SW	OS	Windows7
	Program Language	Java, Python, R

A. 실험 결과 및 성능 평가

실험에 사용한 키워드는 '비핵화'로 선정했으며 제안하는 방법의 상위 연관성 SCORE 범위와 하위 연관성 SCORE 범위에 있는 기사들은 [표 4-2]와 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 4-2] 연관성 SCORE 범위의 기사 내용 예시

범위	기사 내용
상위 연관성 SCORE 범위	'비핵화'에 대한 직접적인 의견이 담긴 기사 - 대통령 통일외교안보특보의 비핵화 협상에 대해 주목해야할 3가지[29] - 감정적인 비핵화 비관론을 경계하는 내용[30]
하위 연관성 SCORE 범위	'비핵화'에 관련된 사건이나 인물에 대한 기사 - 트럼프 대통령의 종전선언 제안[31] - 비핵화에 좋은 영향을 줄 동창리 발사장 해체[32] - 비핵화 협상에 긍정적인 영향이 기대되는 북한의 미군 유해 55구 송환[33] - 비핵화 협상의 중재자로서 강행군하던 문 대통령의 감기 몸살로 인한 휴가[34]

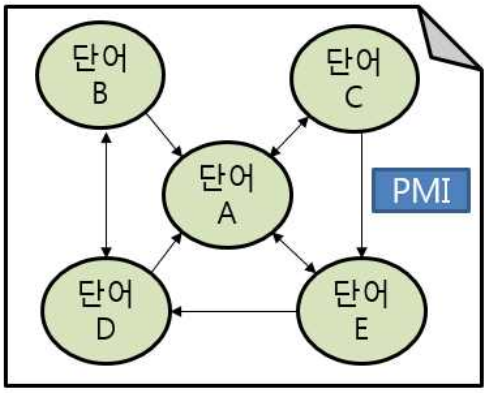
상위 연관성 SCORE 범위에 있는 기사들은 핵, 핵물질, 핵탄두, 핵실험장, 핵국가, 핵무력 등 '비핵화' 키워드와 연관성이 높은 단어들로 구성되기 때문에 주로 직접적인 의견이 담긴 기사들이 나타났다. 하위 연관성 SCORE 범위에 있는 기사들은 종전선언, 유해, 송환, 휴가 등 '비핵화'와 낮은 연관성을 갖는 단어들을 포함하기 때문에 이와 같은 결과가 도출되었다.

본 논문에서 제안하는 방법에 대한 성능을 평가하기 위해 TextRank 알고리즘을 사용하여 문서 내 단어의 중요도를 측정하였다.

[그림 4-1]과 같이 문서 내의 단어들을 정점(Node)로 선택하였고 두 단어의 연관성을 확률적으로 계산하는 PMI(Pointwise Mutual Information) 값을 간선(Edge)으로 지정했다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 4-1] TextRank 알고리즘을 이용한 문서의 핵심어 추출

PMI 값이 클수록 두 단어의 연관성이 높다는 것을 의미한다[35]. PMI의 계산식은 식 (7)과 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.

$$PMI(X,Y) = \log \frac{P(X,Y)}{P(X)P(Y)} \tag{7}$$

성능 평가 또한 두 범위에서 진행하였으며 상위 연관성 SCORE 범위의 결과는 [표 4-3]과 같고 하위 연관성 SCORE 범위의 결과는 [표 4-4]와 같다.

성능 평가 결과 '비핵화' 키워드는 상위 연관성 SCORE에서 더 높은 중요도를 가졌으며 상위 연관성 SCORE 범위의 '트럼프', '김정은'과 하위 연관성 SCORE 범위의 '종전', '선언'과 같이 각 범위에서 중요도가 높은 키워드는 서로 낮은 중요도를 가지는 것 또한 알 수 있었다. 본 논문에서 제안하는 단어 의미 관계를 고려한 연관성 SCORE에 따른 의미 기반 문서 추천의 적합성을 확인할 수 있었다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 4-3] 상위 연관성 SCORE 범위 내 단어 중요도 결과

단어	중요도
북한	0.100335
폼페이오	0.0905634
트럼프, 대통령	0.0881026
비핵화, 협상	0.0684315
미국	0.0676661
핵, 문제	0.0653499
김정은, 위원장	0.0641639
종전, 선언	0.0560652
정상회담	0.0448876
미사일	0.0290696

[표 4-4] 하위 연관성 SCORE 범위 내 단어 중요도 결과

단어	중요도
문, 대통령	0.126457
종전, 선언	0.0956396
미군, 유해	0.0562377
휴가	0.0513322
비핵화, 협상	0.0467537
정상회담	0.0428777
발사장	0.027337
송환	0.0193288
트럼프	0.018463
김정은	0.0167175

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 연관성 비교 실험

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.

[표 4-5]와 [표 4-6]은 비교 실험에 대한 결과를 나타낸다.

[표 4-5] 상위 연관성 SCORE 범위 비교 실험 결과

문서 번호	TF-IDF	LDA	제안 방법
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

[표 4-6] 하위 연관성 SCORE 범위 비교 실험 결과

문서 번호	TF-IDF	LDA	제안 방법
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

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

비교 실험 결과 TF-IDF는 연관성이 불규칙적으로 일치하지 않게 나타났다. LDA의 결과 범위는 0부터 1까지이며 제안하는 방법의 상위 연관성 SCORE 범위에서는 LDA 또한 범위 기준으로 높은 수치 값이 나타났다. 하지만 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. [표 4-7]은 본 연구의 비교실험에 사용한 방법론에 대한 설명을 나타낸다.

[표 4-7] 방법론 비교 설명

	TF-IDF	LDA	제안하는 방법
	단순 빈도수 기반	확률 기반	의미 기반
특징	- 문서 내 중요 단어를 직관적으로 해석 가능 - 새로운 단어에 대한 해석 불가능 - 단어의 의미와 문장에서의 맥락을 고려하지 못함	- 여러 주제가 혼합된 문서 처리 가능 - 휴릭스틱에 의존 - 단어의 의미와 문장에서의 맥락을 고려하지 못함	- 단어의 문맥적 정보 보존 - 단어 간의 의미 관계를 고려 - 새로운 단어에 대한 해석 가능

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

V. 결론 및 제언

본 논문에서는 단어의 의미적 모호성을 해소하고 효율적인 정보 검색을 위해 단어 의미 관계를 고려한 연관 문서 추천 방법을 제안하였다. 실험을 위한 데이터는 네이버에서 제공하는 정치 카테고리의 뉴스 기사를 Java 기반 환경에서 웹 크롤링하여 2018.07.01.부터 2018.07.31.까지 총 24,887개의 기사를 수집하였다.

LDA 기반 토픽모델링을 통해 문서 집합의 주제를 분류할 수 있고 주제 내 키워드들을 추출할 수 있다. Word2vec을 사용하여 단어를 벡터화 한 후 단어 간 유사도를 구해 거리 행렬을 생성한다. 거리 행렬에서 주제 내 키워드에 해당하는 부분만 추출해 가중치 행렬을 생성하여 단어 간의 의미 관계를 고려할 수 있고 DTM 행렬과 가중합을 통해 연관성 SCORE를 도출하였다. 연관성 SCORE의 범위를 0부터 1구간으로 일치시키기 위하여 정규화 과정을 거친 후 연관성 SCORE 수치가 높은 순서대로 문서를 추천하는 방법을 제안하였다.

가중치 행렬을 통해 사용자가 원하는 키워드와 문서집합 내 단어들의 의미적 유사도를 가중치로 부여하였고 키워드와 문서의 연관성을 SCORE로 수치화하였다. LDA는 확률 모델로 단어 간의 관계를 확률적으로 계산하지만 Word2vec을 이용해 가중치 행렬을 생성하여 단어 간의 의미 관계를 고려하여 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 SCORE에 의한 문서 추천 결과로 상위 연관성 SCORE 범위에 있는 문서일수록 질의 키워드와 연관성이 높은 단어들로 구성된 것을 확인할 수 있었다. 문서 내 단어의 중요도를 알 수 있는 TextRank 알고리즘을 사용하여 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. 또한 비교실험을 통해 기존 문서 랭킹 방법론인 TF-IDF와 LDA보다 더 효과적인 의미 기반 문서 추천이 가능하다는 것을 알 수 있었다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

참고문헌

[1] 김지연, “인터넷 검색 엔진 : 사용자의 관심을 흡수하여 전문성을 강화하는 기술”, 한국과학기술학회지, Vol.13, No.1, pp.181-216, 2013.

[2] 오지연, 박승관, “검색 엔진의 공신력과 정보의 순위가 정보의 검색과 활용에 미치는 영향”, 한국언론학회지, Vol.53, No.6, pp.26-49, 2009.

[3] 함경준, “의미 기반 문서 검색 기술 동향”, 대한기계학회지, Vol.55, No.5, pp.38-42, 2015.

[4] KPT, Korea Press Foundation, 언론수용자 의식 조사, 2014.

[5] S. Robertson, “Understanding Inverse Document Frequency: On theoretical arguments for IDF”, Journal of Documentation, Vol.60, No.5, pp503-520, 2004.

[6] http://ko.wikipedia.org/wiki/Tf-idf

[7] 유은순, 최건희, 김승훈, “TF-IDF와 소셜 텍스트의 구조를 이용한 주제어 추출 연구”, 한국컴퓨터정보학회 논문지, Vol.20, No.2, pp.121-129, 2015.

[8] 노연우, 임종태, 복경수, 유재수, “소셜 네트워크 환경에서 변형된 TF-IDF를 이용한 핫 토픽 예측 기법”, 정보과학회 컴퓨팅의 실제 논문지, Vol.23, No.4, pp.217-225, 2017.

[9] 박대서, 김화종, “TF-IDF 기반 키워드 추출에서의 의미적 요소 반영을 위한 결합 벡터 제안”, 한국정보기술학회 논문지, Vol.16, No.2, pp.1-16. 2018.

[10] 김민범, “유의어 가중치를 적용한 논문 표절 유사도 측정 방법”, 조선대학교, 석사학위논문, 2018.

[11] C.H. Papadimitriou, P. Raghavan, H. Tamaki and S. Vempala, “Latent Semantic Indexing: A Probabilistic Analysis”, Journal of Computer and System Sciences, Vol.61, No.2, pp.217-235, 2000.

[12] M. Rosen-Zvi, T. Griffiths, M. Steyvers and P. Smyth, “The Author-Topic Model for Authors and Documents”, Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp.487-794, 2004.

[13] D. Mimno, A. McCallum, “Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression”, Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, pp.411-418, 2008.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[14] D.M. Bleil, A.Y. Ng and M.I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol.3. pp.993-1022. 2003.

[15] 심준식, 김형중, “LDA 토픽 모델링을 활용한 판례 검색 및 분류 방법”, 전자공학회 논문지, Vol.54, No.9, pp.67-75, 2017.

[16] 곽창욱, 김선중, 박성배, 김권양, “무한 사전 온라인 LDA 토픽 모델에서 의미적 연관성을 사용한 토픽 확장”, 정보과학회 컴퓨팅의 실제 논문지, Vol.22, No.9, pp.461-466, 2016.

[17] 조태민, 이지형, “LDA 모델을 이용한 잠재 키워드 추출”, 한국지능시스템학회 논문지, Vol.23, No.2, pp.180-185, 2015.

[18] 김성호, 이성일, 윤한준, 이도훈, 이수철, “토픽 모델링을 활용한 악성 인터넷 트래픽 시그니처 패턴 자동추출 기법”, 정보과학회지, Vol.33, No.6, pp.61-65, 2015.

[19] 강범일, 송민, 조화순, “토픽 모델링을 이용한 신문 자료의 오피니언 마이닝에 대한 연구”, 한국문헌정보학회지, Vol.47, No.4, pp.315-334, 2013.

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

[21] 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.

[22] 이앞길, 최근호, 김건우, “LDA 토픽 모델링을 활용한 유사 특허문서 추천 모델”, 한국지능정보시스템학회, pp.92-104, 2018.

[23] L. Page, S. Brin, R. Motwani, and T. Winograd, “ThePageRank Citation Ranking: Bringing Order to the Web”, Stanford Digital Libraries Working Paper, 1998.

[24] S. Brin, 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.

[25] 김용기, 이상연, 이건명, “페이지간 유사도를 이용한 가중치 PageRank 알고리즘”, 한국정보과학회 2014 한국컴퓨터종합학술대회 논문집, pp.891-893, 2014.

[26] R. Mihalcea and P. Tarau, “TextRank: Brigning Order into Texts”, Proceeding of EMNLP-04 and the 2004 Conference on Empirical Methods in

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Natural Language Processing, 2004.

[27] 안인석, 김현우, 김형주, “텍스트 랭크 알고리즘을 이용한 사용자 타임라인 요약 기법”, 정보과학회논문지, Vol.39, No.4, pp.238-245, 2012.

[28] 이원형, 조성일, 김동회, “트위터에서 형태소 분석과 PageRank 기반 화제 단어 추출 방법 제안”, 한국디지털콘텐츠학회 논문지, Vol.19, No.1, pp.157-163, 2018.

[29] “[김성준의시사전망대] 문정인 ‘북미 비핵화 협상, 3가지를 주목하라’”, SBS 뉴스, https://news.sbs.co.kr/news/endPage.do?news_id=N1004839712&plink=ORI&cooper=NAVER

[30] “[김영희 대기자의 한반도 워치] 감정적인 비핵화 비관론을 경계한다”, 중앙일보, https://news.joins.com/article/22821502

[31] “트럼프 ‘南北美 3자 종전선언’ 제안… 김정은 ‘中포함’ 맞서 무산”, 동아닷컴, http://news.donga.com/3/all/20180731/91312834/1

[32] “청와대 ‘동창리 발사장 해체 동향, 비핵화에 좋은 영향’”, 연합뉴스TV, http://www.yonhapnewstv.co.kr/MYH20180724010300038/?did=1825m

[33] “북한, 미군 유해 55구 송환... 트럼프 ‘김정은 고맙다’”, 한국일보, http://www.hankookilbo.com/News/Read/201807271759791736

[34] “문재인 휴가 구상 키워드... ‘비핵화, 국방·권력기관 개혁, 경제’”, MBN 뉴스, http://mbn.mk.co.kr/pages/news/newsView.php?category=mbn00006&news_seq_no=3595165

[35] 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.

Collection @ chosun