http://www.jsebs.org
pISSN 2288-3908
eISSN 2765-3846

The Journal of Society for e-Business Studies
Vol.26, No.3, August 2021, pp.119-132
https://doi.org/10.7838/jsebs.2021.26.3.119

지식 간 내용적 연관성 파악 기법의 지식 서비스 관리 접목을 위한 정량적/정성적 고려사항 검토

Quantitative and Qualitative Considerations to Apply Methods for Identifying Content Relevance between Knowledge Into Managing Knowledge Service

유기동(Keedong Yoo)*

초 록

내용적 연관성에 기반한 연관지식의 파악은 핵심 지식에 대한 서비스와 보안의 기본적인 기능이다. 본 연구는 내용적 연관성을 기준으로 연관지식을 파악하는 기존의 방식, 즉 키워드 기반 방식과 워드임베딩 방식의 연관문서 네트워크 구성 성능을 비교하여 어떤 방식이 정량적/정성적 측면에서 우월한 성능을 나타내는가를 검토한다. 검토 결과 키워드 기반 방식은 핵심 문서 파악 능력과 시맨틱 정보 표현 능력 면에서 우월한 성능을, 워드임베딩 방식은 F1-Score와 Accuracy, 연관성 강도 표현 능력, 대량 문서 처리 능력 면에서 우월한 성능을 나타냈다. 본 연구의 결과는 기업과 사용자의 요구를 반영하여 보다 현실적인 연관지식 서비스 관리에 활용될 수 있다.

ABSTRACT

Identification of associated knowledge based on content relevance is a fundamental functionality in managing service and security of core knowledge. This study compares the performance of methods to identify associated knowledge based on content relevance, i.e., the associated document network composition performance of keyword-based and word-embedding approach, to examine which method exhibits superior performance in terms of quantitative and qualitative perspectives. As a result, the keyword-based approach showed superior performance in core document identification and semantic information representation, while the word embedding approach showed superior performance in F1-Score and Accuracy, association intensity representation, and large-volume document processing. This study can be utilized for more realistic associated knowledge service management, reflecting the needs of companies and users.

키워드 : 내용적 연관성, 키워드 기반 기법, 워드임베딩 기법, 연관지식 서비스, 핵심 지식 보안
Content Relevance, Keyword-Based Approach, Word Embedding Approach, Associated Knowledge Service, Core Knowledge Security

* Professor, Department of Business Administration, Dankook University(kdyoo@dankook.ac.kr)
Received: 2021-07-26, Review completed: 2021-08-17, Accepted: 2021-08-20

1. 서 론

문제해결을 위해 사용자는 내용적인 면에서 관련된 하나 이상의 지식을 순차적으로 사용한다. 즉, 사용자는 연상과정(Process of association)에 의해 내용적 연관성(Content relevance)을 갖는 다른 지식을 하나씩 조회 및 선택하는 과정을 반복한다. 이는 사용할 지식을 결정하는 사용자의 논리적 패턴, 즉 필요 지식에 대한 사용자의 인지과정(Cognitive process)으로 볼 수 있다[24].

지식은 문서를 통해 표출(Externalize)되므로 지식이 수록된 문서는 지식 자체로 간주될 수 있다. 내용적 연관성을 갖는 하나 이상의 지식을 순차적으로 선택 및 적용하는 사용자의 논리적 과정은, 내용적인 면에서 관련된 문서를 검색 및 추출하는 물리적 과정으로 표현된다. 문제의 해결을 위해 지식 사용자가 탐색 및 추출하는 실질적인 대상은 연관지식(Associated knowledge)을 담고 있는 연관문서(Associated document)이므로, 내용적 연관성에 기반한 연관문서를 파악 및 제공하는 기능은 차세대형 문서관리시스템의 기본적인 기능성 중 하나라고 할 수 있다.

내용적 연관성을 기준으로 연관문서를 파악하기 위한 기존의 방법은 크게 두 가지로 볼 수 있는데, 즉 공통 키워드를 갖는 문서들을 해당 키워드를 기준으로 연결(Linking) 및 군집화(Grouping)하는 '키워드 기반(Keyword-based)' 기법과 문서 내에 포함된 단어를 벡터 공간에 투영하여 문서 벡터 간 의미적 거리(Semantic distance)를 파악하는 '워드임베딩(Word embedding)' 기법이 있다. 키워드 기반 기법은 문서 내에 사용된 단어(Word) 또는 문구(Phrase) 형식으로 구성되는 키워드를 이용하므로 문서의 내용을 직접적이고 정확하게 대변할 수 있다는 장점이 있으나, 자동 추출된 키워드의 부정확성과 동의어사전(Thesaurus) 추가 구성 등의 부담이 있다. 반면 워드임베딩 기법은 동의어사전 등의 추가적인 조치가 없을지라도 정확성이 높은 결과를 산출할 수 있다는 장점이 있으나, 범용적 코퍼스(Corpus)를 이용한 견고한(Robust) 벡터공간 구성에 따른 연산의 부담이 있다. 대용량의 문서 간 내용적 연관성 기반 연관문서를 파악하는 데에는 워드임베딩 기법이 유리한 반면, 문서 간 연관성을 매개(Intermediate)하는 정보를 명시하는 데에는 키워드 기반 기법이 유리하다[18, 20].

이들 기법은 목적과 분야에 따라 선별적으로 다양하게 활용되고 있으나, 실제 활용 중인 지식관리시스템과 같은 실제 정보시스템에 접목되어 내용적으로 관련된 연관문서를 추출 및 제공하는 기능을 구현한 예는 찾아보기 어렵다. 이는 기존의 연구는 주로 내용적 연관성 파악 기법 및 알고리듬의 성능 고도화 및 검증에 관심을 두고 있기 때문으로 볼 수 있다. 즉, 상용 정보시스템에 연관문서 파악 및 추천 기능을 탑재하기 위해서는 산출된 문서 간 내용적 연관성 및 연관문서 네트워크가 기업과 사용자의 현실적 요구사항에 얼마나 부응하는가에 대한 분석과 검토가 필요하다. 각 기법의 핵심 문서 파악 능력, 연관성 강도 표현 능력, 시맨틱 정보 표현 능력, 대량 문서 처리 능력 등의 정성적 관점에서 적용 가능성에 대한 추가적인 고려가 필요하다.

따라서 본 연구는, 문서 간 내용적 연관성 파악을 위한 키워드 기반 및 워드임베딩 기법의 성능을 정량적 및 정성적 관점에서 비교한다.

즉, 연관문서 네트워크를 구성하는 각 기법의 성능을 F1-Score 및 Accuracy, 핵심 문서 파악 능력, 연관성 강도 표현 능력, 시맨틱 정보 표현 능력, 대량 문서 처리 능력 등의 측면에서 비교하여, 어느 기법이 이들 관점에서 연관문서 서비스와 보안 관리에 유리한 기능성을 나타내는가를 검토하고 이유를 분석한다.

2. 관련 연구

2.1 키워드 기반 기법을 이용한 연관문서 파악 연구

키워드 기반 기법은 동일 또는 유사 키워드를 공통으로 포함하는 문서들은 내용적인 면에서 유사하거나 관련되어 있다는 가정 하에 연관문서를 파악하는 방법으로[22, 27], 키워드 자동 추출을 위한 텍스트마이닝 기법이 제시된 이후 광범위하게 활용되고 있다[8, 9].

키워드를 기반으로 연관문서를 파악하는 기존의 연구는, 동일 또는 유사 키워드를 갖는 문서들을 특정 특성을 갖는 하나의 군집(Cluster)으로 정의하여 또 다른 특성을 갖는 문서 군집과 구분하거나 연관시키는 방식으로 진행된다. 즉, 문서로부터 추출된 키워드를 기준으로 문서들을 군집화 하여 해당 키워드의 주제에 대한 정보와 지식의 밀집도 및 관심 정도를 파악하거나[4], 해당 키워드와 함께 출현하는 키워드를 연결하여 특정 주제의 기술과 관련된 기술의 현황을 파악하거나[3], 해당 키워드가 속한 특정 분야에서 거론되는 이슈 등을 파악한다[26]. 또한 키워드를 기준으로 유사한 내용 또는 내용적 영향관계를 갖는 연관 웹페이지를 파악하는 연구도 진행되었다. 즉, 웹페이지 간 상호 영향관계를 대변하는 하이퍼링크(Hyperlink)는 링크를 통해 연결된 웹페이지의 내용적 특성을 설명하지 못하므로[1, 13], 웹페이지에 수록된 내용을 대변하는 키워드를 바탕으로 연관 웹페이지를 파악한다[12, 27].

키워드 기반 기법을 이용한 연관문서의 파악은 문서의 내용을 직접적으로 대변하는 키워드를 이용하므로 문서의 내용을 정확하고 합리적으로 반영할 수 있다는 장점이 있으나 어휘적으로 정확히 일치하는 키워드를 갖는 문서만을 연관문서로 정의하는 한계가 있다. 이를 보정하기 위하여 동의어(Synonym) 또는 관련어(Hypernym, Hyponym) 등을 함께 고려할 수 있도록 추가적인 조치가 반드시 필요한데, 즉 동의어사전(Thesaurus)을 통해 동의어를 고려하거나 Wordnet과 같은 도구를 이용한 관련어 처리를 위한 조치가 필요하다. 그러나 이들 도구를 적용하는 경우일지라도 문서에 포함된 변화 및 신생하는 다양한 단어와 용어에 대한 동의어와 관련어를 모두 고려할 수 있는 것은 아니므로, 매뉴얼(Manual) 방식으로 이를 추가 및 보완하는 작업이 반드시 병행되어야 하는 부담이 있다.

2.2 워드임베딩 기법을 이용한 연관문서 파악 연구

워드임베딩 기법은 대용량 코퍼스를 이용하여 코퍼스 내에 포함된 단어의 벡터값을 기반으로 기준 벡터공간, 즉 학습모델을 형성하고, 새롭게 출현하는 특정 단어 또는 문서의 벡터값을 산출하여 해당 단어 및 문서의 의미적(Semantic) 위치를 벡터공간에서 결정하는 방

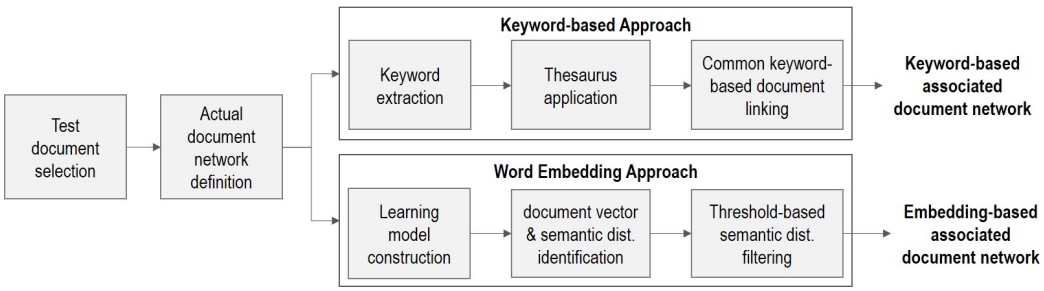
법이다. 각 문서가 갖는 고유의 벡터값을 기준으로 문서 간 의미적 거리를 산출하고 특정 임계값(Threshold) 이상의 의미적 거리값을 갖는 문서들은 연관문서로 판정한다.

키워드 기반 기법에 비해 상대적으로 최근 개발된 워드임베딩 기법은, 연산 장비와 알고리듬의 진보로 인해 대용량 정보를 비교적 빠르게 처리할 수 있는 장점을 이용하여 정보와 문서 간 관계를 파악하거나 요약하는 등의 연구에 활발히 응용되고 있다. 특히 언어모델의 구성을 통해 컴퓨터가 사람의 언어를 모방 또는 이해하도록 하는 자연어처리 분야의 핵심적인 기술로 평가되고 있으며[14, 25], 정확하게 사람의 언어를 이해하고 처리할 수 있도록 다양한 알고리듬이 제시되고 있다. Le and Mikolov[19]는 문맥상 유사한 의미를 가진 단어들이 벡터 공간 내에서 가깝게 위치하게 되는 인공 신경망(Artificial neural network) 기반 언어 모델인 Word2Vec을 확장하여 문장(Sentence), 문단(Paragraph), 문서 등을 처리할 수 있는 Doc2Vec을 제안하였다. Dai et al.[5]은 Doc2Vec이 문서 군집화 및 분류에 있어 기존 토픽 모델인 PLSA(Probabilistic latent semantic analysis) 또는 LDA(Latent dirichlet allocation) 등에 비해 우수한 성능을 나타냄을 증명하였다. Pennington et al.[21]은 Word2Vec과 LSA의 한계를 극복하여 임베딩된 단어 벡터 간 유사도 측정을 용이하게 함과 동시에 코퍼스의 통계 정보를 반영할 수 있는 GloVe 모델을 제안하였다. Bojanowski et al.[2]은 FAIR Lab에서 발표한 텍스트 분류 및 단어 임베딩 라이브러리를 이용하여, 단어를 어근 및 어미 등의 부분단어(Subword)의 벡터로 표현하여 노이즈가 많은 코퍼스의 처리가 가능하도록 FastText를 개발하였다. Kamkarhaghighi and Makrehchi[15]는 Word2Vec 또는 GloVe모델을 기반으로 각 문서의 콘텐트 트리를 구성한 후 업데이트된 단어 벡터의 평균값을 이용하여 문서를 표현하는 Content tree word embedding 모델을 제안하여 단어의 의미적 모호성 완화를 시도하였다. Devlin et al.[7]은 Transformer 인코더를 사용하여 단어의 문맥(Context)상 의미를 state-of-the-art 수준으로 파악하는 BERT (Bidirectional encoder representations from Transformers)를 제시하였다.

그러나 워드임베딩 기법은 문서에 포함된 단어의 벡터값을 기준으로 문서의 벡터값을 결정하므로, 문서의 내용을 대변하는 데에 역할이 상대적으로 낮은 단어도 문서 벡터값 산출에 영향을 주어 결과적인 문서 벡터값의 정확도를 저하시키는 문제가 있다. 이를 위해 문서의 벡터값을 직접 산출하는 Doc2Vec 알고리듬이 개발되었으나 이는 새로운 문서가 출현할 때마다 학습모델을 재구성하는 연산의 부담이 있다. 또한 BERT는 학습모델 재구성의 부담은 없으나, 입력 단어(Token)의 개수가 512개로 제한되므로 많은 수의 단어가 포함된 일반 문서에 적용하기 위하여 별도의 조치가 필요하다.

3. 키워드 기반 및 워드임베딩 기법을 적용한 연관문서 네트워크

<Figure 1>은 키워드 기반 기법과 워드임베딩 기법을 적용하여 문서 간 내용적 연관성을 파악하고 이를 통해 산출된 연관문서 네트워크를 구성하는 절차를 보여준다.



〈Figure 1〉 Procedure to compose associated document network

연관문서 네트워크는 경영정보학 분야 중 지식경영 및 지능형컴퓨팅 주제의 영문 논문을 10편을 대상으로 구성한다. 이들 논문은 구체적으로는 각기 다른 주제를 다루지만, 연구절차, 데이터 분석 및 검증, 프로토타입 구현 등 방법론 측면에서는 상호 관련된 내용을 갖는다. 즉, 제목 및 키워드(저자 지정) 등으로는 파악되지 않는 내용적인 측면에서의 연관성을 갖는 논문으로 선정하였다. <Table 1>은 선정된 논문의 세부 주제 분류 현황이다.

〈Table 1〉 Test Document Classification

Subject		Document ID
Knowledge management	Knowledge map	doc#1, doc#2, doc#3, doc#4
	Knowledge acquisition	doc#5, doc#6
	Knowledge service	doc#7, doc#8
Intelligent computing	Ubiquitous computing	doc#9, doc#10

기존의 연구는 수십만 건의 문서를 대상으로 각 기법을 적용하고 해당 기법의 정량적 성능을 타진하는 것이 일반적이다. 그러나 본 연구는, 각 기법의 정량적 성능을 비교하는 것보다는, 각 기법을 통해 파악된 문서 간 내용적 연관성과 이를 통해 구성되는 연관문서 네트워크의 형태와 의미를 비교하여 각 기법의 행태적(Behavioral) 성능을 분석한다. 따라서 각 기법을 통해 산출된 문서 간 연결과 종합적인 문서 네트워크의 의미와 유효성 여부를 직관적으로 도해 및 설명할 수 있도록 대상 문서의 수를 10편으로 한정하였다.

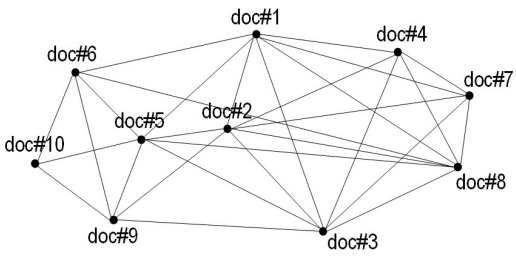
3.1 키워드 기반 연관문서 네트워크

각 문서의 키워드를 추출하고 동의어를 고려하여 공통 키워드를 갖는 문서 간 링크를 형성하여 연관문서 네트워크를 구성한다.

키워드는 자동화된 방식으로 추출하되, 하나 이상의 복합 단어, 즉 구(Phrase) 형식의 키워드 추출에 우수한 성능을 발휘하는 것으로 알려진 TerMine[10]을 적용하여 문서별로 단어 및 구 형식의 키워드를 추출한다.

동의어와 관련어의 고려는 일반적으로 상용화된 도구를 이용하여 자동화된 방식으로 진행되나, 현실적이고 정확한 동의어의 고려를 위하여, 본 연구에서는 분야 전문가의 직접적인 정의 및 검토를 통해 진행한다. 이는 추출된 문서별 키워드에 의해 문서 간 내용적 연관성이 정의되므로, 변화가 빠르고 새롭게 출현하는

	doc #1	doc #2	doc #3	doc #4	doc #5	doc #6	doc #7	doc #8	doc #9	doc #10
doc #1		o	o	o	o	o	o	o		
doc #2	o		o	o	o		o	o	o	
doc #3	o	o		o	o		o	o	o	
doc #4	o	o	o				o	o		
doc #5	o	o	o			o		o	o	o
doc #6	o				o			o	o	o
doc #7	o	o	o	o				o		
doc #8	o	o	o	o	o	o	o			
doc #9		o	o		o	o				o
doc #10					o	o			o	



〈Figure 2〉 Keyword-based Associated Document Network

다양한 개념 및 용어들을 최대한 반영하여 문서 간 내용적 연관성을 파악의 정확성을 극대화하기 위한 조치이다.

동의어 및 관련어를 적용하여 공통 키워드를 포함하는 문서를 연관문서로 정의한다. 또한 연관문서 간의 링크를 종합하여 대상 문서에 대한 연관문서 네트워크를 구성한다. 대상 문서의 수가 많은 경우 '문서-키워드' 형식의 이원(2-mode) 네트워크를 '문서-문서' 형식의 일원(1-mode) 네트워크로 전환하여 연관문서 네트워크를 도출하는 데에 자동화된 도구를 사용할 수 있다[24]. 키워드 기반 기법을 적용하여 파악된 연관문서 네트워크는 <Figure 2>와 같다.

3.2 워드임베딩 기반 연관문서 네트워크

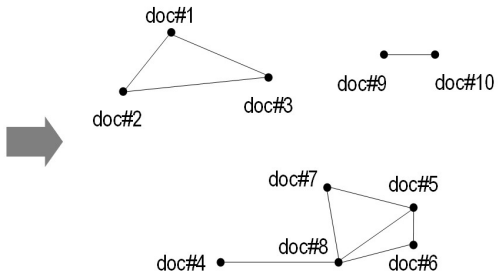
본 연구에서는 워드임베딩 기법의 기본적인 알고리듬인 Word2Vec(Skip gram 방식)을 적용한다. 이는 워드임베딩 기법에 속하는 다양한 알고리듬이 Word2Vec을 근간으로 수정 및 보완 작업을 거듭하며 개발되었으므로, Word2Vec이 워드임베딩 기법의 원리를 가장 충실하게 준수한다고 볼 수 있기 때문이다.

본 연구에서 사용한 코퍼스는 약 3백만 단어를 포함하는 구글뉴스 덤프파일(2016 버전)로, 이에 대해 대소문자 구분 제거, 어근 추출, 불용어 제거 등의 전처리 과정을 거쳐 학습모델을 구성하였다.

Word2Vec은 문서 내 단어의 벡터값을 산출하므로 Word2Vec을 통해 문서의 의미상 위치를 의미하는 문서 벡터값으로 바로 사용할 수 없다. 따라서 본 연구에서는 문서 내에 포함된 단어들의 벡터값을 산술평균하여 해당 문서의 벡터값으로 간주한다[6, 16]. 이는 단어 벡터값의 산술평균을 이용하여 해당 단어가 포함된 문장의 벡터값을 결정하는 방법을 확장한 방식이다.

문서별 벡터값을 기준으로 각 문서의 의미적 위치가 결정되면 이들 문서 간의 의미적 거리를 코사인 유사도를 이용하여 산출한다. 코사인 유사도를 이용하여 문서 간 의미적 거리를 파악하는 것은 일반적인 방법으로, 다차원 벡터공간 상에 위치한 특정 벡터의 의미적 유사성을 파악하는 데에 광범위하게 적용된

	doc #1	doc #2	doc #3	doc #4	doc #5	doc #6	doc #7	doc #8	doc #9	doc #10
doc #1		o	o							
doc #2	o		o							
doc #3	o	o								
doc #4								o		
doc #5						o	o	o		
doc #6					o			o		
doc #7					o			o		
doc #8				o	o	o	o			
doc #9										o
doc #10									o	



〈Figure 3〉 Word Embedding-based Associated Document Network

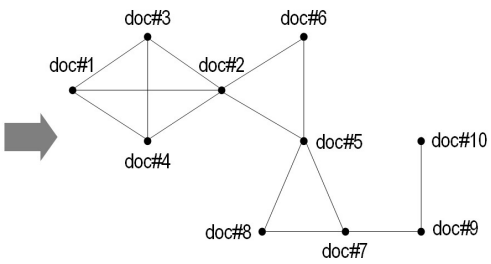
다. 코사인 유사도의 계산은 벡터공간 상에 위치하는 모든 벡터에 대해 행해지므로, 이들 중 실제 연관성을 갖는 벡터를 선별하기 위한 임계치(Threshold)의 설정이 필요하다. 정형화된 임계치 산출 방법이나 값에 대한 기준은 존재하지 않으므로, 본 연구에서는 예측 성능 지표인 F1-Score 값을 최대화하는 최적 임계치를 선형계획(Linear Programming)에 의해 산출하여 적용한다. Word2Vec을 적용하여 파악된 연관문서 네트워크는 <Figure 3>와 같다.

4. 각 기법의 연관문서 네트워크 구성 성능 비교

4.1 정량적 성능

연관문서 네트워크 구성의 정량적 성능 판단은 F1-Score와 Accuracy를 이용하며, 이를 산출하기 위한 연관문서 네트워크의 정답안을 대상 문서의 내용을 정확히 파악 및 이해한 분야 전문가 5인의 검토 및 합의를 통해 구성하였다. 대상 문서 주제 분류(<Table 1>)와 연관문서

	doc #1	doc #2	doc #3	doc #4	doc #5	doc #6	doc #7	doc #8	doc #9	doc #10
doc #1		o	o	o						
doc #2	o		o	o	o	o				
doc #3	o	o		o						
doc #4	o	o	o							
doc #5		o				o	o	o		
doc #6		o			o					
doc #7					o			o	o	
doc #8					o		o			
doc #9							o			o
doc #10									o	



〈Figure 4〉 Actual Answer of the Associated Document Network

네트워크 정답안(<Figure 4>)을 비교하면 각 문서의 주제 분류와는 다르게 문서 간 내용적 연관성이 파악되었음을 알 수 있다. 이는 문서 분류 시 일반적으로 고려하는 주제만으로는 문서의 실제 내용을 대변할 수 없음을 의미한다. 즉, 주제 분류를 기준으로는 관련되지 않는 것으로 판단된 문서일지라도, 내용적인 면에 있어서는 연관되어 상호 검색 및 참조하는 현상이 실제 문제해결 과정에서는 발생할 수 있다. <Table 2>와 <Table 3>은 키워드 기반 기법과 워드임베딩 기법을 통해 구성한 연관문서 네트워크를 정답안과 비교하여 산출한 성능 결과이다.

〈Table 2〉 Performance by Keyword-based Approach

	Actual Positive	Actual Negative
Predicted Positive	22	36
Predicted Negative	6	26

Precision : 37.9 %
Recall : 78.6 %
F1-Score : **51.2 %**
Accuracy : **53.3 %**

〈Table 3〉 Performance by Word Embedding Approach

	Actual Positive	Actual Negative
Predicted Positive	16	4
Predicted Negative	12	58

Precision : 80.0 %
Recall : 57.1 %
F1-Score : **66.7 %**
Accuracy : **82.2 %**

F1-Score와 Accuracy 모두 워드임베딩 기법이 우월한 성능을 보인다. F1-Score 값의 크기를 고려할 경우 본 연구의 Word2Vec 성능이 최근 출현 중인 다른 알고리듬의 성능(최고 약 89%)보다 상대적으로 떨어짐에도 불구하고 이러한 결과를 나타내는 원인은, 키워드의 유사 정도로 문서 간 내용적 연관성을 표현하는 방식 자체가 갖는 한계 때문으로 판단된다. 즉, 동의어와 관련어의 정확성과 다양성을 극대화 할지라도, 동일하거나 유사한 키워드를 갖는 문서 간에 내용적 연관성이 있다고 결론 내리기 어려움을 의미한다.

이는 '관련성이 없는 것을 관련성이 있는 것(Actual Negative - Predicted Positive)'으로 예측하는 Type I Error와 '관련성이 있는 것을 관련성이 없는 것(Actual Positive - Predicted Negative)'으로 예측하는 Type II Error의 값을 통해서도 확인된다. 즉, Type I Error는 워드임베딩 기법이 '36 vs. 4'로 현저히 적은 반면, Type II Error는 '6 vs. 12'로 근소한 차이를 나타내므로, 키워드 기반 방식에서 오류로 인한 성능 저하가 더욱 크게 나타난다. 즉, 문서 간 공통 키워드가 많아도 내용적 연관성이 낮거나, 공통 키워드가 적어도 내용적 연관성이 높은 경우가 키워드 기반 기법에서 상대적으로 빈번히 발생할 수 있음을 의미한다.

4.2 정성적 성능

4.2.1 핵심 문서 파악 능력

네트워크 이론에 따르면 핵심 노드는 허브(Hub) 형태로 관찰되는데, 다른 노드와 연관성이 높아 많은 수의 링크가 형성되는 노드는 파티(Party) 허브로 불린다[11]. 네트워크 이론의

허브 개념을 연관문서 네트워크에 적용하면 핵심 문서(지식)를 판단할 수 있는데, 핵심 문서는 사용 및 참고 가능성과 빈도가 높아 전략적 차원에서 차별적 관리가 필요하다[24]. 즉, 문서관리시스템에서는 핵심 문서로 판단된 문서에 대해 보안 등급, 접근 권한, 보존 연한 등을 차별적으로 설정하는 조치가 필요하다. 이러한 허브에 해당하는 핵심 문서를 연관문서 네트워크에서 파악하기 위하여 네트워크 이론의 중심성(Centrality) 개념을 적용하면, 파티 허브에 해당하는 핵심 문서는 Degree centrality를 산출하여 판단할 수 있다.

<Table 4>는 정답안과 키워드 및 워드임베딩 기반 연관문서 네트워크에 대한 각 문서의 Degree centrality를 보여준다. 또한 각 기법을 통한 연관문서 네트워크의 문서별 Degree centrality가 정답안의 문서별 Degree centrality에 얼마나 부합하는가를 확인하는 상관관계 분석 결과를 보여준다. 정규분포를 따르는 데이터에 적용이 가능한 Pearson 상관계수는, 데이터의 수가 10개로 정규분포를 따르는지 확인할 수 없는 본 연구에는 적용할 수 없다. 따라서, 데이터의 정규분포 준수 여부와 관계없이, 데이터의 상승/하락 경향 간 상관성 확인이 가능한 Spearman 순위상관계수(RCC, Rank correlation coefficient)를 적용하여 상관관계를 확인하였다.

산출된 Spearman 순위상관계수 값을 순위상관계수표에 비교하면 임계값이 '0.564(n=10, 유의수준 = 0.05)'이므로, 이보다 높은 값을 갖는 '정답안-키워드 기반'의 Degree centrality가 유의미한 상관관계를 가짐을 알 수 있다. 즉, 키워드 기반 기법을 통한 연관문서 네트워크가 핵심 문서를 파악하는 데에는 우월한 성능을 나타냄을 의미한다.

〈Table 4〉 Degree centrality & correlation

	Actual	Keyword	Word embedding
doc#1	0.3333	0.7778	0.2222
doc#2	0.5556	0.7778	0.2222
doc#3	0.3333	0.7778	0.2222
doc#4	0.3333	0.5556	0.1111
doc#5	0.4444	0.7778	0.3333
doc#6	0.2222	0.5556	0.2222
doc#7	0.3333	0.5556	0.2222
doc#8	0.2222	0.7778	0.4444
doc#9	0.2222	0.5556	0.1111
doc#10	0.1111	0.3333	0.1111

Spearman RCC(n = 10, a = 0.05)

Actual-Keyword : **0.6578**

Actual-Word embedding : 0.4068

4.2.2 연관성 강도 표현 능력

키워드 기반 기법은 문서 간 공유하는 공통 키워드의 개수를, 워드임베딩 기법은 문서 간 의미적 거리를 이용하여 문서 간 내용적 연관성의 강도(Intensity)를 표현할 수 있다. 따라서 어느 기법이 더욱 효과적으로 연관성의 강도를 표현할 수 있는가를 판단하는 데에 노드 간의 거리를 기준으로 각 노드의 Centrality를 산출하는 Closeness centrality를 활용할 수 있다. Closeness centrality는 노드 간의 직간접적 거리를 이용하여 각 노드의 네트워크 상의 위치를 판단하므로, 이 값이 높은 문서는 연관문서 네트워크의 중앙에 위치하는 문서라 할 수 있다[17].

<Table 5>는 정답안과 키워드 및 워드임베딩 기반 연관문서 네트워크에 대한 각 문서의

Closeness centrality와, 각 기법을 통한 연관문서 네트워크의 문서별 Closeness centrality가 정답안의 문서별 Closeness centrality에 얼마나 부합하는가를 확인하는 상관관계 분석 결과를 보여준다.

〈Table 5〉 Closeness centrality & correlation

	Actual	Keyword	Word embedding
doc#1	0.4091	0.8182	0.2222
doc#2	0.5625	0.8182	0.2222
doc#3	0.4091	0.8182	0.2222
doc#4	0.4091	0.6429	0.2540
doc#5	0.6000	0.8182	0.3556
doc#6	0.4737	0.6923	0.2963
doc#7	0.5000	0.6429	0.2963
doc#8	0.4500	0.8182	0.4444
doc#9	0.3750	0.6923	0.1111
doc#10	0.2813	0.5294	0.1111

Spearman RCC(n = 10, a = 0.05)

Actual-Keyword : 0.4935

Actual-Word_embedding : **0.7492**

산출된 Spearman 순위상관계수 값을 순위상관계수표에 Degree centrality의 경우와 동일하게 비교하면, '정답안-워드임베딩'의 Closeness centrality가 유의미한 상관관계를 가짐을 알 수 있다. 즉, 워드임베딩 기법을 통한 연관문서 네트워크가 문서 간 연관성 강도를 표현하는 데에 더욱 효과적인 성능을 나타냄을 의미한다.

4.2.3 시맨틱 정보 표현 능력

문서 간 내용적 연관성 정보, 즉 시맨틱 정보는 문서 간 링크 형성의 근거를 설명한다. 따라서 문서 간 링크를 통해 연관문서를 연쇄적이고 순차적으로 검색 및 선택하는 상호참조적 네비게이션을 진행하는 사용자의 작업에 필수적인 정보이다[23].

이러한 시맨틱 정보를 문서 간 링크에 표현하는 데에는 키워드 기반 기법이 유리하다. 키워드 기반 기법은 연관문서 간 링크를 해당 문서들이 공통으로 갖는 키워드를 기준으로 구성하므로, 공통 키워드를 해당 링크의 메타데이터(Metadata)로 활용하여 수월하게 시맨틱 링크를 형성할 수 있다. 그러나 워드임베딩 기법을 통해 파악된 연관문서 간 링크는 코사인 유사도 값에 의해 정의되는데, 이 값은 링크의 강도를 설명할 수는 있으나 의미를 설명하지는 못하므로 시맨틱 링크 형성은 불가능하다.

4.2.4 대량 문서 처리 능력

고려하는 문서의 개수와 용량이 증가하면 연관문서 파악은 자동화된 방식으로 진행되어야 하는데, 이에 유리한 기법은 워드임베딩 기법이다. 키워드 기반 기법의 정확성 향상을 위해서는 동의어 및 관련어에 대한 지속적인 업데이트 및 보강이 필수적이나 이를 자동화된 방식으로 진행하는 것은 아직은 한계가 있다. 반면 워드임베딩 기법은 새로운 문서와 용어가 출현할지라도 이를 반영하는 코퍼스의 업데이트를 자동화된 방식으로 진행할 수 있으므로, 고려하는 문서의 개수와 용량에 관계없이 연관문서 파악을 위한 전 과정을 자동화할 수 있다. 또한 본 연구에서 활용한 Word2Vec보다 진보된 알고리듬으로 변경하는 경우에도 학습모델과 문서벡터를 해당 알고리듬에 맞춰 자동화된 방식으로 전환할 수 있으므로, 자동성과 확장성을 보장하는 연관문서 파악 방식은 워드임베딩 방식이라 할 수 있다.

5. 결 론

본 연구는 키워드 기반 기법과 워드임베딩 기법을 적용하여 도출된 연관문서 네트워크를 통해 각 기법의 정량 및 정성 성능을 비교하고, 이를 바탕으로 어느 기법이 정량 및 정성 관점에서 연관문서 서비스와 보안 관리에 유리한 기능성을 나타내는가를 검토하고 이유를 제시하였다.

성능 비교 결과, 키워드 기반 기법은 핵심 문서 파악과 시맨틱 정보 표현 능력이 우월하고, 워드임베딩 기법은 F1-Score 및 Accuracy, 연관성 강도 표현, 대량 문서 처리 능력이 우월한 것으로 나타났다. 따라서 어느 기법이 실제 기업과 사용자의 요구에 충족하는가에 대한 양분적 답을 제시하지는 못하나, 적어도 정량적 성능, 즉 F1-Score만 고려하여 연관문서 파악 기법을 선택해왔던 기존의 관행에 추가적인 고려가 반드시 필요함을 제기한다는 면에서 본 연구의 의의가 있다고 할 수 있다.

본 연구는 대상 문서의 개수를 10개로 한정하였는데, 이는 F1-Score와 Accuracy 산출에 필요한 정답안 구성의 정확성과 현실성을 극대화하기 위함이다. 즉, 전문가를 통해 문서 간 내용적 연관성을 파악하고 이를 기준으로 연관문서 네트워크 정답안을 도출하는 것이 가장 현실적이고 정확한 방법인데, 대상 문서의 개수가 늘어날 경우 쌍대(Pairwise) 비교를 통해 연관성을 파악하는 전문가의 오류 발생 여지를 제거하기 위한 인위적 조치였다. 보다 객관적인 성능의 산출을 위하여 대상 문서의 개수를 충분한 수준으로 증가시키고 이를 각 기법에 동일한 기준으로 적용하는 방법의 개발이 필요하다.

또한 본 연구의 워드임베딩 기법 기반 연관문서 파악 성능(F1-Score = 66.7%)이 상대적으로 낮게 파악되었는데, 이는 본 연구에서 사용한 Word2Vec의 성능에 의한 결과라기보다는, 학습모델의 성능 때문으로 판단된다. 즉, 학습모델 구성에 사용된 구글뉴스 코퍼스의 한계로 인해 전반적인 예측 성능이 저하된 것으로 보인다. 뉴스의 성격 상 학술 및 기술적 용어가 상대적으로 빈약하여 테스트 문서에 포함된 세부 학술 및 기술 용어들에 대해 적절한 대응을 하지 못하는 것으로 판단된다. 따라서 위키피디아 또는 학술 전자저널 등의 코퍼스로 대체하는 경우 향상된 성능 지표값이 산출될 수 있을 것으로 예상한다.

References

[1] Allan, J., “Building hypertext using information retrieval,” Information Processing & Management, Vol. 33, pp. 145-159, 1997.

[2] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T., “Enriching word vectors with subword information,” arXiv preprint arXiv:1607.04606, 2016.

[3] Choi, J. and Hwang, Y. S., “Patent keyword network analysis for improving technology development efficiency,” Technological Forecasting and Social Change, Vol. 83, pp. 170-182, 2014.

[4] Choi, J., Yi, S., and Lee, K. C., “Analysis of keyword networks in MIS research and

implications for predicting knowledge evolution," Information & Management, Vol. 48, pp. 371-381, 2011.

[5] Dai, A. M., Olah, C., and Le, Q. V., "Document embedding with paragraph vectors," arXiv preprint arXiv:1507.07998, 2015.

[6] De Boom, C., Canneyt, S., Demeester, T. and Dhoedt, B., "Representation learning for very short texts using weighted word embedding aggregation," Pattern Recognition Letters, Vol. 80, pp. 150-156, 2016.

[7] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[8] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., Advances in knowledge discovery and data mining, 21, AAAI press Menlo Park, 1996.

[9] Feldman, R. and Dagan, I., "Knowledge Discovery in Textual Databases (KDT)," Proceedings of the 1st Internatinal Conference on KDD, pp. 112-117, 1995.

[10] Frantzi, K., Ananiadou, S., and Mima, H., "Automatic recognition of multi-word terms: The C-value/NC-value Method," International Journal of Digital Libraries, Vol. 3, No. 2, pp. 117-132, 2000.

[11] Han, J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cuslck, M. E., Roth, F. P., and Vidal, M., "Evidence for dynamically organized modularity in the yeast protein-protein interaction network", Nature, Vol. 430, No. 6995, pp. 88-93, 2004.

[12] Haveliwala, T. H., Gionis, A., Klein, D., and Indyk, P., "Evaluating strategies for similarity search on the web," Proceedings of the 11th international conference on World Wide Web, pp. 432-442, 2002.

[13] Henzinger, M. R., "Hyperlink analysis for the web," IEEE Internet Computing, Vol. 5, pp. 45-50, 2001.

[14] Hwang, S. and Kim, D., "BERT-based Classification Model for Korean Documents," The Journal of Society for e-Business Studies, Vol. 25, No. 1, pp. 203-214, 2020.

[15] Kamkarhaghighi, M. and Makrehchi, M., "Content Tree Word Embedding for document representation," Expert Systems with Applications, Vol. 90, pp. 241-249, 2017.

[16] Kenter, T., Borisov, A., and De Rijke, M., "Siamese cbow: Optimizing word embeddings for sentence representations," arXiv preprint arXiv:1606.04640, 2016.

[17] Kil, H., "A Study on the Centrality Types of Reading Fingerprint Text," Journal of Cheongram Korean Language Education, Vol. 74, pp. 39-70, 2020.

[18] Klimek, P., Jovanovic, A. S., Egloff, R., and Schneider, R., "Successful fish go with the flow: Citation impact prediction based on centrality measures for term-document networks," Scientometrics, Vol. 107, pp. 1265-1282, 2016.

[19] Le, Q. and Mikolov, T., “Distributed representations of sentences and documents,” Proceedings of the International Conference on Machine Learning, pp. 1188-1196, 2014.

[20] Lee, D. and Kim, K., “Web Site Keyword Selection Method by Considering Semantic Similarity Based on Word2Vec,” The Journal of Society for e-Business Studies, Vol. 23, No. 2, pp. 83-96, 2018.

[21] Pennington, J., Socher, R., and Manning, C., “Glove: Global vectors for word representation,” Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543, 2014.

[22] Rose, S., Engel, D., Cramer, N., and Cowley, W., “Automatic keyword extraction from individual documents,” Text Mining: Applications and Theory, pp. 1-20, 2010.

[23] Yoo, K., “Application suite for autonomous management and service of verbal knowledge”, The Journal of Society for e-Business Studies, Vol. 21, No. 1, pp. 79-90, 2016.

[24] Yoo, K., “Keyword-based networked knowledge map expressing content relevance between knowledge,” Journal of Intelligence and Information Systems, Vol. 24, No. 3, pp. 119-134, 2018.

[25] Yoo, S. and Jeong, O., “An intelligent chatbot utilizing BERT model and knowledge graph,” The Journal of Society for e-Business Studies, Vol. 24, No. 3, pp. 87-98, 2019.

[26] Zhu, L., Liu, X., He, S., Shi, J., and Pang, M., “Keywords co-occurrence mapping knowledge domain research base on the theory of Big Data in oil and gas industry,” Scientometrics, Vol. 105, pp. 249-260, 2015.

[27] Zhuge, H. and Zhang, J., “Automatically constructing semantic link network on documents,” Concurrency and Computation: Practice and Experience, Vol. 23, pp. 956-971, 2011.

저 자 소 개

유기동	(e-mail: kdyoo@dankook.ac.kr)
2007년~현재	단국대학교 경영학부 교수
2006년	POSTECH 산업경영공학과 (공학박사)
2002년	POSTECH 산업공학과 (공학석사)
1998년	POSTECH 산업공학과 (공학사)
관심분야	경영정보시스템, 지식경영 및 지식관리시스템, 컨텍스트 기반 자율적 컴퓨팅, 지능적 지식 서비스, 정보전략 기획 및 성과평가