Journal of The Korea Society of Computer and Information
Vol. 22 No. 3, pp. 107-113, March 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.03.107

Effective Thematic Words Extraction from a Book using Compound Noun Phrase Synthesis Method

Hee-Jeong Ahn*, Kee-Won Kim**, Seung-Hoon Kim***

Abstract

Most of online bookstores are providing a user with the bibliographic book information rather than the concrete information such as thematic words and atmosphere. Especially, thematic words help a user to understand books and cast a wide net. In this paper, we propose an efficient extraction method of thematic words from book text by applying the compound noun and noun phrase synthetic method. The compound nouns represent the characteristics of a book in more detail than single nouns. The proposed method extracts the thematic word from book text by recognizing two types of noun phrases, such as a single noun and a compound noun combined with single nouns. The recognized single nouns, compound nouns, and noun phrases are calculated through TF-IDF weights and extracted as main words. In addition, this paper suggests a method to calculate the frequency of subject, object, and other roles separately, not just the sum of the frequencies of all nouns in the TF-IDF calculation method. Experiments is carried out in the field of economic management, and thematic word extraction verification is conducted through survey and book search. Thus, 9 out of the 10 experimental results used in this study indicate that the thematic word extracted by the proposed method is more effective in understanding the content. Also, it is confirmed that the thematic word extracted by the proposed method has a better book search result.

▸ Keyword : Book, Text mining, Thematic word, Extraction, Compound Noun Phrase

I. Introduction

인터넷의 발달로 전자 문서의 증가와 함께 도서 분야에서도 온라인 서점을 통해 거래되는 전자책의 개수와 전자책에 대한 독자들의 수요가 증가되고 있다. 현재 대부분의 온라인 서점에서는 도서의 내용이나 분위기와 같은 구체적인 정보보다는 저자 소개, 도서 설명, 목차와 같은 기본 정보만을 제공하고 있다. 사용자가 해당 도서에 대한 사전 지식이 없다면 위의 기본적인 정보만으로 도서 내용을 파악해야하는 것이 현실이다. 따라서 본 논문에서는 '주제어'라는 도서의 핵심 정보를 추출하여 사용자에게 도서에 대한 이해를 높이고 도서 선택의 범위를 넓히는 것이 목적이다.

주제어는 도서의 의미 또는 내용을 요약하거나 표현할 수 있고, 더 나아가 도서 검색 및 분류, 웹 검색 등 많은 분야에서 활

• First Author: Hee-Jeong Ahn, Corresponding Author: Seung-Hoon Kim
*Hee-Jeong Ahn(dreaminghee90@gmail.com), Dept. of Computer Science, Dankook University
**Kee-Won Kim(nirkim@dankook.ac.kr), Dept. of Applied Computer Engineering, Dankook University
***Seung-Hoon Kim(edina@dankook.ac.kr), Dept. of Applied Computer Engineering, Dankook University
• Received: 2017. 03. 07, Revised: 2017. 03. 14, Accepted: 2017. 03. 23.
• This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2016.

용될 수 있다 [1]. 기존의 주제어 추출 방법은 단일 명사를 이용하는 것이 대부분이었고[2-6], 검색 서비스를 위한 색인어 추출 연구에서 복합 명사와 명사구의 중요성이 언급되었다[7][8]. 이는 단일 명사보다 복합 명사와 명사구가 식별력이 크며 문서를 구체적으로 표현할 수 있기 때문이다[9][10]. 강승식(2001)은 복합 명사와 명사구를 인식하여 주제어 추출을 진행하였는데[11], 형태소 분석을 이용하지 않고, 조사의 생김 형태를 이용하였기 때문에 정확한 복합 명사구를 추출하는 데 어려움이 있다.

본 논문에서는 복합 명사와 명사구 합성 방법을 적용한 효과적인 도서 본문 주제어 추출 시스템을 제안한다. 복합 명사와 명사구의 인식이 특정 형태소 분석기에 의존적이지 않도록, 복합 명사와 명사구에 대한 정의를 내려 진행한다. 본 논문에서 복합 명사는 단일 명사가 결합된 형태이며, 명사구의 경우 조사가 생략된 경우와 관형격 조사가 결합된 경우 두 가지를 정의하여 주제어 추출에 사용하였다. 또한, 본 논문에서는 주제어 추출 과정에서 주어와 목적어의 중요성을 포함시킨 가중치 계산 방식을 제안한다. 일반적으로 사용되는 TF-IDF 방식에서는 추출된 명사의 빈도수만을 계산하여 진행하였다면 본 논문에서는 주어로 추출된 빈도와 목적어로 추출된 빈도, 그 외의 역할로 추출된 빈도에 대해 각각의 가중치 값을 주어 텍스트 안에서 더 영향력이 있는 명사가 추출되도록 한다.

논문의 구성은 다음과 같다. 2장에서는 주제어 추출 관련 연구를 기술하고, 3장은 복합 명사와 명사구 합성 방법을 적용한 주제어 추출 방법을 제안한다. 4장은 제안한 주제어 추출 방법을 이용한 실험 및 분석 결과를 제시한다. 마지막으로 5장에서는 결론 및 향후 연구 과제를 논의한다.

II. Related Works

주제어 추출이란 텍스트 마이닝의 한 분야로, 텍스트로부터 텍스트의 주제를 대표하는 단어를 자동으로 추출해내는 것을 말한다. 추출된 주제어는 이용자로 하여금 텍스트에 대한 이해와 텍스트 간의 관계를 쉽게 파악하도록 한다.

주제어 추출 방법에는 출현 빈도를 기반으로 하는 통계적 방법과 한국어의 형태소 분석, 구문 분석과 같은 기법을 사용하는 언어학적 방법, 또한 기계학습(Machine Learning)을 적용한 방법이 있다. 이성직(2009)는 통계적 방법으로 일반적으로 많이 사용되는 TF-IDF(Term Frequency-Inverse Document Frequency) 가중치 모델을 변형하여, 전체 문서 집합에 적용할 수 있는 여섯 가지 TF-IDF 변형식을 제안하였고[2], 한승희(2010)는 단일 문서를 대상으로 용어 클러스터링을 이용하여 빈도분포가 고르면서 주제성이 있는 키워드 추출 알고리즘을 제안하였다[3]. 언어학적 방법으로는 안희정(2015)은 문장 내 중요도와 문장과 문단에 따른 가중치에 근거한 도서 본문 주제어 추출 방법을 제안하였으며[4], 마지막으로 기계학습 방법을 적용한 조제현(2013)은 영어 학술 논문을 대상으로 Bayesian 알고리즘을 적용하여 주제어 추출의 성능을 개선하였다[5]. 주제어 추출에 있어 비교적 정확도가 높은 기계학습 방식의 경우, 추출 정확도가 학습 데이터에 의존적이므로 분야 별로 다양한 내용을 담고 있는 도서 분야에서 고품질의 학습 데이터를 준비하기에는 어려움이 있다.

일반적으로 문서를 대표하는 주제어 혹은 검색에서 사용되는 색인어는 단일 명사보다 복합 명사나 명사구가 식별력이 높으며 문서를 구체적으로 표현할 수 있다[9][10]. 복합 명사는 두 개 이상의 단일 명사가 서로 결합하여 새로운 의미를 가지며 구문적으로 한 단어로 파악되며, 그렇지 않은 경우는 구(Phrase)로 다룬다[7]. 복합 명사구 인식은 정보 검색 분야에서 두각을 나타내는데, 원형석(2000)은 효율적인 한국어 정보 검색을 위해 복합 명사 분할과 명사구 합성 기법을 제안하였고[7], 손기준(2004)은 특허 문헌 검색 시스템 향상을 위해 기존 TF-IDF 방법에 복합 명사 가중치를 적용하였다[8]. 복합 명사와 복합 명사구 인식을 적용한 주제어 추출 연구에는 '은/는', '의'와 같이 조사의 생김 형태로 명사와 명사구를 인식하여 복합 명사에 가중치를 적용한 방법이 있었다[11].

III. Thematic Word Extraction System

본 논문은 복합 명사구 합성 방법을 적용한 효과적인 도서 본문 주제어 추출을 제안한다. 본 논문에서 제안한 주제어 추출 시스템은 Fig.1과 같이 전처리 모듈, 복합 명사를 포함한 명사 및 명사구 인식 모듈, 가중치 부여 모듈로 구성된다. 전처리에서는 텍스트 안에서의 불용어 처리를 통해 쉼표, 따옴표와 같은 특수 기호들이 명사로 추출되지 않도록 처리하고, 형태소 분석기를 이용하여 형태소 분석 및 품사 태깅을 진행한다.

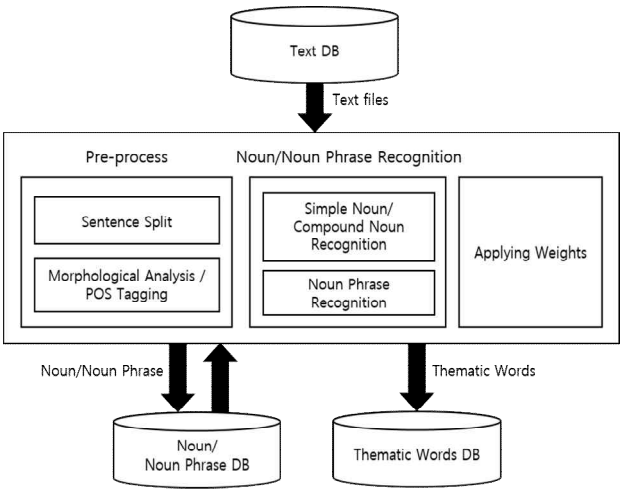

Fig. 1. Thematic Word Extraction System Architecture

1. Recognition of noun and noun phrase

본 논문에서는 '단일 명사', '복합 명사', '명사구'를 인식하여 주제어 추출을 진행한다. 형태소 분석기를 이용하는 경우 단일 명사 추출은 가능하지만, 둘 이상의 단일 명사가 결합된 복합 명사나 명사구의 형태를 자동으로 추출하지 못하기 때문에 본 논문에서는 복합 명사 및 명사구 인식을 다음과 같이 수행한다.

복합 명사 및 명사구 인식 모듈은 전처리에서 태깅된 품사 정보를 이용한다. 복합 명사, 명사구의 인식을 위한 규칙은 Table 1과 같다. 다양한 명사구의 종류가 필요한 검색 서비스와 달리 도서 주제어 추출은 사용자가 이해할 수 있는 명확한 의미를 가진 명사구가 요구되기 때문에 '조사가 생략된 명사구'와 '관형격 조사가 결합된 명사구'만을 인식한다.

Table 1. Recognition Patterns

Type	Pattern	Korean Example
Compound Noun	noun+noun	빅데이터
Noun phrase	noun+" "+noun	데이터 마이닝
	noun+Genitive postposition tag+" "+noun	인간의 뇌

인식을 위해 사용되는 품사는 사용하는 형태소 분석기에 따라 차이가 있다. Table 2는 한글 형태소 분석기 중 일반적으로 사용되는 '한나눔(HanNanum) 형태소 분석기[12]'와 '꼬꼬마(KKM) 형태소 분석기[13]'의 보통명사 및 관형격 조사 태그를 나타낸다. Table 2에서 ncp는 서술성 명사, ncn은 비서술성 명사, jcm은 관형격 조사, NNG는 일반 명사, JKG는 관형격 조사를 의미한다.

Table 2. Common noun and Genitive postposition tag

Type	HanNanum	KKMA
Common noun	ncp	NNG
	ncn	
Genitive postposition tag	jcm	JKG

Fig. 2는 단일 명사, 복합 명사, 명사구 합성 방법에 대한 순서도를 나타낸다. 꼬꼬마 형태소 분석기를 사용하는 경우, 보통명사 태그인 'NNG'와 관형격 조사인 'JKG'를 이용하여 복합 명사와 명사구를 인식하고 저장한다. 예를 들어, '북극곰의 눈물'의 경우 꼬꼬마 형태소 분석기는 [북극곰/NNG+ 의/JKG 눈물/NNG]로 분석되기 때문에 관형격 조사가 결합된 명사구인 '북극곰의 눈물'이 저장된다. 다른 예로 '빅데이터의 중요성'의 경우 [빅/NNG+ 데이터/NNG+ 의/JKG 중요성/NNG]으로 태깅되기 때문에 '빅데이터'라는 복합 명사의 합성 후에 '중요성'과의 명사구 합성이 진행된다.

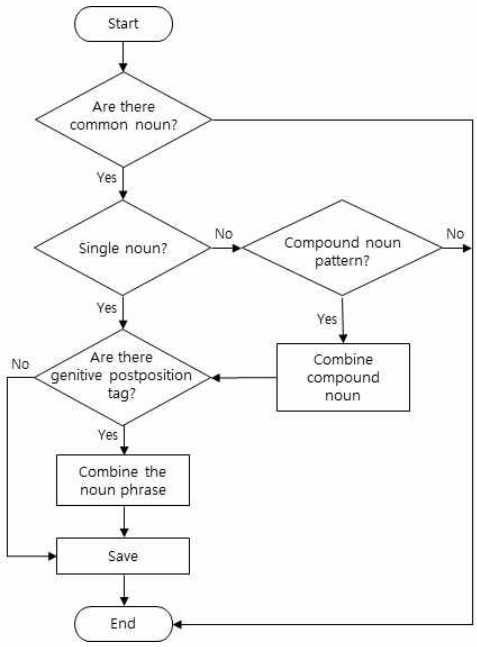

Fig. 2. Flow Chart of Synthesis of Compound Noun Phrase

2. Weighting method

가중치 부여는 앞서 추출된 단일 명사, 복합 명사, 명사구들의 중요도를 계산하는 과정이다. 일반적인 주제어 추출 시스템에서는 단어의 빈도와 역문서 빈도를 이용하는 통계적 가중치 부여 방법인 TF-IDF를 사용하는데, 본 논문에서는 '주어'와 '목적어'의 중요성을 포함시킨 TF-IDF 방식을 제안한다.

Table 3은 가중치 부여를 위해 사용된 기호 정의이며, 식 (1)은 본 논문에서 사용한 가중치 부여 공식을 나타낸다.

Table 3. Definition of Notation

Symbol	Description
$tf^s_{i,j}$	Term frequency of noun i used as the subject in text j
$tf^o_{i,j}$	Term frequency of noun i used as the object in text j
$tf^e_{i,j}$	Term frequency of noun i not used as the subject and the object in text j
df_i	Document Frequency of noun i in text set
$w_{i,j}$	Weight of noun i in text j

$$w_{i,j} = (\alpha \cdot tf^s_{i,j} + \beta \cdot tf^o_{i,j} + \gamma \cdot tf^e_{i,j}) \cdot \log(\frac{N}{df_i}) \quad (1)$$

본 논문에서 사용된 가중치 요소들은 주어와 목적어이며 이는 주어와 목적어가 문장 안에서 주제를 나타내는 표지로 사용된다는 이론을 근거로 진행하였다[14-16]. α는 주어로 사용된 명사 및 명사구에 대한 가중치 값이며, β는 목적어 가중치,

γ는 명사와 목적어 외의 명사 및 명사구에 대한 가중치 값이다. 각 가중치 α, β, γ는 0에서 1사이의 값을 가지며, 총 합은 1이다. 각 가중치의 값은 주제어 추출에 사용되는 텍스트의 종류 및 텍스트 구조에 따라 다르게 적용할 수 있다. 예를 들어, 소설의 경우 전문서적과 비교하여 주어와 목적어가 주제어와 상관없는 인물로 표현되는 경우가 많기 때문에 주어와 목적어의 가중치 α와 β값을 γ보다 상대적으로 작게 부여할 수 있다. 주어와 목적어에 대한 구분도 복합 명사, 명사구와 동일하게 태그로 구분한다. 한국어는 구성 요소들이 위치에 따라 기능을 할당받는 형상적 언어인 영어와 달리 구성요소들이 조사에 의해 그 기능이 표시된다[17]. 따라서 주어는 '은/는/이/가'에 해당되는 주격조사, 보격조사, 보조사로 구분하며, 목적어는 '을/를'에 해당되는 목적격조사를 이용하여 구분한다. Table 4는 한나눔 형태소 분석기와 꼬꼬마 형태소 분석기에 따른 조사 형태를 보여준다.

Table 4. 'Noun' and 'Object' Tag

Role	Type	HanNanum	KKMA
Subject	Nominative case proposition	jcs	JKS
	Complementary case proposition	jcc	JKC
	Auxiliary particle	jxc	JX
Object	Objective proposition	jco	JKO

IV. Experiment

본 논문에서 제안한 복합 명사와 명사구 합성 방법을 적용한 도서 주제어 추출 실험을 진행한다. 실험 데이터는 경제 분야 도서 텍스트 50권이며, 형태소 분석기는 한나눔 형태소 분석기를 사용하였다. 한나눔 형태소 분석기는 카이스트에서 개발한 한국어 형태소 분석기이며, 한국어 문장을 형태소 단위로 분석하여 품사 태그를 제공해주는 오픈 소프트웨어이다. 본 실험의 결과 비교는 기존의 주제어 추출 방법인 단일 명사만으로 TF-IDF를 진행한 결과와 비교한다. 본 실험에서는 주어, 목적어, 그 외의 경우에 대한 가중치 α,β,γ 값은 동일하게 부여한다.

1. Result of experiment

Table 5는 본 논문에서 제안한 방법을 적용한 도서 '2000만원으로 연봉 버는 경매투자'의 주제어 추출 결과 일부이며, Table 6은 실험 도서 50권의 결과 중 기존의 단일 명사만을 이용하여 주제어를 추출한 결과와 본 논문에서 제안한 방식의 주제어 추출 결과 일부를 나타낸다.

Table 5. The Result of Korean Thematic Word Extraction of 'Earn a year salary by investing 20 mil won in auction'

2000만원으로 연봉 버는 경매투자	$tf^s_{i,j}$	$tf^o_{i,j}$	$tf^e_{i,j}$	$w_{i,j}$
경매	65	45	224	0.006388
경매물건	42	36	52	0.004167
입찰	7	49	206	0.003501
세입자	71	12	66	0.003233
경매투자	9	9	34	0.003127
낙찰자	67	8	14	0.002285
경매시장	10	6	52	0.002207
부동산	47	77	55	0.002135
경매정보	8	8	31	0.002114
상가	37	14	37	0.001882

도서 '2000만원으로 연봉 버는 경매투자'의 주제어 추출 실험 결과를 비교해보면, 기존 단일 명사만을 이용한 결과에서는 볼 수 없는 복합 명사인 '경매물건', '경매투자', '경매시장', '경매정보'를 유추할 수 있다. 해당 도서의 경우 제목에서 유추할 수 있듯이 경매투자에 대한 내용을 담고 있기 때문에 '경매투자'에 대한 내용을 담고 있지 않은 기존의 주제어 추출 결과보다 더 나은 방법으로 판단할 수 있다. 또한, '3일 만에 주식 프₩로가 되는 책'의 경우 기존 TF-IDF에서 추출된 불용어 '주'가 본 논문에서 제안한 방식으로 적용한 결과에서는 상위 주제어로 추출되지 않은 것을 확인할 수 있다.

추가적으로 정확률과 재현율로 정확한 명사 추출과 불용어 처리에 대한 결과를 파악한다. 정확률은 추출된 상위 주제어 15개 중 정확하게 추출된 명사 및 명사구의 비율을 의미하며, 재현율은 정확하게 추출된 주제어 중 해당 텍스트와 관련이 없는 불용어를 제외한 주제어의 비율을 의미한다. 주제어가 추출된 도서 50권의 대한 정확률, 재현율의 결과는 정확률 평균 95.2%, 재현율 평균 98.5%를 보였다.

Table 6. Comparison of Korean Thematic Word Extraction Result

2000만원으로 연봉 버는 경매투자		3일 만에 주식 프로가 되는 책	
TF-IDF	Proposed	TF-IDF	Proposed
경매	경매	주가	주식
입찰	경매물건	주식	주식투자
낙찰	입찰	주식투자	이동평균선
상가	세입자	종목	종목
세입자관계	경매투자	증권주	증권주
부동산	낙찰자	주	기본적 분석
명도	경매시장	상투	투자기준
세입	부동산	일급	투자기법
말소	경매정보	주가폭락	상투
배당	상가	편람	위험관리

2. Verification of thematic words extraction

특정 도서 및 텍스트 안에서 추출된 주제어는 해당 단어가 다른 단어들보다 도서를 정확하게 나타내는지 대한 검증이 쉽지 않다. 이는 주제어라는 것이 주관성을 띄고 있으며, 추출된 주제어를 비교할만한 정답이 없기 때문이다. 따라서 본 논문에서는 구글에서 제공하는 설문조사 도구인 폼(form)을 이용하여 기존의 단일 명사만으로 추출된 주제어 세트와 본 논문에서 제안한 방식으로 추출된 주제어 세트의 결과 비교를 진행하였다.

2.1 Survey procedure

설문은 기존 방법으로 추출된 주제어 세트와 본 논문에서 제안한 방법으로 추출된 주제어 세트 중에서 주어진 제목의 도서 내용을 유추하는 데 더 효과적이라고 생각되는 것을 선택하도록 하였다. 설문에 사용된 도서는 위의 결과 데이터 50개 중 무작위 10개를 골라 진행하였다. Table 7은 설문지에서 사용된 주제어 세트 일부를 나타내며, 주제어는 추출된 주제어 중 상위 10개를 가지고 진행하였다. 선택지는 세 가지로 기존 TF-IDF 주제어 추출 결과와 본 논문에서 제안한 방법으로의 주제어 추출 결과, '잘 모르겠다.'로 구성하였다. Fig. 3은 설문지 일부를 나타낸다.

Table 7. Part of Korean Thematic Word List

Book	Thematic Words
A	주가, 주식, 주식투자, 종목, 증권주, 상투, 주가폭락, 편람, 일반투자자, 매도
	주식, 주식투자, 이동평균선, 종목, 증권주, 기본적 분석, 투자기준, 투자기법, 상투, 위험관리
B	디자인의 중요성, 감성, 제품, 이미지, 문화, 인간, 디자인의 역할, 아이덴티티, 디자인 경영, 기업문화
	디자인, 감성, 휴머니즘, 제품, 아이덴티티, 르네상스, 디자이너, 조형, 이미지, 프로세스
C	사업계획서, 작성가이드, 기재, 자금수지표, 목차, 표기, 손익계산서상, 로직트리, 자사, 투자유치
	사업계획서, 목차, 자금수지, 대차대조표, 손익계산서, 자사, 예산편성, 예제, 기재, 편성
D	빅데이터, 하둡, 데이터 마이닝, 데이터, 구글, 정용찬, 데이터 센터, 데이터 과학자, 정보통신정책, 데이터 시각화
	빅데이터, 데이터, 구글, 마이닝, 클라우드, 하둡, 개인정보, 정용찬, 웹, 비정형
E	대예측, 부동산, 임대아파트, 재건축, 용적률, 상가, 하노이시, 주택, 아파트, 리모델링
	부동산 대예측, 부동산 시장, 노무현 정권, 용적률, 토지사용권, 상가, 상가주택, 리모델링, 판교 신도시, 개발권 양도제

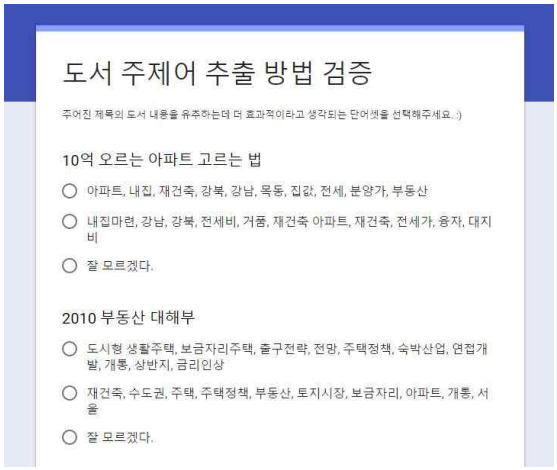

Fig. 3. Part of Survey

2.2 Survey result

설문은 본 논문과 관련이 없는 대학생 및 일반인, 총 35명 대상으로 실시하여 35명이 응답하였다. Table 8은 도서 10권에 대한 응답자들의 결과이다. 도서 10권 중 9권이 본 논문에서 제안한 방법으로 추출한 주제어가 효과적임을 나타내었으며, 응답 결과, 기존 방법으로 추출된 주제어보다 약 20%~55% 사이의 높은 선택을 받은 것을 확인할 수 있었다. 선택을 받지 못한 한 권의 도서의 경우, 단일 명사로 이루어진 주제어 세트가 복합 명사 및 명사구로 이루어진 주제어 세트보다 약 17% 정도 높게 선택되었다. 이는 기존 방법을 이용한 주제어 세트 중 해당 도서의 제목과 일치하는 단어가 존재하여 더 많은 선택을 받은 것으로 파악된다. 이 문제점은 추후 도서 제목, 주어, 목적어에 대한 가중치 연구로 개선될 수 있을 것으로 예상된다.

Table 8. Selection Ratio of Each Book

Book	A	B	C	D	E
1	70.6%	60.0%	35.3%	61.8%	51.4%
2	23.5%	31.4%	52.9%	17.6%	31.4%
3	5.9%	8.6%	11.8%	20.6%	17.1%
Total	100%	100%	100%	100%	100%
Book	F	G	H	I	J
1	71.4%	58.8%	74.3%	71.4%	57.1%
2	17.1%	26.5%	22.9%	17.1%	22.9%
3	11.4%	14.7%	2.9%	11.4%	20.0%
Total	100%	100%	100%	100%	100%

(1-Proposed method 2-Existing method 3-Abstention)

2.3 Verification of thematic words using Google

추출된 주제어들이 도서 본문의 내용을 효과적으로 담고 있는지에 대한 추가 검증으로 구글 도서 검색을 진행하였다. 추출된 주제어들로 검색을 하여 해당 도서가 몇 번째로 검색되는지 확인하였다. 본 실험은 경제경영 도서 50권 중, 구글 도서에서 본문 검색이 가능한 11권으로 2017년 1월 23일 진행하였다. 구글은 기본적인 서지 정보를 포함한 기본 검색과 함께 본문 검색 서비스를 제공하는데 이는 본문 데이터를 가지고 있는 경

우에 가능하다[18].

본문 검색은 추출된 주제어들 중 상위 3개를 가지고 진행하였으며, Table 9는 검색에 사용된 주제어 리스트이다. 검색 실험 결과, 본 논문에서 제안한 방식으로 추출된 주제어로 검색한 결과가 TF-IDF의 검색 결과 순위보다 높게 추출되는 것을 확인할 수 있었다. Table 10에서 볼 수 있듯이, 도서 3번을 제외한 모든 경우에서 본 논문 방식을 사용한 결과가 높은 순위를 나타내었다.

Table 9. List of Korean Thematic Word used in Book Search

Book	TF-IDF	Proposed method
1	뇌, 그레츠키, 청크	뇌, 청크, 어드밴티지
2	아파트, 내집, 재건축	내집마련, 강남, 강북
3	아파트, 주택, 주택보급률	통화량, 본원통화, 주택시장
4	경매, 빌라, 뉴타운	경매, 감정 가격, 서울 지역
5	데이터, 하둡, 과학자	빅 데이터, 데이터, 데이터 과학자들
6	빅데이터, 데이터, 마이닝	빅데이터, 하둡, 데이터 마이닝
7	정보관, 첩보원, 첩보	정보관, 첩보원, 첩보
8	예측지, 지능, 포캐스트	예측지, 예측경영, 예측지능
9	경매, 낙찰, 농지	경매, 낙찰, 농지
10	청약, 경매, 내집	내집마련, 부동산투자, 경매투자
11	현지, 브릭스, 글로벌	신흥국, 신흥국 시장, 볼륨 존

Table 10. Ranking Result of Book Search

Book	Ranking of search results	
	TF-IDF	Proposed method
1	1	1
2	3	1
3	2	4
4	1	2
5	11	2
6	3	2
7	1	1
8	1	1
9	4	4
10	32	1
11	7	1
Average	6	2

V. Conclusions

본 논문은 복합 명사 및 명사구 합성 방법을 적용한 효과적인 도서 텍스트의 주제어 추출을 제안하였다. 복합 명사는 둘 이상의 단일 명사가 결합된 형태로 인식되며, 명사구는 조사가 생략된 명사구와 관형격 조사가 결합된 명사구를 인식한다. 인식된 단일 명사, 복합 명사, 명사구는 TF-IDF 가중치를 통해 계산되어 주제어로 추출되게 된다. 본 논문에서는 추가적으로 TF-IDF의 계산 방식에서 모든 명사의 빈도의 합이 아닌 주어와 목적어, 그 외의 역할에 대한 빈도를 분리하여 계산하는 방식을 제안하였다. 실험은 경제경영 분야 도서 50권으로 진행되었으며, 주제어 추출 검증은 설문을 통해 실시하였다. 설문은 기존 단일 명사만으로 추출한 주제어와 본 논문에서 제안한 방법으로 추출된 주제어 중 도서를 파악하는데 효과적인 것을 선택하도록 하였다. 그 결과, 기존 방식으로 추출되었던 주제어보다 본 논문에서 제안한 방식으로 추출된 주제어가 약 20%~50% 사이의 높은 선택을 받았다.

향후에는 도서 유형 및 텍스트의 종류에 따라 본 논문에서 제안한 주어, 목적어에 대한 적합한 가중치를 찾아 주제어 추출을 더 정확하게 진행할 수 있도록 연구한다. 또한, 추출된 주제어를 이용하여 유사도서 분류, 군집을 통해 사용자에게 맞춤형 도서 추천을 해줄 수 있을 것으로 기대한다.

REFERENCES

[1] H. Shin, U. Yun, and K. H. Ryu, "Efficient Blog Retrieval System by Topic-based Weighting," Journal of the Korea Society of Computer and Information, Vol. 15, No. 4, pp.1-9, Apr. 2010.

[2] S. Lee and H. J. Kim, "Keyword Extraction from News Corpus using Modified TF-IDF," The Jounal of Society for e-Business Studies, Vol. 44, No. 4, pp.59-73, Nov. 22009.

[3] S. H. Han, "A Study on Keyword Extraction From a Single Document Using Term Clustering," Journal of the Korean Society for Library and Information Science, Vol. 44, No. 3, pp.155-173, Aug. 2010.

[4] H. J. Ahn, G. H. Choi, and S. H. Kim, "Thematic Word Extraction from Book Based on Keyword Weighting Method," Proceedings of the Korean Society of Computer Information Conference, Vol. 23, No. 1, pp.19-22, Jan. 2015.

[5] J. Cho and E. Paek, "Performance Improvements in Keyphrase Extraction via Candidate Phrase Selection Based on Natural Language Processing Techniques," Korean Institute of Information Scientists and Engineers, Vol. 40, pp.729-731, Nov. 2013.

[6] E. S. You, G. H. Choi, and 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, Feb. 2015.

[7] H. Won, M. Park and G. Lee, "Integrated Indexing Method using Compound Noun Segmentation and Noun Phrase Synthesis," Journal of KISS : Software and Applications, Vol. 27, No. 1, pp84-95, Jan. 2000.

[8] K. Son and S. Lee, "Weighting Methods for compound Nouns in Patent Retrieval System," Korean Institute of Information Scientists and Engineers, Vol.31, Issue 1, pp.895-897, Apr. 2004.

[9] C. E. Park, B. Ryu, and S.B. Kim, "A Segmentation Method of Compound Nouns Using Syllable Preference," Journal of Korea Multimedia Society, Vol. 9, No. 2, pp151-159, Feb. 2006.

[10] M. H. Cho and D. H. Jeong, "A Method Of Compound Noun Phrase Indexing for Resolving Syntactic Diversity," The Journal of the Korea Contents Association, Vol. 11, No. 3, pp.467-476, Mar. 2011.

[11] S. S. Kang, H. Lee, S. H. Son, G. C. Hong, and B. J. Moon, "Term Weighting Method by Postposition and Compound Noun Recognition," Korean Institute of Information Scientists and Engineers, Vol. 28, No. 2, pp.196-198, Oct. 2001.

[12] HANNANUM, http://semanticweb.kaist.ac.kr/hannanum/

[13] KKMA, http://kkma.snu.ac.kr/

[14] H. B. Lim, "Discourse-pragmatic notion of topic and syntactic analysis in Korea,", Seoul National University Press, 2007.

[15] Y. Jun, "On 'i/ka' as a Topic Marker," Discourse and Cognition, Vol. 16, No. 3, pp.217-238. Dec. 2009.

[16] D. H. Pak, "How to analyse and teach Korean specific particles 'i/ka' and 'eul/leul'," Foreign languages education. Vol. 14 No. 2, Jun. 2007.

[17] I. S. Choe and Y. M. Chung. "A Study on an Automatic Summarization System Using Verb-Based Sentence Patterns," Journal of the Korean Society for Information Management, Vol. 18. No. 4, pp.37-55. Dec. 2001.

[18] Google Books, https://books.google.co.kr/

Authors

Hee-Jeong Ahn received the B.S degree in Multimedia Engineering from Dankook University, Korea, in 2015 and M.S degree in Computer Science from Dankook University, Korea, in 2017. She is interested in Natural Language Processing, Text Mining, Data Analytics, etc.

Kee-Won Kim received his Ph.D. degree in Computer Engineering from Kyungpook National University in 2006, Republic of Korea. Currently, he is an assistant professor of Dept. of Applied Computer Engineering, Dankook University, Korea. His current research interests are cryptography, VLSI, network security, data mining, etc.

Seung-Hoon Kim received his Ph.D. degree in Computer Science and Engineering from Pohang University of Science and Technology (POSTECH), Korea in 1998.
Dr. Kim is currently a professor of Dept. of Applied Computer Engineering, Dankook University, Korea since 2001. From 1989 to 1990 he was a member of technical staff in Electronics and Telecommunications Research Institute(ETRI), Taejon, Korea. From 1991 to 1993 he was a member of technical staff in POSDATA, Seoul, Korea. His current research interests include data computing and networking, IoT, distributed systems, etc.