Journal of Information Technology and Architecture
Vol. 13. No. 2, June 2016, Pages 317-329

학술연구에서의 텍스트 마이닝 활용 현황 및 주요분석기법

김성근[1] · 조혁준[1] · 강주영[2]
[1]아주대학교 경영정보학과
[2]아주대학교 e-비즈니스 학과
kimsungguen7@gmail.com; ataraxia@gmail.com; jykang@ajou.ac.kr

(2016년 6월 10일 접수; 2016년 6월 28일 수정; 2016년 6월 30일 채택)

요약: 웹과 소셜미디어의 활용이 활발해짐에 따라 온라인에서 생성되는 비정형데이터 역시 기하급수적으로 증가하고 있으며, 이러한 비정형데이터 중에서도 텍스트 데이터에 대한 분석이 다양한 분야에서 이루어지고 있다. 본 논문은 학술논문 검색 사이트 디비피아 ('www.dbpia.co.kr')에서 텍스트 마이닝을 주제로 한 학술논문들의 제목과 연도, 학회 등의 정보를 수집하여, 국내 연구현황, 연구주제, 연구분야, 및 주요분석기법을 텍스트 마이닝 기법을 활용하여 분석하였다. 또한 학술논문들에서 많이 활용된 주요 텍스트 마이닝 기법들의 특징 및 분석방법을 조사하였다. 본 연구는 국내 연구 현황에 대한 객관적인 분석을 통해 텍스트 마이닝 연구의 현주소를 이해하고 향후 발전방향을 모색함으로써, 텍스트 마이닝 교육 및 연구에 기여하는 참고자료를 제공할 수 있을 것으로 기대된다.

키워드: 텍스트 마이닝, 토픽모델링, 연구동향, 워드투벡, 오피니언 마이닝

The Status of Using Text Mining in Academic Research and Analysis Methods

Sungguen Kim[1], Hyukjun Cho[1], and Juyoung Kang[2]*
[1]Ajou University, Department of MIS
[2]Ajou University, Department of e-Business

(Received June 10, 2016; Revised June 28, 2016; Accepted June 30, 2016)

Abstract : A large amount of unstructured data are generated online. Text data are analyzed in various fields. In this study we use DBpia (www.dbpia.co.kr) to collect data about academic research papers related to text mining. Based on that data, we then analyze the status of text mining research, using a text mining technique to find the topics, methods, and trends of text mining research. We also investigate the characteristics and method of analysis of the text mining techniques, confirmed by analysis of the data. This study provides reference materials for text mining education. It seeks to understand the current status of text mining research, through objective analysis of domestic research status, then exploring directions for future development.

Keywords : Text mining, Topic modeling, Research trend, Word2vec, Opinion mining

*교신저자
본 연구는 2015학년도 아주대학교 일반연구비 지원에 의하여 연구되었음.

1. 서 론

최근 수년 사이에 통신기술의 급격한 발전과 다양한 디지털 기기들의 등장으로 온라인 공간에서 수많은 정보들이 손쉽게 생산되고 빠르게 확산 될 수 있는 환경이 만들어 졌다. 계량화된 정형데이터뿐 아니라 다양한 비정형 데이터들이 축적되고 있으며, 그 중 텍스트 데이터는 가장 대표적인 비정형 데이터라고 할 수 있다. 이러한 비정형 데이터는 개개인의 삶에 엄청난 영향을 미치고 있다. 대표적으로 온라인 쇼핑몰에서 사람들은 물건을 구매할 때 다른 구매자가 남긴 제품리뷰 텍스트로부터 제품에 대한 정보를 수집한다. 또한 페이스북(FaceBook)과 같이 사람들 사이의 관계망을 구축해 주는 사회연결망 서비스(Social Network Service : SNS)는 방대한 텍스트 데이터들이 생산하고, 빠르게 텍스트데이터를 확산시킨다. 국내에서도 SNS사용자는 지속적으로 증가하고 있으며, 기업들은 SNS를 마케팅 채널로 활발히 활용하고 있다.

이처럼 방대한 텍스트 데이터가 개인의 일상뿐 아니라 유통, 제조, 서비스, IT 영역 등 다양한 영역에서 축적되면서 텍스트데이터에 대한 분석이 점점 더 중요해지고 있다. 이러한 시대적 변화에 따라 텍스트 데이터를 분석하는 연구기법과 방법론들이 등장하기 시작했다. Jaeseok et al. (2008)은 상품리뷰를 분석하는 시스템을 구축하고 시스템을 활용해 상품순위를 예측하였고, Yukun and Seogjun (2016) 은 비정형데이터의 수집을 통해 가치를 창출하는 빅데이터 비즈니스 모델에 대해 분석하였다. 또한 Yoosin et al. (2012)은 뉴스 텍스트 데이터의 분석을 통해 투자의사결정 모형 개발하였으며, Jaepil et al. (2016)은 텍스트 마이닝을 기반으로한 주식 투자전략 수립하여 우수한 성과를 입증하였다. 그리고 많은 연구들이 사람들의 의견이 표출된 텍스트로부터 특정 감정을 추출하고자 하였다 (Liu 2010; Narayanan et al. 2009; Sadamitsu et al. 2008). 이와 같이 많은 연구들을 통해 텍스트 데이터의 중요성이 더욱 더 높아지고 있음을 알 수 있다. 따라서 기업이나 대학 교육에서도 텍스트 마이닝 기법이나 방법론에 대해 체계적인 접근이 요구된다. 본 연구에서는 텍스트 데이터 분석을 주제로 한 학술연구 현황과 연도별 추세, 주제 및 분야를 분석하고, 학술연구에서 주로 사용되는 텍스트 마이닝 기법의 방법과 절차를 기술하였다. 또한 최근 구글에서 발표한 딥러닝 기반 텍스트 마이닝 기법인 워드투벡 (Word2vec)의 특징 및 분석절차를 소개하고, 향후 학술 연구에서의 활용범위에 대해서 알아보고자 한다.

본 연구는 학술논문 검색 사이트 디비피아를 활용하여 국내 텍스트 마이닝 연구 현황 및 기법에 대해 분석하고, 분석결과로 확인 된 주요텍스트 마이닝 기법과 이슈들을 추가적으로 조사하였다. 본 연구는 특정 연구분야나 기법에 초점을 두지 않고 국내 학술연구 분야 전반에 걸쳐 사용되는 텍스트 마이닝 기법들과 연구주제를 파악함으로써 향후 텍스트 마이닝 교육이나 연구에 참고가 됨은 물론 국내 연구현황을 객관적으로 파악하고 추후 국외연구 현황과 비교하여 텍스트 마이닝 연구의 발전방향을 모색하는데 유용한 자료로 활용 할 수 있다.

본 연구의 구성은 다음과 같다. 2장에서는 데이터 수집 절차 및 분석방법에 대해 설명하였으며, 3장에서는 데이터 연구현황분석 및 연구주제분석 결과를 기술하였다. 4장에서는 분석결과를 통해 확인 된 주요 텍스트 마이닝 기법들과 관련연구들에 관해 설명하였고, 5장에서는 연구결과를 바탕으로 국내 텍스트 마이닝 연구의 현황과 향후 텍스트 마이닝분야의 과제, 그리고 본 연구의 기여와 한계점에 대해서 논하였다.

2. 연구방법

본 연구는 학술논문 검색 사이트 디비피아를 활용하여 텍스트 마이닝 연구 현황에 대해 분석하였다. 연구현황분석을 위해 연도 별, 학회 별 논문 수를 집계하였다. 키워드 분석 및 토픽분석을 통해 주요 연구주제 및 분야와 분석기법들을 확인 하였다. <Figure 1>은 본 연구의 연구단계 및 방법을 정리한 것이다.

2.1 데이터수집

2.1.1 분석대상 선정

학술연구 제공서비스로는 국회도서관의 전자문서 검색서비스, 한국교육학술정보원의 RISS, ㈜한국학술정보 KISS, 누리미디어의 디비피아(DBpia)가 대표적이며, 이 중 국내에서 가장 많은 논문을 보유하고 있는 학술연구 제공 서비스는 디비피아(DBpia)이다. 디비피아(DBpia)는 2016년 5월 26일 기준으로 1,989,544건의 논문을 보유하고 있다. 따라서 본 연구는 데이터 수집 대상으로 가장 많은 논문을 보유하고 있는 디비피아(DBpia)를 사이트로 선정하였고, 사이트 내의 웹 페이

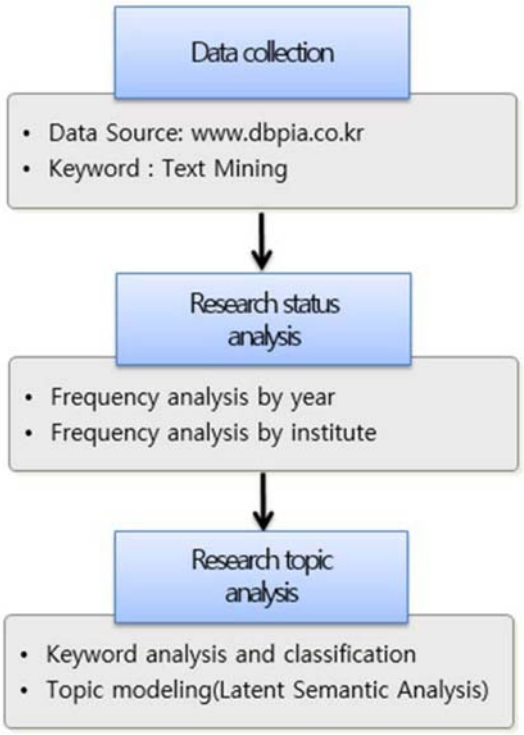

Figure 1. Research Procedure

지 URL을 추적하여 웹 크롤링을 통해 데이터를 수집하였다.

2.1.2 데이터 수집 및 전처리

데이터는 웹크롤링을 통해 수집되었으며, 수집 도구로는 R(버전 3.2.5)를 사용하였다. 데이터 수집을 위한 검색 키워드로는 '텍스트 마이닝'키워드를 사용하였다. 디비피아(DBpia)의 논문 검색기능은 두 종류의 검색기능을 지원한다. 첫 번째는 키워드 검색 시 본문을 제외한 논문의 정보를 바탕으로 검색결과를 제공하는 일반적인 검색기능이며, 두 번째는 본문을 포함한 정보를 바탕으로 검색결과를 제공하는 기능이다. 본 연구는 본문을 포함한 검색기능을 활용하였다. 본문을 포함한 검색기능을 활용하면 여러 키워드를 사용하지 않고 '텍스트마이닝' 한가지 키워드의 검색만으로도 텍스트 마이닝과 관련된 대부분의 논문을 조회할 수 있다는 장점이 있다.

실제 본문을 제외한 일반적인 검색을 통해 '텍스트 마이닝'키워드 검색 시 총 270편의 논문이 조회되었으며, 본문을 포함한 검색에서는 1750편의 논문이 조회되었다. 본 연구는 디비피아(DBpia)에서 '텍스트 마이닝' 키워드를 제목과 본문에 포함하고 있는 모든 논문의 제목, 학회, 저널, 발행일을 웹크롤링을 통해 수집하였으며, 총 1750편의 논문에 대한 데이터가 수집되었다. 1750개의 데이터 중 '00학회 학술대회 논문집', '00학회 논문집' 등이 논문의 제목으로 수집된 경우가 있었으며, 이러한 경우 제목으로부터 텍스트 마이닝과 관련된 세부적인 내용을 확인할 수 없기 때문에 데이터 전처리 과정을 통해 삭제되었다. 그 결과 최종 1570개의 논문에 대한 데이터가 수집되어 분석에 사용되었다. 학술연구의 제목이나 키워드 등은 띄어쓰기나 맞춤법이 정확하고, 줄임말이나 비속어가 사용되지 않았기 때문에 별도의 텍스트분석을 위한 전처리는 거치지 않았다.

2.2 분석방법

2.2.1 연구현황 분석(빈도분석)

학술연구에서 텍스트 마이닝의 활용현황을 분석하기 위해 연도별 논문현황, 학회 별 논문현황에 대한 빈도분석을 실시하였다. 그리고, 연구분야 및 주제를 확인하기 위해 논문제목에 대한 형태소 분석 후 명사와 용언에 해당하는 형태소를 추출하여 추출 된 키워드에 대한 빈도분석을 실시하였다. 형태소 분석에는 R의 KoNLP패키지에서 제공하는 SimplePos09 형태소분석기를 사용하였다. 분석결과는 3.2장에서 제시한다.

2.2.2 연구주제 분석 (잠재의미분석)

잠재의미분석 (Latent Semantic Analysis, LSA)은 단어의 의미를 동일한 문장에서 동시에 제시되는 단어들의 동시성 (co-occurrence)으로 정의되는데, 한 단어는 맥락들을 대표하는 축들로 구성된 다차원 상의한 점으로 표상되며, 단어 의미는 각 단어가 맥락 속에서 등장한 빈도로 정의된다(Landauer and Dumais 1997). LSA는 단어를 벡터로 표현하는 계량기법으로 문장을 구성하고 있는 개별단어는 각 문장에서 출현한 빈도로 기재된다. 10개의 문장에 특정 단어가 발생한 빈도가 0, 0, 0, 0, 2, 4, 5, 6, 7, 8 이라고 하면, 해당 단어는 10차원의 공간에서 (0, 0, 0, 0, 2, 4, 5, 6, 7, 8)의 벡터로 표현 될 수 있다. 이렇게 10차원의 공간에 표상된 벡터는 특이치 분해(Singular Value Decomposition, SVD) 정리를 이용해 더 작은 차원으로 축소하여 표상할 수 있다. 이

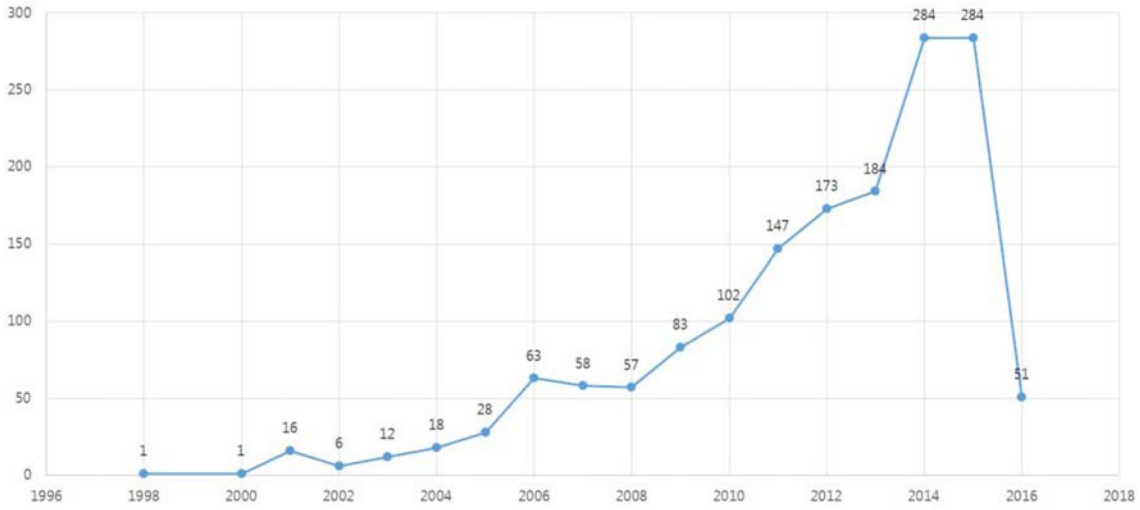

Figure 2. Frequency Analysis by Year

렇게 차원을 축소할 경우 원래의 차원에서 관찰되지 않았던 단어들이 같은 차원에서 관찰되게 되며, 같은 차원에서 표상된 단어들은 높은 상관을 보인다. 각각의 축소된 차원은 문서 내에 존재하는 하나의 주제를 의미하며 분석가는 단어목록을 바탕으로 주제를 추론할 수 있다.

본 연구에서는 텍스트 마이닝을 활용한 학술연구의 주요 주제들을 도출하기 위해 1570개의 논문 명에 대해 LSA 분석을 실시하였다. 분석결과는 3.2장에서 제시한다.

3. 분석결과

3.1 연구현황 분석

2.2절에서 언급한 바와 같이 텍스트 마이닝 연구 현황을 분석하기 위해 탐색적 분석 방법으로 가장 많이 쓰이는 방법 중 하나인 빈도 분석을 이용하여 수집된 1570개의 논문 자료들을 바탕으로 다음과 같이 세가지 형태의 분석을 수행하였다. 첫째, 연도별 텍스트 마이닝 논문현황에 대한 빈도분석을 수행하였고, 둘째, 학회 별 텍스트 마이닝 논문 현황에 대한 빈도분석을 수행하였으며, 셋째, 논문 제목에 자주 사용되는 키워드에 대한 빈도 분석을 수행하였다. 각각에 대한 상세한 분석 결과는 다음과 같다.

3.1.1 연도별 연구현황 분석

<Figure 2>는 수집된 1570개의 논문 자료들을 바탕으로 과거부터 2016년 5월 26일까지의 연도별 논문의 빈도를 분석 한 것이다. 텍스트 마이닝을 활용한 연구는 1998년부터 매년 꾸준히 증가하는 추세이며, 2011년의 증가율은 44%2014년의 증가율은 55%로 비교적 최근인 2011년 이후 텍스트 마이닝 관련 연구가 빠르게 증가 하고 있음을 알 수 있다.

3.1.2 학회 별 연구현황 분석

학회 별 텍스트 마이닝 관련 논문 게재현황을 파악하기 위해 학회 별 논문의 빈도를 분석하였다. 전체 177개 학회에서 텍스트 마이닝을 활용한 논문이 게재 되었으며, <Figure 3>은 텍스트 마이닝 논문이 가장 많이 게재된 상위 15개 학회에 대한 자료이다. 한국정보과학회에서 게재된 논문의 수는 390건으로 가장 많은 텍스트 마이닝 관련 논문이 게재되었으며, 상위7개의 학회에서 게재된 논문의 비율이 전체 논문의 약 50%를 차지하고 있다. 비교적 다양한 학문분야에서 텍스트 마이닝을 활용한 연구가 실시되고 있지만, 게재빈도가 50회 이상인 학회는 8곳밖에 되지 않는다는 것을 확인할 수 있다. 특히 산업공학, 컴퓨터공학 분야의 학회가 상위 15개 학회의 대부분을 차지하고 있으며 경영 및 인문, 사회과학 영역의 학회는 상위 15개 학회에 포함되어있지 않다.

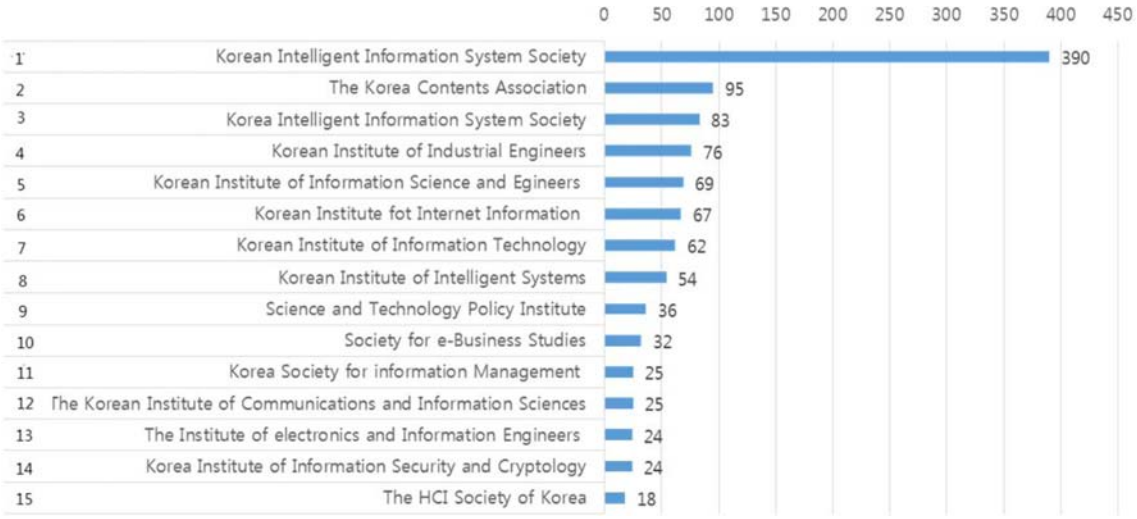

Figure 3. Frequency Analysis by Institute

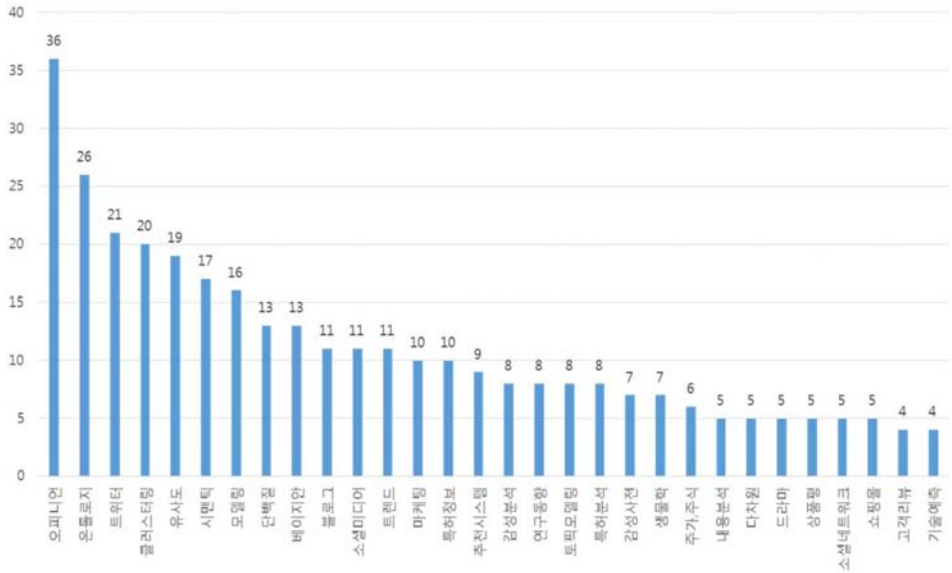

Figure 4. Research Topic Keyword Analysis

3.2 연구주제 분석

3.2.1 연구주제 키워드 분석

본 연구에서는 텍스트 마이닝이라는 키워드를 가진 학술 논문들의 연구 주제를 파악하기 위해 키워드 분석을 실시하였다. 이를 위해 본 연구는 수집 된 1570편의 논문제목에 대해 형태소 분석을 실시하였다. 전체 1570개의 논문 제목에서 명사와 용언을 추출하였고, 추출된 키워드의 빈도분석을 실시 하였다. 형태소 분석 결과 총 2020개의 형태소가 추출되었으며, '텍스트 마이닝', '텍스트', '마이닝', '빅데이터' 등과 같이 텍스트 마이닝 연구와 관련된 의미를 내포하지 않는 명사와 논문에서 자주 사용되는 '활용한', '기반한', '이용한', '고려한' 등의 용언은 키워드 분석에서 제외 하였다.

<Figure 4>는 논문의 제목에 자주 등장한 상위 30개 단어를 보여준다. 키워드 분석결과 '오피니언'이라는 용어가 가장 많이 등장하였으며 이러한 분석기법관련 용어와 '트위터', '블로그'와 같은 분석대상과 관련된 용

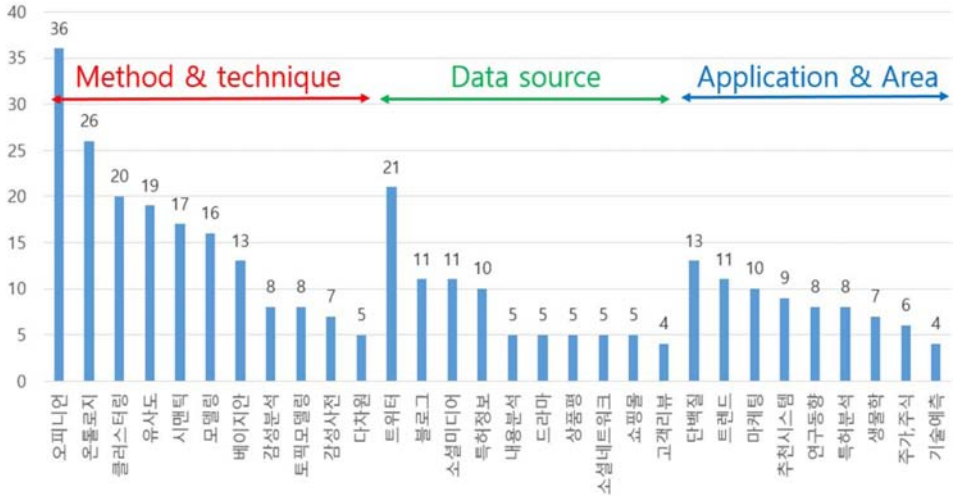

Figure 5. Classification of Keyword Sets

Table 1. Re-Classification of Keyword Sets

Method & Technique			Data Source			Application & Area		
Classification	Keyword	Frequency	Classification	Keyword	Frequency	Classification	Keyword	Frequency
Ontology	온톨로지	26	SNS	트위터	21	Trend prediction	트렌드	11
	클러스터링	20		블로그	5		연구동향	8
	시맨틱	17		소셜미디어	5		기술예측	4
Opinion mining	오피니언	36		소셜네트워크	5	Bio	단백질	13
	감성분석	8	Product review	상품평	10		생물학	7
	감성사전	7		쇼핑몰	5	Marketing	마케팅	10
Topic modeling	모델링	16		고객리뷰	4		추천시스템	9
	토픽모델링	8	etc	특허정보	5	Finance & Patent	주가,주식	6
etc	베이지안, 다차원	18		드라마	11		특허분석	8

어 그리고 '마케팅', '생물학' 등 활용 분야 및 영역을 나타내는 용어들이 혼재되어 있는 것을 볼 수 있다.

3.2.2 키워드 집합의 구성

본 연구에서는 분석결과 분석기법, 분석대상, 활용영역과 관련된 키워드가 혼재되어 있어, 보다 유의미한 분석을 하기 위해 내용 검토를 통해 키워드를 분석방법론, 분석대상, 활용분야 3가지 범주로 분류하여 키워드 집합을 구성하였다. <Figure 5>는 논문제목에 자주 등장한 상위 30개의 단어를 세 가지 범주로 재 정렬하여 그래프로 그린 것이다. 전체 키워드 중 분석기법 범주에 해당하는 키워드가 가장 많았으며 다음으로 분석대상 범주의 키워드와 활용영역 범주의 키워드가 많았다. 분석방법론 범주에서 가장 많이 활용 되고 있는 텍스트 마이닝 분석 기법은 오피니언 마이닝이며, 이와 관련성이 높은 '감성분석', '감성사전' 키워드를 함께 고려한다면 텍스트 마이닝을 활용한 국내연구 중 오피니언 마이닝기법이 가장 많이 활용되고 있음을 알 수 있다.

분석대상 키워드로는 '트위터', '블로그, '소셜미디어'가 가장 높은 빈도를 보였으며, 이는 SNS에서 생산된 텍스트데이터가 주된 분석 대상임을 파악할 수 있다. 활용분야 키워드로는 '단백질', '생물학' 키워드를 바탕으로 생물학 분야의 연구에서 텍스트 마이닝 기법이 활발히 활용되고 있음을 알 수 있으며, '마케팅', '추천시스템'을 바탕으로 마케팅 영역에 서도 텍스트 마이닝 기법을 연구에 활용하고 있는 것을 알 수 있다. <Table

Table 2. Result of Latent Semantic Analysis

Topic1	Topic2	Topic3	Topic4	Topic5
방법론	이용한	활용한	시스템	데이터
콘텐츠	한국어	빅데이터를	실시간	비정형
온톨로지	효과적	방법론	시맨틱	미디어
빅데이터를	키워드	환경민원	모바일	시각화
모델링	유사도	텍스트마이닝을	적용한	소프트웨어
적용한	빅데이터를	토픽모델링	트위터	영향요인
키워드	계층적	인터넷	프레임워크	하이퍼네트워크
실시간	클러스터링	패턴분석	필터링	콘텐츠
모니터링	마이닝을	탐색적	지능형	마이닝을
시맨틱	이미지	특허정보	개인화	오피니언
Topic6	Topic7	Topic8	Topic9	Topic10
텍스트	서비스	빅데이터	사용자	마이닝을
마이닝을	모바일	활용방안	프로파일	오피니언
오피니언	콘텐츠	프로세스	온톨로지	지능형
시각화	프레임워크	시각화	콘텐츠	효과적
마이닝	애플리케이션	보건복지	필터링	감성사전
디자인	온톨로지	과학기술	고려한	마이닝과
트렌드	큐레이션	처리과정	키워드	이미지
빅데이터를	활성화	추천시스템	트위터	방법론
온톨로지	유비쿼터스	활용효과	기반한	상품평
모바일	테크놀로지	요구사항	선호도	투자의사결정모형
Topic11	Topic12	Topic13	Topic14	Topic15
네트워크	온라인	알고리즘	마이닝	기반한
키워드	방법론	유사도	오피니언	미디어
전문가	고객리뷰	효율적	마이닝을	인터넷
베이지안	콘텐츠	맵리듀스	효율적	키워드
신뢰성	실시간	클러스터링	계층적	온톨로지
전자무역	효과적	베이지안	시각화	시각화
활용방안	마케팅	기초한	텍스트	오피니언
고려한	감성분석	유전자	서브토픽	클러스터링
방법론	텍스트마이닝	기계학습	모니터링	이미지
동기화	쇼핑몰	그래프	콘텐츠	지능형

1>은 각각의 범주 내에서 키워드의 내용검토를 통해 세부범주로 키워드를 재 분류한 것이다.

3.2.3 토픽분석

<Table 2>는 LSA 분석 결과를 보여주고 있다. 토픽분석을 위해 1570개의 논문 제목에 대해서 LSA 분석을 실시하였으며, 총 15개의 차원으로 분석하였다. LSA 분석은 유사도가 높은 단어를 같은 주제로 분류해 주며 같은 주제로 분류 된 단어집합을 보고 분석자가 직접 주제를 추론하게 된다. 1570개의 논문제목을 바탕으로 분석하였기 때문에 데이터가 충분하지 않아 정확한 주제를 추론하는데 한계가 있으나, 일부 주제는 단어집합을 바탕으로 주제를 추론이 가능하다. 예를 들어 토픽12의 경우 온라인, 쇼핑몰, 고객리뷰, 감성분석 등의 단어를 바탕으로 “고객리뷰에 대한 감성분석”을 주제로 추론할 수 있으며, 토픽13의 경우 맵리듀스, 클러스터링, 알고리즘, 기계학습, 베이지안 등의 키워드를 바탕으로 “데이터분석과 관련된 기술과 방법론”을

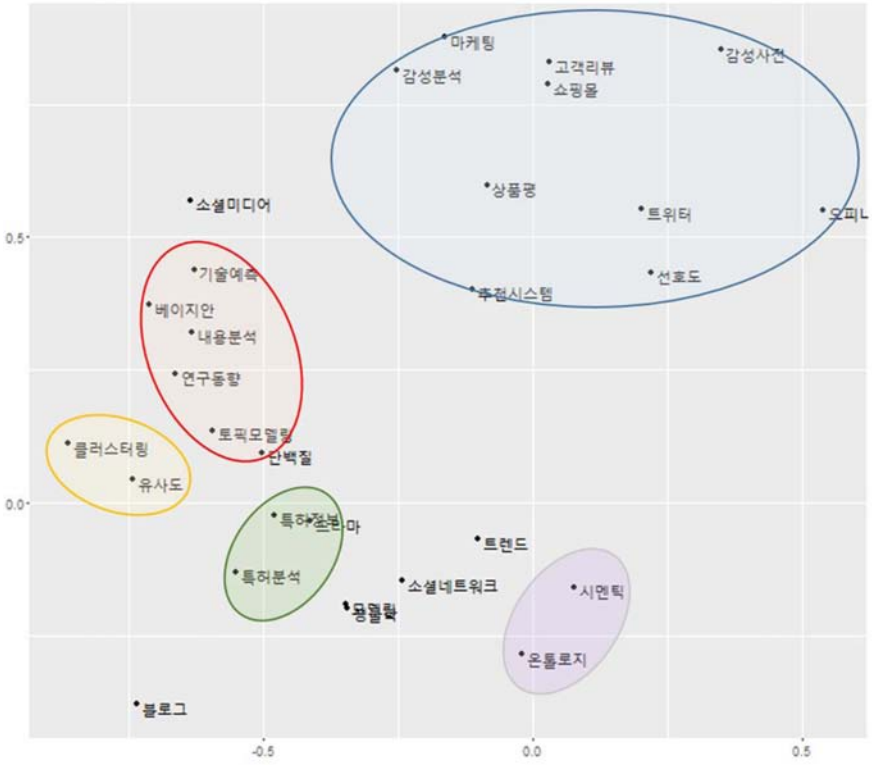

Figure 6. Visualization of Latent Semantic Analysis and PCA

주제로 추론할 수 있다.

<Figure 6>은 LSA의 결과를 주성분 분석을 통해 2개차원으로 축소하여 2차원상에 단어들의 관계를 표현한 것이다. <Figure 6> 에서 위치가 근접한 단어들은 유사도가 높은 단어들이다. 토픽12에서 관찰되었던 고객리뷰, 쇼핑몰, 상품평, 감성분석 등 마케팅관련 키워드는 그림의 우측 상단에 분포하고 있으며, 기술예측, 연구동향, 토픽모델링 등의 토픽 및 동향분석 관련 키워드는 그림의 좌측에 분포하고 있는 것을 볼 수 있다. 키워드 빈도분석결과에서 분석기법관련 키워드를 오피니언 마이닝, 토픽모델링, 시멘틱웹 및 온톨로지로 세 가지로 분류한 것과 유사하게 군집이 형성된 것을 <Figure 6>에서 확인 할 수 있다.

4. 주요 텍스트 마이닝 분석기법

텍스트 마이닝은 비정형 또는 반 정형 텍스트 데이터에서 자연어 처리 기술에 기반하여 유용한 정보를 추출, 가공하는 데이터분석 기법이다. 4장에서 키워드 빈도분석 및 주제분석 결과, 학술연구에서 가장 많이 활용하는 분석기법으로는 오피니언 마이닝, 토픽모델링, 시멘틱 웹 및 온톨로지 기법으로 나타났다. 시맨틱 웹 기술은 온톨로지를 활용한 지식 표현 모델로, 추론엔진을 지능형 확장 도구로 사용하는 의미 기반 지능형 웹을 말한다. 이는 빅데이터의 등장 이전부터 존재했던 지능형 웹 기술로 텍스트데이터에 대한 자연어처리 및 기계학습의 개념을 포함하는 텍스트 마이닝과는 다른 특성의 기술로 구분될 수 있다. 따라서 4장에서는 가장 기초적인 텍스트 마이닝 기법인 형태소 및 키워드 분석과 3장의 분석결과로 확인된 주요 텍스트 마이닝 기법으로 오피니언 마이닝(감성분석), 토픽모델링(LDA, Word2vec)에 대해 설명하도록 한다.

4.1 형태소 및 키워드 분석

텍스트를 분석하기 위해서는 대부분의 경우 문장을 개별단어로 분리하는 토큰화 과정(토크나이즈)이 필요하다. 한국어의 경우 문장을 개별단어로 분리하기 위해

서 형태소분석을 실시하며, 형태소 분석은 R이나 Python에서 제공하는 자연어 처리 패키지를 통해 분석할 수 있다. 형태소분석을 통해 명사, 용언 등의 형태소를 추출하여 발생빈도를 파악하여 문서의 주제 또는 문서에 대한 태도나 감성을 추론 할 수 있다. 또한 단어들의 동시발생 빈도를 파악하여 단어들간의 관련성이나 상관관계를 파악하고 네크워크 그래프로 표현할 수 있기 때문에 다양한 연구에서 활용되고 있다 (Cheolwoo and JaeJun 2012; HaJin and Min 2014; Manuk et al. 2009; Sunghoon and Hakyoun 2015).

4.2 오피니언 마이닝

오피니언 마이닝은 글쓴이가 서술 대상에 대하여 말하고자 하는 의견을 임의의 문서로부터 찾아내는 분야로서 최근 여러 연구에 의하여 성숙되고 있으며 감성분석 (Sentiment Analysis), 감성 분류 (Sentiment Classification), 오피니언 추출 (Opinion Extraction)로 표현되기도 한다. 오피니언 마이닝 기법의 활용을 통해 소셜 미디어와 같은 온라인상의 여론을 신속하게 파악할 수 있으며, 기존의 오픈 라인 여론조사에 보다 시간과 비용을 절약하여 사람들의 의견을 빠르게 파악하고 예측할 수 있기 때문에 활용도가 높은 분야로 평가 받고 있다(Kweon Yang and Chang Suk 2009; Sang-Do et al. 2010).

오피니언 마이닝은 다음과 같이 세가지 하위 단계로 구성된다. 첫째는 긍정 또는 부정의 극성을 판별하는 방향을 결정하는 것이며, 두 번째는 텍스트에서 의견표현부분을 추출하는 것이고, 세 번째는 사용자의 의견을 나타내는 언어학적 자원을 정의하고 구축하는 것이다 (Jaeseok et al. 2008).

이 중, 극성을 판별하는 감성 분석은 일반적으로 긍정 또는 부정 단어의 수를 합산하거나 감정사전에서 정의한 값을 합산하여 극성을 판별하며, 형태소분석을 통해 용언과 명사를 추출하여 사용자의 의견표현을 파악하고, 기존에 구축되어있는 범용감성사전을 사용하거나 직접 감성사전을 구축하여 사용할 수 있다. 영어텍스트를 기반으로 한 감성분석의 경우 SentiWordNet을 활용한 연구들이 활발히 이루어져 왔다(Hamouda and Rohaim 2011; Hanhoon et al. 2012; In-Su 2013; Ohana and Tierney 2009; Saggionα and Funk 2010). 한국어의 경우 연세대학교에서 개발된 한글 감성어 사전 API인 '오픈한글' (Jungkook and Hee-Woong 2015)

또는 서울대에서 개발한 KOSAC (Korean Sentiment Analysis Corpus) (Jang et al. 2013) 등이 사용 가능하다. 그러나 범용 감성사전보다 도메인 감성 사전을 구축하였을 때 더 높은 분류 성능을 보인다는 연구결과들이 있다 (Dahae et al. 2015; JungHo et al. 2015). 이러한 연구에도 불구하고 한글의 경우, 외국에 비해 코퍼스나 감성사전이 부족한 실정이며 더 활발한 연구가 필요하다(JungYeol and Chan 2014).

4.3 토픽모델링(LDA)

토픽모델링은 방대한 텍스트자료로부터 특정 주제를 추출하는 알고리즘이다. 대표적인 토픽모델링 알고리즘으로는 잠재의미분석(Latent sementic analysis)과, 잠재디리클래할당(Latent Dirichlet Allocation)이 있으며, 4.4장에 설명 된 Word2vec알고리즘을 활용해서 주제를 추론할 수도 있다. 3.2장에서 LSA대해서 기술하였으므로 토픽모델링의 방법론으로 LDA와 Word2vec에 대해 설명하도록 하겠다.

Blei et al. (2003)는 확률기법을 기반으로 문서의 토픽을 확인할 수 있는 알고리즘인 LDA (Latent Dirichlet Allocation)를 제안하였다. LDA는 주어진 문서 내 주제를 잠재적인 가정하는 확률모델로 문서에 내에 주제들의 확률분포 θ와 각 주제를 구성하는 단어들의 확률 분포z가 주어졌을 때, 문서를 구성하는 주제를 확률적으로 선택하고 선택된 주제에 존재하는 단어를 확률적으로 선택하는 샘플링과정을 반복함으로써 임의의 문서를 생성하는 모델이다 (Blei et al. 2003).

LDA는 주어진 문서와 사전에 정의된 문서 내에 토픽들의 사전확률 분포인 α와 토픽 내에서 단어의 사전확률분포인 β의 파라미터 값을 활용해 z와 θ를 추정한다. Blei (2012)는 'Science'저널과 'Yale Law'저널을

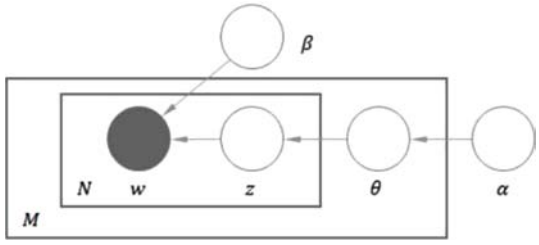

Figure 7. Latent dirichlet Allocation Model

```
In [11]: print (sentences[0])

['with', 'all', 'this', 'stuff', 'going', 'down', 'at', 'the', 'moment', 'with', 'mj', 'i',
've', 'started', 'listening', 'to', 'his', 'music', 'watching', 'the', 'odd', 'documentary',
'here', 'and', 'there', 'watched', 'the', 'wiz', 'and', 'watched', 'moonwalker', 'again']
```

Figure 8. Example of Word2vec Input

```
Out[114]: array([ 0.114587  ,  0.14587821,  0.10506593, -0.03043565,  0.00371807,
                 -0.04472743,  0.05764203,  0.01767993, -0.04527798,  0.02006411,
                  0.11546931,  0.01672839, -0.01837232, -0.04834358,  0.08837188,
                  0.06663876, -0.07159595,  0.0070584 ,  0.02129613,  0.05653455,
                  0.07622128,  0.00130508, -0.01912388,  0.01011143, -0.01728549,
                 -0.03013886,  0.05815636,  0.03524212, -0.01099172, -0.01827165,
                  0.11007821,  0.02696958, -0.06023923,  0.07584872, -0.01576053,
                  0.00376411,  0.04731355,  0.00038145,  0.04461492, -0.04201531,
                  0.11486857,  0.01165507, -0.0649289 , -0.01950942, -0.08592417,
                  0.01449627,  0.03251954,  0.03993464,  0.07306941, -0.03070936,
                 -0.07778312,  0.06415024, -0.02020006,  0.09563055, -0.03555943,
                 -0.06125034, -0.13431963, -0.05056937, -0.07600672,  0.03526962,
```

Figure 9. Example of Word2vec Output

대상으로 LDA기법을 활용하여 토픽모델을 도출하였으며, Keeheon et al. (2015)은 생의학분야 연구동향을 분석하는 등 LDA기법은 다양한 분야의 기술 및 연구동향을 분석하는데 자주 활용 된다(Byeongmun et al. 2015; Ja-Hyun and Min 2013; Kwang-Seob et al. 2012; Kyeonghee et al. 2016; Seung Hoon et al. 2015).

4.4 워드투벡 (Word2vec)

Word2vec(Mikolov et al. 2013)은 구글의 연구원들이 작성한 “Efficient Estimation of Word Representations in Vector Space” 에서 제안된 방법을 구현한 알고리즘이다. Word2Vec은 텍스트 문서를 통해 학습을 진행하며 한 단어에 대해 인접하여 출현하는 다른 단어들을 관련 단어로서 인공 신경망에 학습시킨다. 즉, 단어의 순서가 근접하여 자주출현 할수록 단어들은 유사한 벡터 값을 갖게 된다. 단어를 벡터 값으로 계량화 한다는 점은 LSA와 동일하지만 특정 단어의 앞뒤에 위치한 단어들의 분포를 고려하여 신경망에 학습시킨다는 것에서 차이가 있다.

학습 알고리즘으로 CBOW(continuous bag of words)와 Skip gram이 있으며 CBOW는 특정 단어의 주변단어를 이용해 특정단어의 다차원 벡터를 생성하며, Skip gram은 특정 단어를 바탕으로 주변 단어의 다차원 벡터를 생성한다. 대부분의 텍스트 마이닝 기법에서는 데이터 전처리 과정에서 불용어를 모두 삭제하는 경우가 많으나, Word2vec기법은 특정 단어들의 주변에 발생한 단어들을 바탕으로 학습을 하기 때문에 불용어가 삭제되지 않은 상태의 텍스트를 그대로 학습 데이터로 사용한다. Word2vec기법은 분석결과로 각 단어에 대해서 최대 300차원의 벡터 값을 생성하기 때문에 기존의 기계학습 알고리즘을 적용할 수 있으며, 토픽모델링, 기계번역, 주가예측이나 추천시스템 등 다양한 분야에 있어서 활용이 가능하다(Choi and Koo 2015; Dongsuk et al. 2016; Hyosup et al. 2015; Jonghee et al. 2015). Word2vec의 인풋데이터의 형태는 <Figure 8>과 같으며, 아웃풋의 형태는 <Figure 9>와 같다.

5. 결 론

본 연구에서는 웹크롤링을 통해 수집된 텍스트 마이닝 관련 논문들에 대해서 연도별 텍스트 마이닝 연구추이, 학회 별 게재 논문 수, 논문의 주제 탐색을 위한 키워드 분석과 토픽분석을 실시 하였다. 연구 결과 비교적 최근인 2011년 이후 텍스트 마이닝을 활용한 학술연구가 꾸준히 증가하였고, 전체 177개 의 다양한 학회에서 텍스트 마이닝을 활용한 논문들이 개재되었다. 그러나 3.1.2절의 결과와 같이 상위 7개학회 논문이 전체논문의 50% 이상을 차지하는 것은, 아직 일부 학술분야에서만 텍스트 마이닝 기법을 연구에 활용하고 있음을 의미한다. 또한, 키워드 및 토픽분석을 통해 오피니언 마이닝, 감성분석, 토픽모델링, 온톨로지기법 등

이 텍스트 마이닝 연구에 주로 활용되고 있음을 확인 할 수 있었다. 또한 분석자료로는 SNS, 고객리뷰, 특허정보가 있었으며, 활용분야로는 기술예측, 마케팅, 생물학 등의 분야에서 텍스트 마이닝 기법이 활용되고 있음을 확인 할 수 있었다.

텍스트 마이닝은 다양한 산업영역에서 유용하게 활용될 수 있을 것으로 기대된다. 개인화된 마케팅, 추천시스템, 기업 브랜드 이미지관리, 기업과 공공기관의 고객의 소리 (Voice of Customer, VOD), 기계번역, 문서분류 및 평가의 자동화 등 텍스트 마이닝 방법론은 매우 넓은 활용 범위를 가지고 있다. 이러한 이유로 텍스트 마이닝은 빅데이터 분석의 중요 기법 중 하나로 각광 받고 있지만, 실제 국내 연구의 경우 감성분석이나 토픽분석 위주의 매우 국한된 주제로 연구가 이뤄지고 있다. 연구 대상 또한 트위터와 제품리뷰 등으로 국한되어 있는 실정이다.

그 이유는 아직까지 텍스트 분석의 정확도가 상대적으로 정형화된 데이터 분석에 비해 낮기 때문이며, 이를 보완하기 위해서는 한국어 분석을 위한 언어학적 자원이 확보되어야 한다. 영어의 경우 많은 코퍼스와 감성사전 및 텍스트 분석 툴이 존재하지만 한국어의 경우 이러한 언어학적 자원이 절대적으로 부족한 상황이다. 글로벌 IT 기업들이 Word2vec과 같은 진보된 기술을 만들고 공급해 줄 수 있지만 한국어의 언어학적 자원을 개발하는 것은 국내연구자들만이 할 수 있는 일이다. 이러한 문제가 해결된다면, 보다 정교한 텍스트분석이 가능해질 것이고 더욱 폭넓은 영역에서 텍스트 마이닝 활용이 가능해 질 것이다.

본 연구의 기여는 다음과 같다. 첫째, 국내 텍스트 마이닝 연구의 주된 주제와 기법을 분석함으로써 국내 텍스트 마이닝 연구 현주소를 객관적으로 확인 할 수 있었다. 둘째, 본 연구는 연구자들에게 학술연구에서 활용하는 주요 텍스트 마이닝 기법을 소개함으로써 텍스트 마이닝 연구에 적용할 수 있는 기법을 판단하는데 참고자료를 제공하고 있다. 셋째, 추후 해외 텍스트 마이닝 연구들에 대한 조사 및 분석이 이뤄질 경우 본 연구는 이를 비교할 수 있는 국내자료를 제공함으로써 국내외 차이점을 분석하고 발전 방향을 모색 할 수 있다.

본 연구의 한계점으로는 비록 본문을 포함하여 검색하였음에도 불구하고, '텍스트마이닝' 한 가지 키워드를 활용하여 데이터를 수집하였기 때문에 해당 키워드를 포함하지 않은 논문은 수집하지 못하였다는 점이다. 실제 온톨로지, 시맨틱 웹과 같이 해당 키워드를 포함하지 않은 텍스트 마이닝 연구도 있기 때문에 향후 다양한 키워드 목록을 구성하여 좀 더 면밀히 데이터를 수집할 필요가 있다. 후속 연구로는 해외 텍스트 마이닝 현황 및 영역, 목적 등을 조사하여 이를 국내 연구현황과 비교한다면, 국내 텍스트 마이닝 연구의 발전방향을 모색하는데 기여할 수 있을 것이라 생각된다.

References

[1] Blei, D.M., "Probabilistic Topic Models", *Communications of the ACM,* Vol. 55 No. 4, pp. 77-84, 2012.

[2] Blei, D.M., A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation", *The Journal of machine Learning research,* Vol. 3, pp. 993-1022, 2003.

[3] Byeongmun, J., K. Taehwan, L. Jin, and K. Jungsun, "Twitter Topic Extraction and Topic Category Decision Using Lda Model", paper presented at the Korea Information Science Society, 2015.

[4] Cheolwoo, J. and K. JaeJun, "Analysis of Foresight Keywords in Construction Using Complexity Network Method", *Journal of the Korean Digital Architecture Interior Association,* Vol. 12 No. 2, pp. 15-23, 2012.

[5] Choi, I. and P. Koo, "Prediction of Stock Fluctuation Using Web News Text Mining", paper presented at the Korea Institute of Industrial Engineers, 2015.

[6] Dahae, K., C. Taemin, and L. Jee-Hyong, "A Domain Adaptive Sentiment Dictionary Construction Method for Domain Sentiment Analysis", paper presented at the The Korean Society Of Computer And Information, 2015.

[7] Dongsuk, O., K. Sangwoo, and S. Jungyun, "An Iterative Approach to Graph-Based Word Sense Disambiguation Using Word2vec", *KOREAN JOURNAL OF COGNITIVE SCIENCE,* Vol. 27 No. 1, pp. 43-60, 2016.

[8] HaJin, K. and S. Min, "A Study on the Research Trends in Domestic/International Information Science Articles by Co-Word Analysis", *Journal of the Korean Society for Information Management,* Vol. 31 No. 1:99-118, 2014.

[9] Hamouda, A. and M. Rohaim, "Reviews Classification Using Sentiwordnet Lexicon", paper presented at the World Congress on Computer Science and Information Technology, 2011.

[10] Hanhoon, K., Y. Seong Joon, and Y. Dongil, "Senti-Lexicon and Improved Naïve Bayes Algorithms for

Sentiment Analysis of Restaurant Reviews", *Expert Systems with Applications,* Vol. 39 No. 5:6000-6010, 2012.

[11] Hyosup, S., Y. Hwan jo, and L. Gary Geunbae, "Disambiguation of Relation in Knowledge Base Question Answering Using Ann-Based Parallel Semantic Model", paper presented at the Korea Information Science Society, 2015.

[12] In-Su, K., "A Comparative Study on Using Sentiwordnet for English Twitter Sentiment Analysis", *Journal of Korean Institute of Intelligent Systems,* Vol. 23 No. 4, 317-324, 2013.

[13] Ja-Hyun, P. and S. Min, "A Study on the Research Trends in Library & Information Science in Korea Using Topic Modeling", *Journal of the Korean Society for Information Management,* Vol. 30 No. 1, pp. 7-32, 2013.

[14] Jaepil, R., H.C. Hoon, and S.H. Joon, "Sector Investment Strategies Using Big Data Trends", *Information Technology and Architecture,* Vol. 13 No. 1, pp. 111-121, 2016.

[15] Jaeseok, M., L. Dongjoo, and L. Sang-goo, "A Korean Product Review Analysis System Using a Semi-Automatically Constructed Semantic Dictionary", *Journal of KISS : Software and Applications,* Vol. 35, No. 6, pp. 392-403, 2008.

[16] Jang, H., M. Kim, and H. Shin, "Kosac: A Full-Fledged Korean Sentiment Analysis Corpus", *Sponsors: National Science Council, Executive Yuan, ROC Institute of Linguistics, Academia Sinica NCCU Office of Research and Development,* Vol.:366, 2013.

[17] Jonghee, P., P. Eunjeong, and J. DongJoon, "Automated Text Analysis of North Korean New Year Addresses, 1946-2015", *Korean Political Science Review,* Vol. 49 No. 2, pp. 27-61, 2015.

[18] JungHo, K., O. YeanJu, and C. SooHoan, "The Construction of a Domain-Specific Sentiment Dictionary Using Graph-Based Semi-Supervised Learning Method", *Science of Emotion and Sensibility,* Vol. 18 No. 1:97-104, 2015.

[19] Jungkook, A. and K. Hee-Woong, "Building a Korean Sentiment Lexicon Using Collective Intelligence", *Journal of Intelligence and Information Systems,* Vol. 21 No. 2, pp. 49-67, 2015.

[20] JungYeol, S. and K. Chan, "Big Data Analysis by Sensitivity Analysis", *The Society of Convergence Knowledge Transactions,* Vol. 2 No. 1, pp. 15-21, 2014.

[21] Keeheon, L., H. Go Eun, and S. Min, "Biomedical Text Mining : A Trend in Biomedical Knowledge Discovery Based on Text Mining", *Communications of the Korean Institute of Information Scientists and Engineers,* Vol. 33 No. 4, pp. 30-38, 2015.

[22] Kwang-Seob, K., J. Ho-Gyeong, L. Hyun-Jong, and K. Hyung-jun, "Collaborative Filtering Using Topic Models for Rating Based Recommender Systems", paper presented at the Korea Information Science Society, 2012.

[23] Kweon Yang, K. and K. Chang Suk, "A String Kernel Based Sentiment Classification for Blog Text", paper presented at the Korean Institute of Intelligent Systems, 2009.

[24] Kyeonghee, S., L. Minsu, and O. Sangyoon, "Proceedings of Symposium of the Korean Institute of Communications and Information Sciences", paper presented at the Korea Institute of Communication Sciences, 2016.

[25] Landauer, T.K. and S.T. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge", *Psychological review,* Vol. 104 No. 2:211, 1997.

[26] Liu, B., "Sentiment Analysis and Subjectivity", *Handbook of natural language processing,* Vol. 2, pp. 627-666, 2010.

[27] Manuk, H., L. Sunghui, and L. Woohyoung, "A Study on the It R&Demerging Technology Detection through Information Analysismethod - Focus on Next Generation Computing Field -", paper presented at the Korea Institute of Industrial Engineers, 2009.

[28] Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", *arXiv preprint arXiv:1301.3781,* Vol., 2013.

[29] Narayanan, R., B. Liu, and A. Choudhary, "Sentiment Analysis of Conditional Sentences", paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, 2009.

[30] Ohana, B. and B. Tierney, "Sentiment Classification of Reviews Using Sentiwordnet", paper presented at the 9th. IT & T Conference, 2009.

[31] Sadamitsu, K., S. Sekine, and M. Yamamoto, "Sentiment Analysis Based on Probabilistic Models Using Inter-Sentence Information", paper presented at the LREC, 2008.

[32] Saggiona, H. and A. Funk, "Interpreting Sentiwordnet for Opinion Classification", paper presented at the Proceedings of the seventh conference on international language resources and evaluation LREC10, 2010.
[33] Sang-Do, K., P. Seong-Bae, P. Se-Young, L. Sang-Jo, and K. Kweon-Yang, "A Syllable Kernel Based Sentiment Classification for Movie Reviews", *Journal of Korean Institute of Intelligent Systems,* Vol. 20 No. 2, pp. 202-207, 2010.
[34] Seung Hoon, C., L. Jay Ick, and K. Juyoung, "A Comparative Analysis of Social Commerce and Open Market Using User Reviews in Korean Mobile Commerce", *Journal of Intelligence and Information Systems,* Vol. 21 No. 4, pp. 53-77, 2015.
[35] Sunghoon, S. and L. Hakyoun, "Fintech Trend Analysis Using Topic Modeling of Bm Patents", paper presented at the, 2015.
[36] Yoosin, K., K. Namgyu, and J. Seong Ryoul, "Stock-Index Invest Model Using News Big Data Opinion Mining", *Journal of Intelligence and Information Systems,* Vol. 18 No. 2, pp. 143-156, 2012.
[37] Yukun, H. and L. Seogjun, "The Distinctiveness of Big Data Business Model in Its Components: A Comparative Analysis of Korea-Us Cosmetic Big Data Business Cases", *Information Technology and Architecture,* Vol. 13 No. 1, pp. 63-75, 2016.

김성근
Sung Guen Kim received the B.A. degree in Psychology from Catholic university in 2009. He is a Master Candidate in Management Information System at Ajou University. His current research interests include Recommender system, Big data analysis and text mining.
e-mail: kimsungguen7@gmail.com

조혁준
Hyukjun Cho received the B.S. degree in Media and Business Administration in 2014. He is a Master Candidate in Management Information System at Ajou University. His current research interests include text mining, big data analysis.
e-mail: haha2432@gmail.com

강주영
Professor Juyoung Kang is currently a Full Professor of e-Business at School of Business, Ajou University. She received her Ph.D. in Management Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2005. She has more than 50 refereed publications in academic journals and conference proceedings and has developed Intelligent Systems and E-Commerce applications with various industrial partners. Her current research interests are in the fields of text mining, cloud computing, big data, and intelligent systems. and etc.
e-mail: jykang@ajou.ac.kr