지능정보연구 제18권 제3호 2012년 9월(pp.53~77)

텍스트 마이닝을 활용한 신문사에 따른 내용 및 논조 차이점 분석

감미아
연세대학교 문헌정보학대학원 석사과정
(makiyma@hanmail.net)

송 민
연세대학교 문과대학 문헌정보학과 부교수
(min.song@yonsei.ac.kr)

본 연구는 경향신문, 한겨레, 동아일보 세 개의 신문기사가 가지고 있는 내용 및 논조에 어떠한 차이가 있는지를 객관적인 데이터를 통해 제시하고자 시행되었다. 본 연구는 텍스트 마이닝 기법을 활용하여 신문기사의 키워드 단순빈도 분석과 Clustering, Classification 결과를 분석하여 제시하였으며, 경제, 문화 국제, 사회, 정치 및 사설 분야에서의 신문사 간 차이점을 분석하고자 하였다. 신문기사의 문단을 분석단위로 하여 각 신문사의 특성을 파악하였고, 키워드 네트워크로 키워드들 간의 관계를 시각화하여 신문사별 특성을 객관적으로 볼 수 있도록 제시하였다.

신문기사의 수집은 신문기사 데이터베이스 시스템인 KINDS에서 2008년부터 2012년까지 해당 주제로 주제어 검색을 하여 총 3,026개의 수집을 하였다. 수집된 신문기사들은 불용어 제거와 형태소 분석을 위해 Java로 구현된 Lucene Korean 모듈을 이용하여 자연어 처리를 하였다. 신문기사의 내용 및 논조를 파악하기 위해 경향신문, 한겨레, 동아일보가 정해진 기간 내에 일어난 특정 사건에 대해 언급하는 단어의 빈도 상위 10위를 제시하여 분석하였고, 키워드들 간 코사인 유사도를 분석하여 네트워크 지도를 만들었으며 단어들의 네트워크를 통해 Clustering 결과를 분석하였다. 신문사들마다의 논조를 확인하기 위해 Supervised Learning 기법을 활용하여 각각의 논조에 대해 분류하였으며, 마지막으로는 분류 성능 평가를 위해 정확률과 재현률, F-value를 측정하여 제시하였다.

본 연구를 통해 문화 전반, 경제 전반, 정치분야의 통합진보당 이슈에 대한 신문기사들에 전반적인 내용과 논조에 차이를 보이고 있음을 알 수 있었고, 사회분야의 4대강 사업에 대한 긍정-부정 논조에 차이가 있음을 발견할 수 있었다. 본 연구는 지금까지 연구되어왔던 한글 신문기사의 코딩 및 담화분석 방법에서 벗어나, 텍스트 마이닝 기법을 활용하여 다량의 데이터를 분석하였음에 의미가 있다. 향후 지속적인 연구를 통해 분류 성능을 보다 높인다면, 사람들이 뉴스를 접할 때 그 뉴스의 특정 논조 성향에 대해 우선적으로 파악하여 객관성을 유지한 채 정보에 접근할 수 있도록 도와주는 신뢰성 있는 툴을 만들 수 있을 것이라 기대한다.

논문접수일 : 2012년 08월 10일 논문수정일 : 2012년 09월 07일 게재확정일 : 2012년 09월 11일
투고유형 : 국문일반 교신저자 : 송 민

1. 서론

1.1 연구 배경

언론은 표현의 자유를 보장받고 있으며, 특히 신문의 경우 편집부의 영향이나 광고주의 영향을 받는 등의 특성을 가지고 있어 신문사마다의 논조 차이가 확연히 드러남은 한국 사회에서 공공연히 알려져 왔다. 과거부터 2012년 현재까지 부동으로 신문 구독률 1위에서 3위를 차지하고 있는 조선일보, 중앙일보, 동아일보는 구독률만 보아도 영향력이 얼마나 큰지를 짐작할 수 있다. 본 연구에서 초

점을 맞추고 있는 조선일보, 중앙일보, 동아일보의 경우, 한국 사회에서 세 신문사가 하나의 명사처럼 묶여서 쓰이기도 하는 등 보수 성향을 가진 신문사로서 자리매김을 한지 오래이고, 한겨레나 경향신문은 진보 성향을 가지고 있는 신문사로 유명하다. 구독하는 신문의 논조에 따라 사람들의 인식이 바뀌기도 하고, 역으로 구독자가 구독하는 신문에 따라 구독자의 정치적 성향을 판단하게 되는 것이 대부분이다. 요즘은 인쇄신문 구독률이 줄어들고 인터넷이 발달하여 인터넷신문을 활용하는 사람들이 늘어남에 따라, 포털 사이트에서 사람들이 뉴스를 접할 때 특정 주제를 가진 뉴스를 클릭하도록 유도하는 방법을 이용하여 쉽게 사람들의 인식 변화를 가져올 수 있게 되었다. 그렇기에 인쇄 신문만이 아닌 인터넷을 통해 접근하는 신문기사 논조의 중요성은 날로 증가하고 있다.

특히 신문사의 논조는 신문사별로 지지하는 특정 정부가 어떤 사회적인 이슈를 일으켰는지에 따라 달라짐을 지금까지 있어왔던 다양한 사례를 통해 확인할 수 있다. 한 예로, 4대강 사업에 대해 보수와 진보진영의 신문사들이 서로 다른 시각에서 접근했음은 구독자들 사이에서 이미 잘 알려져 있는 사실이다. 사실상 어떤 논조이든지 간에 무비판적으로 받아들이게 되면 글을 읽는 네티즌들의 생각은 점점 그 논조에 맞게 물들어갈 것이다. 네티즌들이나 신문 구독자들이 자신이 읽고 있는 신문의 논조와 성향을 파악하면서 읽는 것과, 무비판적으로 흡수하는 것의 영향력 차이는 크다. 각 언론이 가지고 있는 표현의 자유는 허락되어야 하지만, 그것을 받아들이는 사람들도 표현의 자유만큼 매스미디어를 자유롭게 바라볼 줄 알아야 한다. 그런 의미에서 각각의 신문이 어떤 논조를 가졌는지 혹은 어떤 위치에 있는지를 알고서 글을 읽을 수 있도록 돕는 새로운 툴이 절실한 시점이다.

1.2 관련 문헌 및 사례 분석

2010년 국정감사 기간 동안 이루어진 신문사들의 논조 분석에 대해 다루고 있는 기사를 발견할 수 있었다. 미디어투데이에서는 2010년 10월 4일부터 23일까지 전국단위 종합일간지 10곳의 1면에 실린 국정감사자료 기사를 분석했다. 분석결과, 신문의 이념적인 성향에 따라 정당 편향적인 국정감사보도가 이루어짐을 발견할 수 있었다. 자료의 출처가 보수인지 진보인지에 따라 각 여야정당에 대한 비판을 달리 했음을 알 수 있다. 경향신문이나 한겨레의 경우, 한나라당 의원의 자료는 1면에 싣지 않았으며 조선일보, 중앙일보, 동아일보에는 민주당에 대한 문서가 한 건도 없었다. 특히 보수신문의 경우 참여정부를 비판하기 위해 야당을 언급했을 뿐이었다. 또한 진보성향의 신문들은 현재 정부를 비판하기 위해 신문기사를 썼음을 알 수 있었다. 이 연구는 신문기사의 1면에서 언급된 기사의 횟수를 분석한 것으로, 단순빈도를 사용하였다.

보수와 진보의 성향을 뚜렷하게 가지고 있는 한국의 신문사들을 분석한 신문방송학 분야에서의 연구들은 예로부터 매우 활발하게 진행되어왔다. 특히 보수적인 성향을 대표하는 조선일보와, 진보성향을 대표하는 한겨레의 비교가 많았음을 알 수 있었다. 대부분의 연구들은 직접 텍스트를 하나하나 분석하고 코딩하여 연구결과를 도출해냈다. 정치인이 연루된 사건 분석(Kim and Chae, 2008)이나 호주제 폐지에 대한 논조 분석(Lee and Kim, 2006), 언론개혁에 관한 기사 담론 분석(Chung, 2002)을 통한 기사의 과장이나 축소 및 배제를 살펴본 연구 등 신문기사의 논조를 분석한 다양한 연구들이 있었으며, 연구 결과 해당 신문들의 보수와 진보의 성향이 뚜렷하게 드러남이 밝혀졌다.

앞선 연구들과 같이 정치 분야의 특정 주제에서

만 보수-진보의 이분법으로 논조를 나누는 것에 그치지 않고 '신문시장의 다양성과 주요 신문들의 보도 성향에 대한 비교 고찰'이라는 연구(Choi, 2011)에서 조선일보, 중앙일보, 동아일보, 한겨레신문, 경향신문, 한국일보 6개 주요 일간지를 대상으로 정치 이념(북한 핵과 대북 지원), 경제적 가치의 배분(종합부동산세 폐지), 사회·문화(호주제 폐지)의 세 가지 이슈에 대한 각 신문들의 이념을 분석했다. 최현주는 정치적인 입장뿐만 아니라 경제와 사회·문화적인 부분에서도 보수와 진보로 나뉘어지는지를 살펴보았다. 그 결과 6개 주요 일간지들은 각 분야에서 보수-진보의 이분법으로만 나뉘어지는 것이 아니라 스팩트럼에 따라 서로 다른 견해를 가지고 있었음을 밝혀내었다.

앞에서 제시한 국내의 신문방송학에서 진행된 연구들은 직접 기사를 분석하고 판단해내는 작업을 하였다. 그렇기에 보다 정확한 결과를 낼 수는 있었겠지만, 반면 연구자가 직접 코딩 작업을 하는 과정에서 연구자의 가치가 들어갈 위험성과 많은 분량의 데이터는 소화해내지 못한다는 단점이 있다. 텍스트 마이닝 기법을 활용한다면, 이러한 단점을 극복하여 더욱 다양한 분류가 가능해지고 각 분류에 따라서 신문사들을 다각도로 조명해볼 수 있다. Kim et al.(2012a)는 빅데이터의 오피니언 마이닝을 활용하여, 뉴스와 주가 사이의 관계를 통해 이용자들이 투자기회를 찾고 투자이익을 얻을 수 있는 툴을 개발한 연구다. 너무 많은 뉴스들이 실시간으로 생성되어, 뉴스가 주가에 미치는 영향을 알아보기 힘든 단점을 파악하여 이를 빅데이터 감성분석 기법을 활용하여 지능형 투자의사결정 모형을 제시하였다. 모형의 유효성을 검증하기 위하여 마이닝 결과와 주가지수 등락 간의 관계를 통계분석하여, 뉴스 컨텐츠의 감성분석 결과값과 주가지수 등락이 유의한 관계를 가지고 있음을 입증하였다. Choi et al.(2011)는 LED 분야의 특허들을 대상으로 텍스트 마이닝을 통해 중요한 기술정보를 추출한 다음, 키워드 네트워크를 구축하고, 이를 대상으로 커뮤니티 네트워크 분석을 수행하였다. 네트워크의 밀도와 클러스터링 지수를 살펴보았으며, 분석 결과 멱함수 분포를 따르고 있음을 발견하여 연구 및 기술의 발명이 이루어지는 절차를 확인하였다. 이 연구는 텍스트 마이닝 기법을 활용하여 특허의 키워드를 중심으로 네트워크를 시각화하였다는 데서 의의가 있다. 본 연구에서는 Choi et al.(2011)의 연구와 같은 방법을 활용하여 신문기사 키워드의 네트워크를 시각화하고, 더 나아가 각각의 네트워크가 가지고 있는 차이가 어느 정도의 신뢰성을 가지고 있는지도 Classification의 기법을 사용하여 함께 제시하고자 하였다. 또한 새로운 툴을 개발하여 다량의 문서를 분류하기 위해 다양한 분류기들을 활용한 국내 연구들 중, Naïve Bayes를 사용하여 분류 성능을 높인 Kim et al.(2012b)의 연구를 통해 한글 문서의 분류기 성능 향상의 가능성을 발견할 수 있었다. 기존 논문에서 나타난 개인정보의 유형들을 분석하여, 개인정보 관련 문서로 분류된 학습데이터에 영향력이 있는 개인정보 유형들을 추가 학습시켜 알고리즘이 학습하는 문서 자질의 질을 높이고자 하였다.

국내의 텍스트 마이닝을 활용한 다양한 분석 이외에도, 텍스트 마이닝을 활용하여 신문사들 간의 상반된 패턴을 알아보는 연구도 국외에서 다수 진행되었다. Balahur et al.(2009)은 Opinion analysis를 소개하였다. 주어진 주제에 대한 긍정이나 부정적인 표현을 기자, 독자, 기사의 수준으로 나누어 각각을 분석하였다. Pollak et al.(2011)은 Corpus를 활용한 담론분석과, bag of words, Decision tree를 사용하여 케냐의 선거에 대한 지역신문과 국제신문의 상반된 패턴과 성향을 분석해냈다.

앞선 다양한 관련 연구들에서 의미와 한계점을 발견함을 통해, 한국 신문사들의 분야별 논조에 따른 다차원적인 분석이 필요함을 알 수 있었다. 즉 본 연구는 정치적 이슈에 대한 보수-진보라는 이분법적이고 이차원적인 접근에서 벗어나서 다양한 분야에 따른 Classification 및 Clustering 분석 기법을 도입하여, 신문사별로 그리고 분야별로 내용과 논조를 비교해보고자 한다. 각 신문사가 해당 분야에 대해 어떤 내용을 주로 언급하는지 혹은 어떤 논조로 해당 주제를 구성하고 있는지 각 신문사의 논조 주관성을 파악하고 이를 객관적인 데이터로 제시하고 설명하여, 구독자들이 각 신문사의 기사들의 논조를 객관적으로 파악할 수 있도록 돕는 지표를 마련하고자 한다.

1.3 연구 질문

이번 연구에서 알아보고자 하는 연구문제는 다음과 같다.

"다량의 데이터를 사용한다면, 막연하게 알아오던 신문사별 성향과 논조 및 내용의 차이가 눈으로 확인이 가능한가"

그리고 세부 연구질문을 통해서 알아보고자 하는 것은 다음과 같다.

(1) 신문사별 키워드의 단순 빈도를 분석하면 신문사 간, 분야별 차이가 있는가?

(2) 각 신문사의 주요 키워드를 네트워크로 시각화한다면, 신문사별로 가지고 있는 논조와 내용이 보수-진보로 구분되던 기존의 분류와는 다르게 분류될 수 있을 것인가?

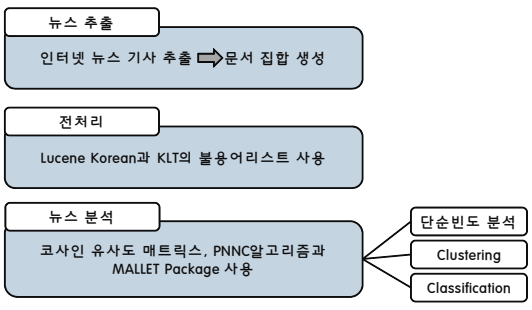
(3) 긍정-부정의 논조 판단이 필요한 이슈의 경우, 자동 분류를 시도했을 때 얼마나 의미있는 결과가 나오며 앞으로 어떻게 활용할 수 있을 것인가?

1.4 연구 목적

비교 대상은 진보진영의 동아일보, 보수진영의 한겨레신문, 경향신문 총 3개이며, 인터넷 상에 게재된 신문들을 텍스트 마이닝하여 각각 분야에 따른 논조가 어떻게 달라지는지를 살펴보았다. 각각의 분야는 정치, 사회, 경제, 국제, 문화의 5가지와, 신문사의 논조를 가장 잘 나타내주는 것으로 파악되는 사설을 살펴보아 총 6개의 분야를 분석했다. 신문사별로 각 분야에서는 어떤 내용을 주로 다루고 있는지, 그리고 각 신문사들 간의 논조의 차이가 있는지 혹은 비슷한 성향이 보이는지를 신문사별로 비교해보며 파악해보고자 한다.

2. 연구 설계

연구 설계는 다음과 같은 방법을 거쳤다. 우선 인터넷에 있는 뉴스를 추출하여 문서집합을 생성한 후, 전처리 과정을 거쳤다. 전처리에서는 Java로 구현된 Lucene Korean 모듈을 활용하였고, 그 모듈에 KLT(Korean Language Technology; 한글형태소분석기)의 불용어 리스트를 연동시켜 불용어를 제거하였다. 분석을 위해 신문기사 키워드들의 단순빈도를 분석하여 순위를 확인하였으며, 코사인 유사도 매트릭스와 PNNC 알고리즘을 활용



<Figure 1> The Overall Procedure

하여 군집화를 한 후 Mallet Package를 사용하여 Naïve Bayes 분류기로 분류를 하였다.

2.1 신문기사 수집

신문기사 데이터베이스 시스템인 KINDS (Korea Integrated News Database System, www.kinds.or.kr)에서 주제어 검색을 하여 해당 뉴스들을 수집하였다. KINDS에서 서비스되고 있는 동아일보, 한겨레, 경향신문을 대상으로 하였고, 2008년부터 2012년까지 해당 주제로 검색하여 수집한 신문기사들은 총 3,026개이다.

KINDS의 검색결과로 나온 신문기사의 해당 URL을 스파이더링(Spidering)하기 위해 HTML 문서로 저장한 후, 태그는 삭제하고 텍스트만 남겨 새로 저장하였다. 그리고 분류에 영향을 주지 않는 작성일자와 신문사, 기사 제목, 기자 이름 및 e-메일 주소는 삭제하여 저장하였다.

국제, 사회, 정치 분야는 각 분야별로 신문사 간 논조의 차이가 있을 것이라 여겨지는 키워드가 이슈화 된 시기를 기준으로 하여 한 달 내외로 수집하였다. 경제와 문화 분야의 경우 내용상의 차이를 분석하기 위해 검색 기간 한 달 내에 있는 모든 관련 뉴스를 추출하여 각 신문들이 어떤 부분을 주로 다루는지를 살펴보았으며, 사설도 마찬가지로 본 연구를 수행한 시기의 최근 한 달 간 기사들을 모두 수집하여 각 신문사들의 논조와 내용을 키워드 네트워크를 통해 분석하고자 하였다. 각각의 분야에 따른 검색 내용은 다음 <Table 1>과 같다.

사회 분야에서는 4대강 사업과 관련된 기사를 분석하였으며, 2008년 12월 29일 낙동강 지구에서 착공식이 진행된 시기를 중심으로 신문기사들을 검색하여 분석하였다. 정치 분야에서는 최근 통합진보당의 문제가 불거진 2012년 4월 전후의 뉴스를 수집하였다. 국제 분야는 2012년 3월 핵안보정상회의 개최일 근처 북한의 핵에 대한 기사를 분석하였다. 문화, 경제, 사설은 2012년 4월 25일부터 5월 25일 한 달 동안 신문기사화 된 전반적인 이슈와 논조에 대해 살펴보았다.

<Table 1> Retrieved Date, Keywords and the Number of News Articles

분야에 따른 검색 내용	
1. 사회	
검색 일자	2008.12. 29~2009. 01. 29
검색 내용	키워드 : 4대강 사업
신문기사 수	경향 : 42개 / 한겨레 : 31개 / 동아일보 : 15개
2. 정치	
검색 일자	2012. 04. 25~2012. 05. 25
검색 내용	키워드 : 통합진보당
신문기사 수	경향 : 218개 / 한겨레 : 181개 / 동아일보 : 191개
3. 경제	
검색 일자	2012. 04. 25~2012. 05 .25
검색 내용	검색 기간에 있는 모든 경제 관련 뉴스
신문기사 수	경향 : 226개 / 한겨레 : 263개 / 동아일보 : 633개
4. 국제	
검색 일자	2012. 03. 15~2012. 04. 15
검색 내용	키워드 : 북한 핵
신문기사 수	경향 : 68개 / 한겨레 : 71개 / 동아일보 : 114개
5. 문화	
검색 일자	2012. 04. 25~2012. 05. 25
검색 내용	검색 기간에 있는 모든 문화 관련 뉴스
신문기사 수	경향 : 174개 / 한겨레 : 173개 / 동아일보 : 210개
6. 사설	
검색 일자	2012. 04. 25~2012. 05. 25
검색 내용	검색 기간에 있는 모든 사설
신문기사 수	경향 : 75개 / 한겨레 : 54개 / 동아일보 : 58개

2.2 텍스트 데이터 처리

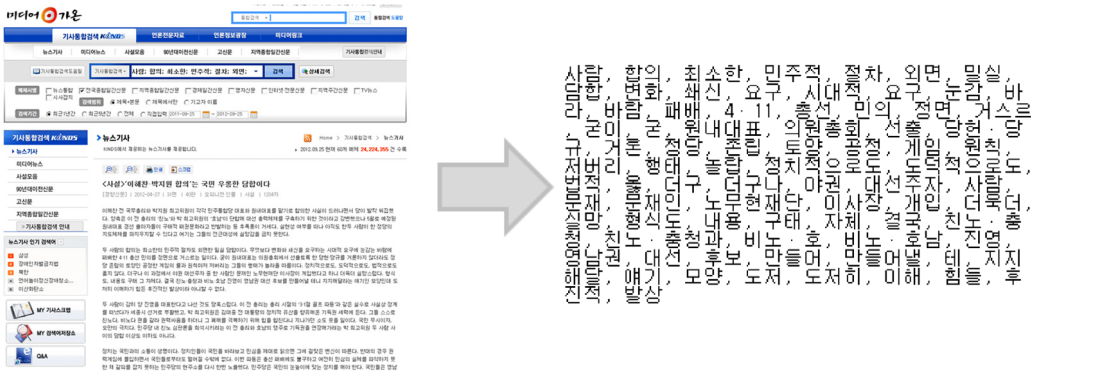
데이터 수집 후, 분석 시에 불필요한 데이터를 삭제하고 데이터의 형식을 통일했다. HTML, 스

크립트 언어 등을 삭제하여 기사의 순수 텍스트를 추출한 후, 불용어 제거와 형태소 분석을 위해 Java로 구현된 Lucene Korean 모듈을 이용하여 자연어 처리를 하였다. 모듈에 KLT(Korean Language Technology; 한글형태소분석기)의 불용어 리스트를 연동시켰으며, 그 외에도 텍스트를 분석하면서 발견되는 불용어를 처리하였다. 신문기사는 특정 단어를 나타내거나 문단을 나타낼 때 기호(■, ○, △, ▶ 등)를 빈번히 사용하는 특징이 있었기에 해당 기호들을 모두 삭제하였다. 또한 앞의 불용어 제거 및 형태소분석에서도 처리되지 못한 불필요한 단어들을 삭제하였다. 이런 과정을 거쳐 <Figure 2>와 같이 인터넷 상에 게재되어 있는 신문기사를 텍스트파일로 처리하고, 불용어를 제거한 후 의미 있는 단어만 선정하여 분석에 활용하였다. 분석하고자 하는 3,000여 개 신문기사 전체에 기재되어 있는 단어의 수가 많았기에, 본 연구의 특성을 고려하여 단어들을 보다 효과적으로 추출해야 했다. 본 연구는 신문사별로 차별화 되는 신문사의 특정 중요 단어를 밝혀내는 것뿐만이 아니라 신문사별 비슷한 내용이 어떻게 발견되는지를 분석하여 밝혀내고자 했으므로 가중치를 두지 않은 단순빈도를 이용하였다. 가장 빈번하게 나타나는 단어를 추출하여 각 신문사 별 차이점을 분석하고자 하였기에, 이와 같은 연구의 특성을 고려하여 단어 삭제 처리 및 동일한 단어 통합 절차를 거쳤다.

본 연구에서는 분석의 기본 단위를 신문기사가 아닌 문단으로 설정하였다. 파일은 한 기사 당 하나의 파일로 저장이 되었고, 기사 내의 문단으로 새롭게 파일을 저장했다. 이들은 Java 프로그램을 활용하여 처리하였으며, 각각 처리된 단어들은 쉼표(,)로 구분하고 문단 별로 구분된 파일은 단어들의 Clustering의 기준으로 활용되거나 Classification에서의 학습과 테스트 데이터로 사용되었다.

2.3 데이터 분석 방법

신문기사의 내용 및 논조를 파악하기 위해, 경향신문, 한겨레, 동아일보가 정해진 기간 내에 일어난 특정 사건에 대해 언급하는 단어들의 빈도를 측정하여 빈도 순위 상위 10위 안에 드는 단어들을 제시한 후 신문사 간의 차이점을 분석하였다. 각 신문기사의 문단에 있는 코사인 유사도를 활용한 co-word 분석을 하여 네트워크 지도를 만들었고, 단어들의 네트워크를 통해 Clustering 결과를 분석하였다. 마지막으로 신문사들마다의 논조를 확인하기 위해 Supervised Learning 기법을 활용



<Figure 2> Modification of Online News Articles into Texts

하여 각각의 논조에 대해 분류하였다. 논조는 기존 연구들에서 사용해왔던 보수-진보를 활용해 특정 이슈에 대한 긍정-부정의 수준으로 나누어 분석해 보았다.

해당 사항들을 분석하기 위해 다음 <Table 2>와 같이 총 6개의 분야를 3가지 범주로 나누어 분석을 시도하였다.

<Table 2> Different Analysis Method of Issues

이슈의 특성	해당 분야의 전반적인 내용 파악용	논조 판단이 필요 없는 경우	논조 판단이 필요한 경우
뉴스 분야 (키워드)	경제(전반) 문화(전반) 사설(전반)	국제(북한 핵) 정치(통합진보당)	사회(4대강)
분석 방법	단순빈도 Clustering	단순빈도 Clustering	단순빈도 Clustering Classification

표와 같이 전반적인 내용 및 논조를 파악하기 위해 대상으로 삼은 경제 전반, 문화 전반, 사설 전반에 대해서는 단순빈도분석과 Clustering을 하여 신문사가 주로 다루고 있는 내용을 살펴보았다. 북한 핵이나 통합진보당에 관해서는, 가치 판단이 크게 필요 없이 신문사들 모두가 부정적인 목소리를 내고 있으므로 논조에 대한 판단은 의미가 없다고 가정하여 단순빈도분석과 Clustering 분석을 하였다. 마지막으로 신문사들끼리 서로 다른 논조를 가지고 있다고 알려져 있는 4대강 사업과 관련한 신문기사는, 신문사들 간 확연한 논조의 차이를 객관적인 데이터로 밝혀내기 위해 단순빈도와 Clustering 뿐만이 아닌 긍정-부정 Classification 기법을 활용하여 확인해보고자 하였다. 즉 Classification을 통해 긍정인지, 부정인지를 분류하여 해당 신문들의 긍정-부정 성향을 도표를 통해 파악해보고자 하였다.

2.3.1 단순빈도 분석

특정 신문사 내에서 등장한 키워드들의 빈도를 분석하였다. 단순하게 해당 신문사가 다루고 있는 특정 주제분야에서 자주 등장한 단어의 순위를 매겨 상위 10위를 분석해보았다. 다른 신문사와의 순위 비교를 통해 신문사들은 어떤 논조나 내용을 주로 나타내고 있는지를 분석하고자 하였다. 해당 주제와 직접적으로 관련이 있는 단어, 즉 모든 신문사에서 높은 순위로 함께 등장하는 단어가 1순위에 있는 경우 결과 해석에 의미가 없다고 판단하여 삭제한 후 다시 순위를 매겨 분석하였다. 예를 들어 북한 핵에 대한 신문기사를 분석할 시 '북한'이라는 단어가 1순위에 있고 '핵'이라는 단어가 2순위에 있을 경우, 1순위와 2순위를 제외한 나머지를 대상으로 순위를 매겨 분석하였다.

2.3.2 Clustering

Clustering으로 키워드를 분석하기 위해 코사인 유사도 매트릭스를 활용하여 키워드들의 정방대칭행렬을 만들고 패스파인더 네트워크를 생성하였다. 이 네트워크는 단어들끼리의 문단 내 동시출현빈도를 활용하여 코사인 유사도를 산출한 것이다. 출현 빈도가 100위 내외로 드는 단어들을 중심으로 네트워크 분석을 했으며, 그 중 삼각매개중심성(Triangle Betweenness Centrality; 이하 CTB, Lee, 2006a)이 높은 경우 노드의 크기를 확대시켜 강조하였다. CTB의 경우는, 한 노드가 다른 노드들 사이를 결속시켜주는 능력을 측정하는 것으로, 주변에 있는 다른 단어들을 연결시켜주는 빈도가 높아질수록 중요한 역할을 한다고 파악하여 노드의 크기에 반영을 하였다.

네트워크 시각화에는 NodeXL 프로그램을 활용하였다. Harel-Koren Fast Multiscale을 사용하였

으며, 각 값에 가중치 부여를 하여 링크 굵기를 설정해주었다. PNNC(최근접 이웃 클러스터링) 알고리즘(Lee, 2006b)에 의해 군집화가 나타나면 서로 다른 색을 부여해 각 집단을 구분해주었다. PNNC는, 모든 노드를 크기가 1인 군집으로 간주하여 각 군집의 최근접 이웃 군집을 찾는 기법이다. 최근접 이웃이 가장 가까운 군집으로부터 시작하여 모든 군집을 각자의 최근접 이웃 군집과 연결되도록 하며, 군집이 하나로 결집되지 않았으면 앞의 단계를 반복하여 군집을 형성하는 기법으로, 코사인 유사도로 측정하여 네트워크를 그렸을 경우 군집을 가시적으로 잘 나타내 보여주기에 코사인 유사도를 활용하였다.

2.3.3 Classification

Java로 구현되어 있는 MALLET Package를 사용하여 Classification을 진행하였으며, 앞서 Clustering한 결과가 실제로 각 신문사별 논조와는 얼마나 차이가 있는지를 확인해보고자 하였다. Naïve Bayes 분류기를 사용하여 테스트 결과를 분류하였다. Naïve Bayes 분류기는 베이즈 정리(Bayes' theorem)에 근거한 확률적 분류기로서, 학습문헌을 이용하여 각 단어가 특정 범주를 대표할 확률을 계산한 다음 분류할 입력문헌에 출현한 단어들을 단서어로 하여 이 문헌의 범주를 예측하는 기법이다(Chung, 2005). 그 외의 분류 기법으로는 Maximum Entropy나 Decision Tree, 신경망 등의 기법도 있는데, 그 중 Naïve Bayes 분류기를 쓴 이유는 Baek(2003)의 연구에 기반하였다. 한글 인터넷 뉴스 기사의 자동분류시스템에 관해 연구한 백용규에 따르면, 여러 분류기법을 썼을 경우 200개 이하로는 비슷하나 200개 이상의 단어인 경우에 베이지언 네트워크 분류자가 다른 분류자에 비해 분류율이 높았음(200단어의 경우 62%의 분류율)을 확인했기 때문이다.

이와 같이 분류한 것을 바탕으로, 각 분야에 따라 신문 별로 분류가 되는지를 살펴보기 위해 정확률(Precision)과 재현률(Recall), F-value로 분류성능을 평가해 보았다. Classification을 하기 위해서는, 학습된 데이터가 필요하다. Naïve Bayes 분류기로 데이터를 학습시킨 후에, 테스트를 하는 데이터를 투입하여 그 성능이 어느 정도인지를 판단하였다. 즉 경향신문의 일부 문단들을 분류기 상에 경향신문으로 학습시킨 후, 나머지 경향신문의 문단들을 학습된 데이터들을 기반으로 하여 테스트해보아 테스트한 데이터들이 어느 정도의 확률을 가지고 경향신문으로 분류가 되는지를 확인해보았다. 경향신문의 테스트 데이터를 넣었을 경우 경향신문으로 분류되면 제대로 분류된 것이라 하여 True positive 영역에 넣었고, 한겨레나 동아일보로 잘못 분류된 경우는 False negative 영역으로 두어 얼마나 제대로 분류가 되었는지를 확률상 계산해보았다. 전체 경향신문 기사 100개 중 경향신문으로 분류된 것이 50개라면 재현률이 0.5가 되고, 나머지 한겨레와 동아일보를 모두 학습시키고 테스트했을 경우 전체 기사들 중 경향신문으로 분류된 것이 150개가 있을 때 이 중 실제로 경향신문 기사인 것이 50개가 있다면 정확률은 0.33이 된다. 즉 테스트 결과 올바르게 나올 확률을 알아보는 것이다. F-value는 다음의 식과 같이 계산하며 이는 정확률과 재현률 둘의 특성을 조합한 신뢰도를 측정해주기 위해 제시하였다.

$$\text{F-value} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Classification을 위한 학습데이터와 테스트 데이터는 모두 신문기사 내의 문단을 기본 단위로

하여 분석하였으며, 각각을 학습하고 테스트하기 위한 문단들의 수는 다음의 <Table 3>과 같다. 학습데이터와 테스트 데이터의 비율은 70:30이 되는 것이 적당하므로(Carlos and Lucio, 2003), 이를 유념하여 문단의 수를 지정하였다. 또한 학습데이터의 경우 신문사별로 학습의 수를 통일시켜 지정하기 위해 문단의 수가 많은 신문사라도 문단이 적은 신문사에 맞추어 학습 데이터를 저장하였다. 즉 <Table 3>과 같이 신문기사를 문단으로 나누었고, 경제의 경우를 예로 들면 세 신문사 모두 학습 문단의 수를 750개씩 넣었고 테스트 문단의 수를 150개로 맞추어 비교해본 후 결과를 분석하였다.

<Table 3> The Number of Learning and Test Data(Paragraphs)

분야	학습 데이터 (문단)의 수	테스트 데이터 (문단)의 수
경제(전반)	750	150
문화(전반)	1400	150
국제(북한 핵)	350	100
정치(통합진보당)	150	50
사회(4대강 사업)	2000	150
사설(전반)	150	50

정확률이나 재현률, F-value가 0.6 이상의 분류 성능을 보이는 경우에는 의미가 있는 분류 결과라고 보았으며 이를 해석하였다. 반대로 0.25 이하의 낮은 분류 성능을 가진 경우에는 이를 다른 신문사들과의 논조 차이가 거의 없다고 판단하였고, 이 또한 역으로 의미가 있다고 판단하여 해석을 시도하였다. 즉 낮은 분류 성능을 가진 경우 해당 신문만의 논조나 내용 상 특성이 나타나지 않는다고 해석하였다. 또한 신문사별로 특정 주제(4대강 사업)에 대해 긍정-부정 성향이 어떻게 다르게 나타나는지를 분석하는 데에도 Classification이 사용되었다.

3. 데이터 분석 결과

앞서 설명한 단순빈도, Clustering, Classification의 3가지 분석 방법을 활용하여 경제, 문화, 국제, 사회, 정치, 사설 총 6개의 분야에 관해 분석해보았다. Clustering 분석에서는 경제, 문화, 사설, 국제, 정치, 사회의 6개 분야로 나누어 각 분야를 군집화 하였다. 군집화된 신문사별 특성이 어느 정도 신뢰성이 있는지를 측정하기 위해, 실제 텍스트 마이닝 기법을 활용하여 신문사별로 논조에 따라 다르게 분류가 될 확률을 객관적인 데이터로 보여주었다. 각 분야별로 신문사별 논조의 차이를 Mallet Package를 사용하여 2.3.3의 설명과 같이 Naïve Bayes 분류기로 분류를 하였다.

3.1 경제(전반)

동아일보는 경제 분야에서 기업에 대해 가장 높은 빈도로 언급함을 알 수 있었고, 그에 반해 경향신문은 시장, 한겨레는 정부의 경제에 대해 가장 높은 빈도로 언급하였다. 또한 한겨레나 경향신문과는 달리 동아일보는 투자, 상품에 관련된 기사를 높은 빈도로 다룬다면, 한겨레와 경향신문은 삼성

<Table 4> Top 10 Keywords of Three Newspapers in Eco-business Section

순위	경향신문	한겨레	동아일보
1	시장	정부	기업
2	삼성전자	시장	국내
3	정부	제품	한국
4	관계자	삼성전자	시장
5	제품	회장	미국
6	국내	가격	정부
7	한국	미국	제품
8	사업	서울	투자
9	기업	국내	고객
10	미국	대기업	상품

전자에 관한 기사를 자주 다루고 있다고 해석할 수 있다.

<Figure 3>과 같이 Clustering을 해보았을 때 경향신문은 대체적으로 4개의 클러스터로 나누어졌고, 기업 전반, 세계로의 진출 확대, 은행, 공급 및 분양을 크게 다루고 있으며, 한겨레는 6개의 클러스터로 하이마트, Toyota, KT, 삼성전자와 Apple의 스마트폰 경쟁, 정부의 부동산 및 주택 가격 문제, 기업들의 수출 등을 주로 다루고 있음을 알 수 있다. 동아일보는 크게 4개의 클러스터로 나누어졌고 가격의 상승이나 매출의 증가, 삼성전자, 현대차, Apple 등과 같은 대기업이나 중소기업, 정부의 지원, 고객의 투자 상품, 인수와 매각에 관한 기사를 주로 다루고 있음을 알 수 있다.

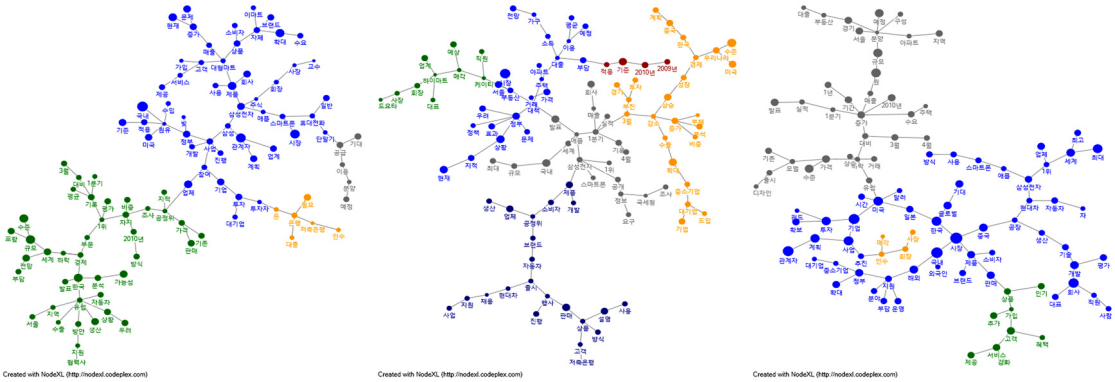
신문사별 경제 전반에 대한 논조 및 내용의 차이를 Classification 기법을 통해 검증해본 것은 다음의 <Table 5>와 같이 나타낼 수 있다. 앞서 설명하였듯이 각 신문사의 신문기사들을 신문사별로 학습시키고, 테스트를 한 기사들이 자신의 신문사에 올바르게 분류될 확률을 정확률과 재현률, F-value를 사용하여 분석해보았다. 경제 전반에 관한 것은, 전체 평균적으로 0.47 정도의 확률을 보이고 있어 각 신문사별로 논조 및 내용에 크게 차이가 나지 않는다고 판단할 수 있다. 이는 세 신문사의 경제에 관한 신문기사에서는, 내용상 각자만의 특성이 크게 드러나지 않음을 시사한다.

<Table 5> Precisions, Recall and F-value of Three Newspapers in Eco-business Section

	정확률	재현률	F-value
경향일보	0.479339	0.386667	0.428044
한겨레	0.460674	0.546667	0.5
동아일보	0.483444	0.486667	0.48505

3.2 문화(전반)

문화에 대한 기사에서는, 경향신문은 삶에 대해, 한겨레는 학생에 대해, 동아일보는 작품에 대해 주로 다루고 있음을 알 수 있다. 즉 경향신문은 사람의 삶과 인간에 초점을 맞춘 기사를 주로 쓰고, 동아일보는 서울 내에 있는 여러 공연에 관한 정보를 제공하는 기사를 주로 다루며, 한겨레는 아이들이나 학생들이 공부를 할 때 유용한 정보를 제공함을 파악할 수 있었다.



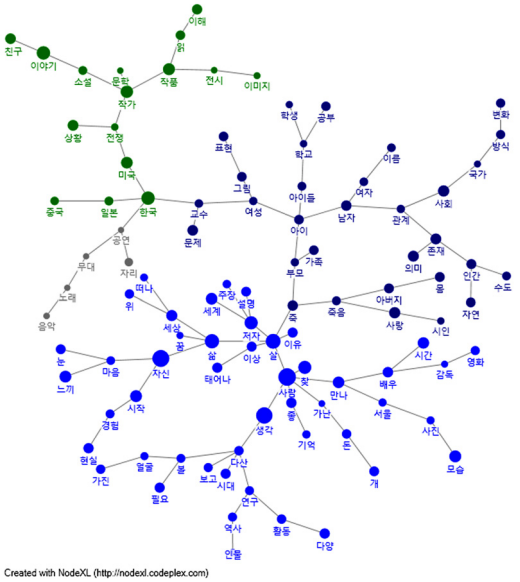
<Figure 3> Keyword-Network of Eco-business Section of Kyunghyang(Left), HanKyoreh(Middle) and Dong-A(Right)

<Table 6> Top 10 Keywords of Three Newspapers in Culture Section

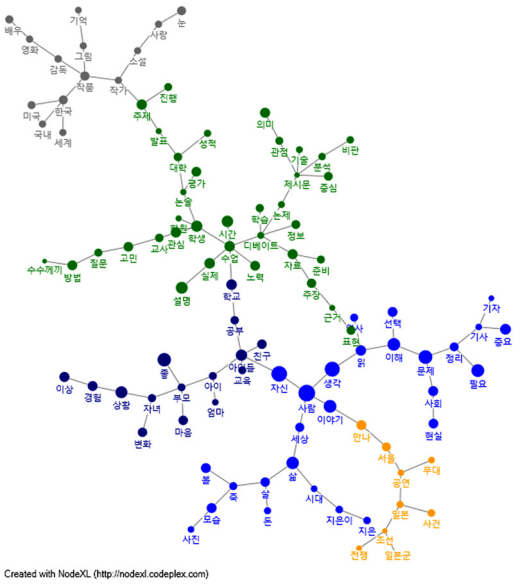
순위	경향신문	한겨레	동아일보
1	사람	사람	사람
2	삶	학생	작품
3	다산	삶	서울
4	자신	생각	한국
5	생각	문제	학생
6	살	자신	공연
7	저자	공부	미국
8	한국	감독	자신
9	죽	영화	교수
10	미국	아이들	저자

Clustering 결과를 확인하였을 때도 빈도 분석과 비슷하였다. <Figure 4>를 보면 경향신문의 경우 4개의 클러스터로 나누어져 다양한 역사나 시대에 대한 기사를 다루고 있으며, <Figure 5>에서 한겨레의 경우 주로 수업이나 대학의 논술, 평가, 교사들의 관심사나 자녀교육과 같은 기사를 다루고 있음을 알 수 있다. <Figure 6>을 분석하면 동아일보는 주로 서울에서 볼 수 있는 각종 공연에 대해 상세히 다루는 것으로 파악되어, 교육 분야에 대한 정보를 주로 다룬 신문을 보고자 할 때는 한겨레를, 인간의 삶에 대해 조명한 신문을 보고자 할 때는 경향신문을, 서울의 각종 공연에 관한 정보를 얻고 싶을 때는 동아일보를 확인하면 되는 것으로 분석이 되었다.

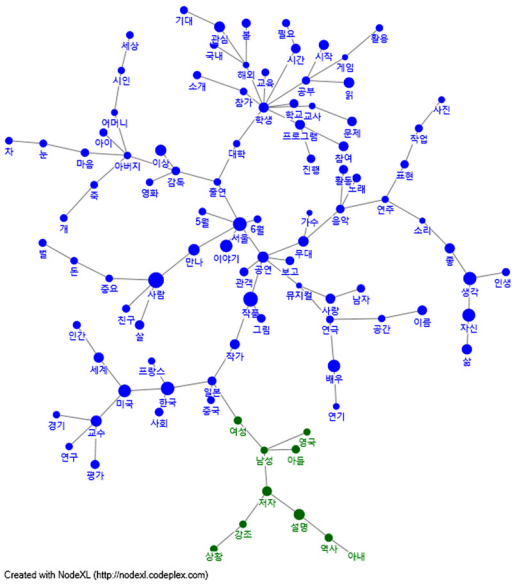
<Table 7>에서는, 경향신문의 재현률이 0.7 이상, F-value가 0.5 이상을 보이고 있어서 경향신문이 다른 두 신문과 비교했을 때 현격한 수치상 차이를 보여 논조나 내용상 뚜렷한 특성을 가지고 있는 것으로 나타난다. 하지만 한겨레의 경우, 재현률이 0.18이고 F-value도 0.23 정도를 나타내고 있는데, 이는 분류기에 의해 한겨레로 분류가 잘 되지 않음을 뜻한다. 즉 한겨레는 문화 전반에 대



<Figure 4> Keyword-Network of Culture Section of Kyunghyang



<Figure 5> Keyword-Network of Culture Section of HanKyoreh



<Figure 6> Keyword-Network of Culture Section of Dong-A

해 한겨레만의 뚜렷한 특성을 가지고 있지 않다고 분석할 수 있다. 그에 비해 경향일보는 다른 두 신문에 비해 문화 분야에서 특성 있는 기사를 쓰고 있음을 분류 성능 평가를 통해 발견할 수 있다.

<Table 7> Precisions, Recall and F-value of Three Newspapers in Culture Section

	정확률	재현률	F-value
경향일보	0.401515	0.706667	0.512077
한겨레	0.346154	0.18	0.236842
동아일보	0.425926	0.306667	0.356589

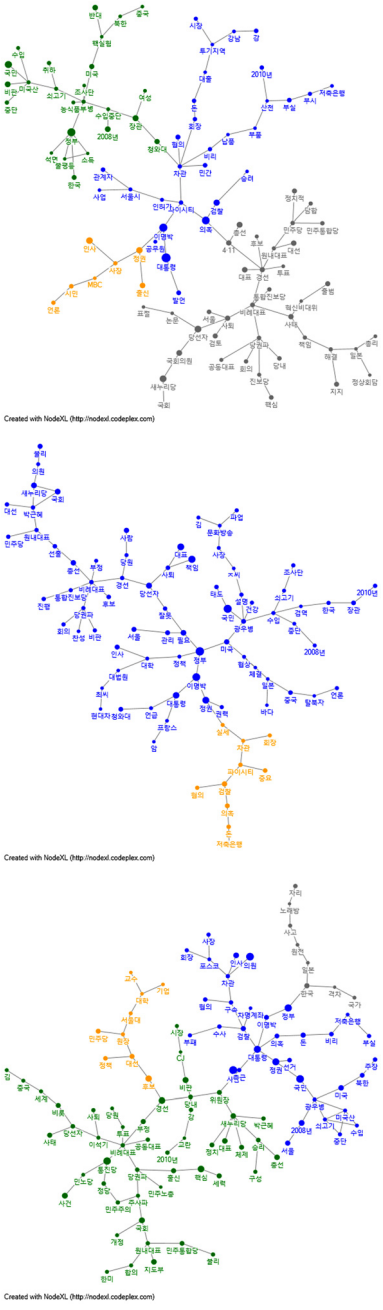
3.3 사설(전반)

사설 전반에 대한 분석을 통해, 2012년 4월과 5월 사이 있었던 사회 전반적인 이슈들을 살펴볼 수 있었다. 경향신문과 한겨레는 검찰과 광우병, 미국을 높은 빈도로 다루고 있었다. 동아일보는 통합진보당의 사건을 사설에서 주로 다루고 있으며, 북한이나 민주당에 대한 언급 빈도도 다른 두 신문사에 비해 높았음을 알 수 있다.

<Table 8> Top 10 Keywords of Three Newspapers in Editorial-opinion Section

순위	경향신문	한겨레	동아일보
1	대통령	정부	대통령
2	정부	미국	통진당
3	검찰	대통령	북한
4	광우병	국민	정부
5	미국	검찰	경선
6	당권파	광우병	미국
7	의혹	국회	한국
8	국민	비례대표	국민
9	당선자	당선자	비례대표
10	파이시티	경선	민주당

<Figure 7>과 같이, 주로 다루고 있는 내용적인 측면에서 세 신문이 모두 비슷하게 구성되어 있음을 알 수 있었고, 이 부분은 각각 신문사의 특성을 파악한다기 보다는 최근의 이슈를 한 눈에 파악할 수 있는 자료로서 활용할 수 있을 것이라 보았다. 경향신문의 경우에는, 파이시티를 중심으로 이명박 대통령과 서울시 인허가에 대해 다루고 있으며 소득 불평등, 미국산 쇠고기, 북한 핵실험 반대에 대해 논하고 있다. 그 외 MBC의 파업에 대한 언급과, 크게는 통합진보당의 사태에 대해 상세히 다루고 있음을 알 수 있다. 한겨레도 경향신문과 비슷하게 정부의 정책이나 미국, 광우병, MBC 사장, 통합진보당의 부정행위, 파이시티에 대해 주로 다루고 있음을 알 수 있다. 동아일보는 가장 크게 통합진보당의 사건에 대해 언급하고 있으며 그 외에 민주당, 서울대, 새누리당, 대통령 의혹, 검찰의 부패 척결 및 구속, 광우병 등에 대해 다루고 있었다.



<Figure 7> Keyword-Network of Editorial-opinion Section of Kyunghyang(Left), HanKyoreh(Middle) and Dong-A(Right)

<Table 9>에서 보이듯이 사설 분야는 높지 않은 분류 성능을 보여준다. 동아일보가 0.59로 0.6과 비슷한 수치를 보여 다소 높은 확률로 다른 신문사에 비해 어느 정도 특성을 나타내고 있음을 알 수 있지만, 나머지는 확률상 현저하게 분류되지 않아 신문사별로 사설의 내용 및 논조에 큰 차이가 없음을 확인할 수 있었다.

<Table 9> Precisions, Recall and F-value of Three Newspapers in Editorial-opinion Section

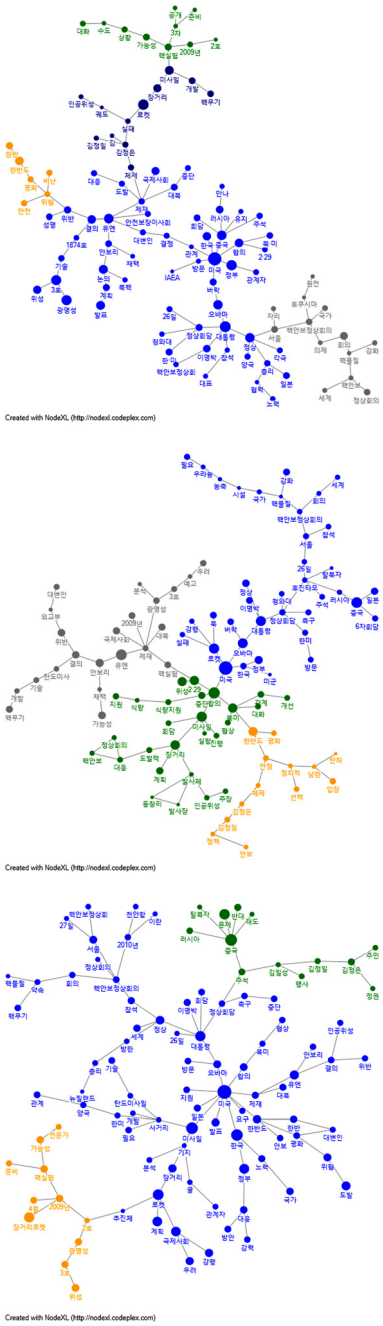
	정확률	재현률	F-value
경향일보	0.384615	0.5	0.434783
한겨레	0.511628	0.44	0.473118
동아일보	0.595238	0.5	0.543478

3.4 국제(북한 핵실험)

북한 핵에 대한 키워드로 분석해본 결과, 동아일보가 해당 신문 기사를 가장 많이 다루고 있었다. 핵안보정상회의가 개최되는 시기를 근처로 분석했으므로 세 가지 신문사 모두 미국을 중심으로 한 신문기사 내용을 담고 있음을 알 수 있고, 그 외에 중국과 대통령이 중요한 역할을 함을 알 수

<Table 10> Top 10 Keywords of Three Newspapers in International Section

순위	경향신문	한겨레	동아일보
1	미국	미국	미국
2	대통령	대통령	중국
3	로켓	중국	한국
4	중국	합의	대통령
5	오바마	로켓	미사일
6	3호	오바마	로켓
7	광명성	미사일	김정은
8	핵실험	유엔	문제
9	정부	위성	정상
10	합의	장거리	서울



<Figure 8> Keyword-Network of Nuclear Weapon of North Korea in Kyunghyang(Left), HanKyoreh(Middle) and Dong-A(Right)

있다. 경향신문과 한겨레는 합의에 대해 높은 빈도로 언급하는가 하면, 동아일보는 다른 두 신문사에서 높게 언급하지 않는 김정은에 관한 기사를 많이 제공하고 있음을 알 수 있다.

<Figure 8>의 Clustering 결과를 분석해본 결과, 경향신문은 미국과 유엔의 입장에 대해 주로 다루고 있었고, 로켓의 실패와 핵실험의 가능성을 보도하였다. 한겨레의 경우 미국과 오바마 대통령이 네트워크의 중심에서 다른 노드들을 연결해주고 있어 기사의 내용상 중요한 역할을 한다고 판단할 수 있다. 동아일보의 경우에도 미국을 중심으로 북한 핵 이슈에 대해 논하고 있음을 알 수 있고, 다른 신문사와 차별적으로 중국, 탈북자나 러시아와의 관계에 대해 주로 언급하고 있음을 발견할 수 있었다.

전반적으로 국제적인 분야나 핵안보정상회의, 북한 핵에 대해 다루는 내용이나 논조가 신문사별로 비슷하기에 내용이나 논조 상 큰 특성을 보이지는 않음을 <Table 11>에서 확인할 수 있다.

<Table 11> Precisions, Recall and F-value of Three Newspapers in International Section

	정확률	재현률	F-value
경향일보	0.430233	0.37	0.397849
한겨레	0.5125	0.41	0.455556
동아일보	0.395522	0.53	0.452991

3.5 정치(통합진보당)

통합진보당에 대한 신문기사를 분석해본 결과 세 신문사 모두 대표, 당선자, 당원, 사퇴, 경선에 대해서 언급하고 있었다. 순위가 하위로 내려갈수록 차이가 생겼는데, 동아일보에서는 다른 두 신문과는 달리 국회와 북한이라는 단어가 빈번하게 나타냈으며 경향신문에서는 다른 신문사와는 다르

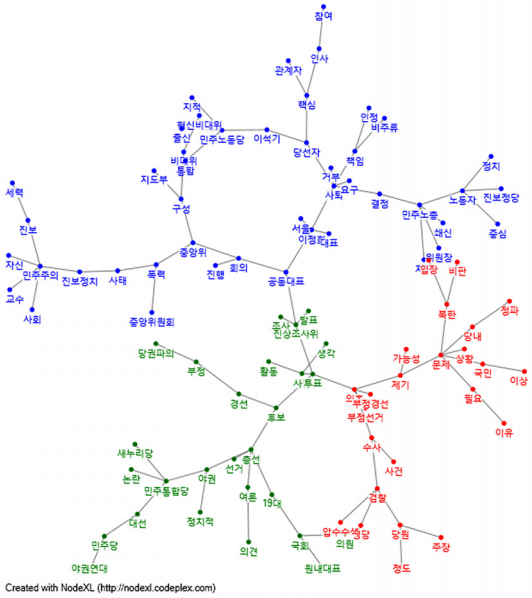
게 검찰에 대해 높은 빈도로 언급함을 알 수 있었다. 한겨레는 국민이라는 단어를 다른 신문사들에 비해 비중 높게 언급하고 있었다.

<Table 12> Top 10 Keywords of Three Newspapers in Politics Section

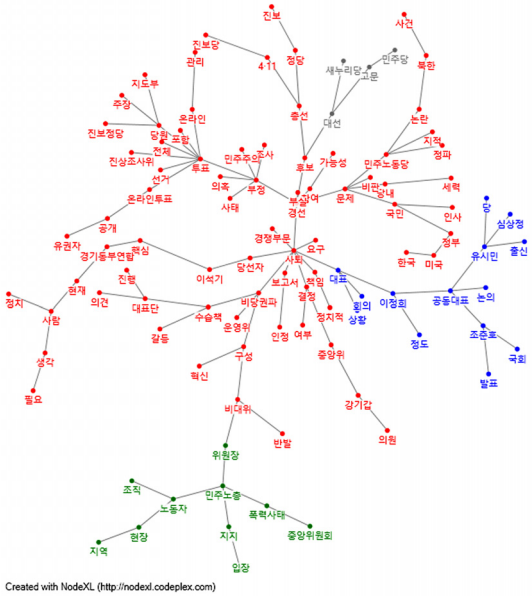
순위	경향신문	한겨레	동아일보
1	당선자	대표	당선자
2	문제	투표	대표
3	당원	당원	경선
4	후보	문제	후보
5	사태	경선	당원
6	사람	후보	사퇴
7	사퇴	사퇴	국회
8	경선	부정	부정
9	총선	당선자	의원
10	검찰	국민	북한

경향신문은 <Figure 9>와 같이 크게 세 그룹으로 나뉘었으며, 주로 민주노동당의 사퇴나 비대위 통합, 중앙위, 폭력, 부정선거의 문제와 수사, 총선, 선거, 민주통합당, 새누리당에 대해 언급하고 있음을 알 수 있다. 한겨레는 <Figure 10>과 같이 총 4개의 그룹으로 나누어졌고, 경선에서의 부정행위와 문제점을 비판하고 있으며 관련된 인물들의 이름이나 경기동부연합에 대해 언급하고 있음을 알 수 있다. 또한 민주노총의 폭력사태나 새누리당, 대선에 대해서 언급하고 있다. 동아일보에는 <Figure 11>과 같이 총 3개의 그룹으로 나누어지고, 크게는 경선에서 후보의 부정에 대한 의혹 제기와 당선자에 대한 비판, 경기동부연합에 대한 언급을 하고 있다. 또한 검찰의 압수수색에 대한 자료 등이 다른 두 신문사에 비해 많이 차지를 하고 있고, 중앙위원회와 폭력사태, 그리고 공동대표에 대한 조사에 대해 언급함을 알 수 있었다.

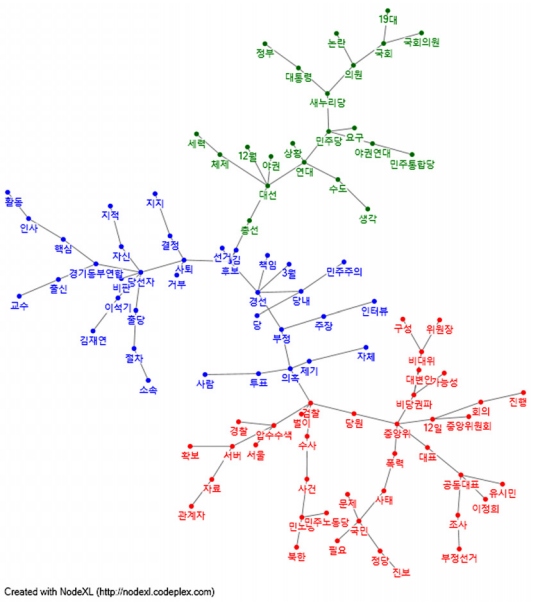
<Table 13>을 살펴보면, 통합진보당에 대한 내



<Figure 9> Keyword-Network of Tonghap-Jinbo Dang in Kyunghyang



<Figure 10> Keyword-Network of Tonghap-Jinbo Dang in HanKyoreh



<Figure 11> Keyword-Network of Tonghap-Jinbo Dang in Dong-A

용은 모든 신문이 정확률 부분에서 0.6 이상의 분류 성능을 보이고 있음을 알 수 있다. 특히 경향신문의 정확률은 0.84이고 F-value는 0.81, 한겨레의 재현률은 0.8로 매우 높은 수준의 분류 성능을 보여주고 있다. 이를 통해 통합진보당에 대한 각 신문사들의 내용이나 논조는 세 신문 간에 확연히 차이가 난다고 해석할 수 있다. 즉 통합진보당과 관련한 이슈에 대해서는 세 신문사가 각각 논조 및 내용에 각자만의 특색을 가지고 있음을 객관적인 수치를 통해 발견해낼 수 있었다.

<Table 13> Precisions, Recall and F-value of Three Newspapers in Politics Section

	정확률	재현률	F-value
경향일보	0.847826	0.78	0.8125
한겨레	0.655738	0.8	0.720721
동아일보	0.627907	0.54	0.580645

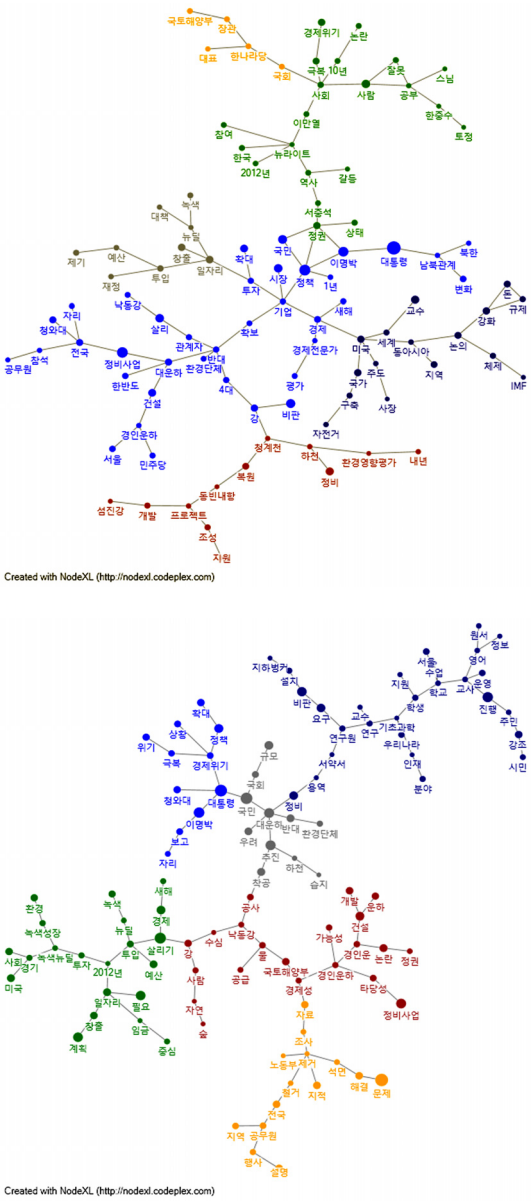
3.6 사회(4대강 사업)

4대강 사업에 대한 기사들을 분석해본 결과, 단어 빈도 순위에 있어서 세 신문사 간 확연한 차이를 보임을 알 수 있었다. 동아일보는 살리기, 낙동강, 일자리, 녹색, 자전거 등과 같이 4대강에 대해 긍정적인 단어들을 많이 사용하여 기사를 제공하고 있음을 알 수 있다. 반면 동아일보에서는 빈도수 10위 안에 들지 않은 단어 '대통령'이 경향신문과 한겨레에서는 빈도가 가장 높다는 차이점을 보여, 경향신문과 한겨레는 주로 대통령과 4대강을 연결하여 언급하고 있음을 알 수 있다.

<Table 14> Top 10 Keywords of Three Newspapers in National Issues Section

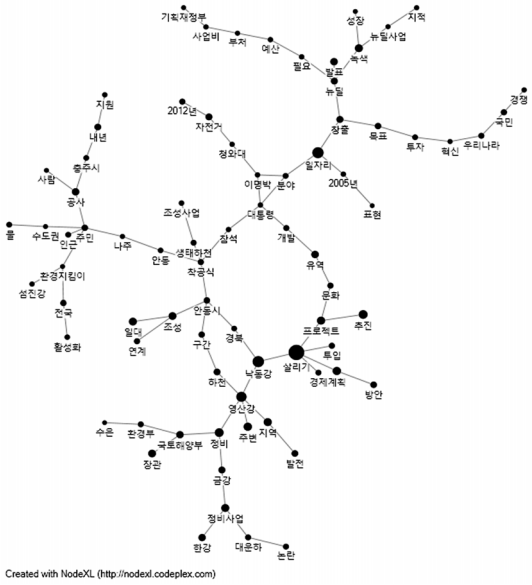
순위	경향신문	한겨레	동아일보
1	대통령	대통령	살리기
2	국민	영어	낙동강
3	교수	일자리	지역
4	정책	국민	일자리
5	일자리	대운하	녹색
6	이명박	이명박	영산강
7	녹색	녹색	물
8	정비사업	석면	자전거
9	경제	청와대	주민
10	시장	건설	계획

<Figure 12>의 Clustering 결과를 분석해보면, 경향신문은 6개의 클러스터로 나누어지며, 이명박 대통령의 정책에 대해 중점적으로 다루고 있다. 4대강 사업이 가져올 수 있는 결과 혹은 투자대비 효과를 논하고 있음을 추측할 수 있다. 환경단체의 반대와 대운하 사업, 청계천, 뉴라이트와 한나라당, 국토해양부 장관에 대해 주로 다루고 있다. 한겨레는 6개의 클러스터로 나누어지며, 주로 대운하에 대해 중점적으로 언급하고 있으며, 4대강 사업의 대운하 사업으로의 확장과 정책으로서의 실



<Figure 12> Keyword-Network of 4-major-river Project in Kyunghyang(Left) and HanKyoreh(Right)

효성 분석, 경제적인 투자나 일자리 창출, 녹색성장 등에 대해 조명하고 있다. 동아일보는 17개의 군집이 발견되어 <Figure 13> 상에 해당 군집들을 모두를 표시하지는 않았다. 해당 기간(한 달) 내에 있던 4대강 사업에 대한 기사 자체가 총 15개로 다른 신문사들에 비해 비교적 적었으며, 그렇기에 다량의 군집화 현상이 일어나는 것으로 판단이 된다. 동아일보는 <Table 14>의 단어 빈도 상의 순위에 대한 결과와 동일하게, 살리기나 경제계획과 같은 단어가 크게 차지하고 있으며 일자리, 뉴딜 정책, 투자에 관해 상세히 다루고 있다. 또한 충주, 나주, 안동, 섬진강 등과 같이 지역을 주로 언급함으로써 해당 지역에 사업을 진행할 때 주로 기사를 제공한다고 파악할 수 있다.



<Figure 13> Keyword-Network of 4-major-river Project in Dong-A

<Table 15>를 보면 4대강 사업의 경우 전체적으로는 낮은 확률의 분류 성능을 보여주고 있어 신문사별 특별한 차이를 보이고 있지 않다고 해석할 수 있으나, 사실상 내용을 확인해보았을 때는

해당 키워드에 대해 부정적이나 긍정적인 정도가 신문사별로 확연히 차이가 남을 발견했다. 분류기 상으로 이와 같은 신문사별 확연한 차이가 발견되지 않는 이유를 분석해본 결과, 경향신문과 한겨레 두 신문의 논조나 내용이 거의 비슷하여 경향신문이 한겨레로, 한겨레가 경향신문으로 분류가 되는 경우가 많았음을 발견하였다. 그렇기에 정확률과 재현률의 결과가 수치상 낮아졌지만, 실은 경향신문과 한겨레가 비슷한 논조 성향을 가졌기에 서로에 교차되어 분류됨으로써 이런 현상이 나타났다고 해석할 수 있다. 반면 동아일보는 이들 두 신문사에 비해 다른 논조와 내용으로 기사를 쓰고 있어 그 신문 만의 특징이 있으므로 높은 정확률의 분류 결과를 보여주었음을 알 수 있다. 내용이나 논조상 많은 차이가 나는 것을 Clustering을 통해서만 밝혀내기는 한계가 있다고 판단이 되었기에, 긍정 및 부정의 논조를 파악한 후 분류기에 학습시켜 그 정도를 다시 분석해보았다.

<Table 15> Precisions, Recall and F-value of Three Newspapers in National Issues Section

	정확률	재현률	F-value
경향일보	0.435897	0.34	0.382022
한겨레	0.349398	0.58	0.43609
동아일보	0.642857	0.36	0.461538

3.6.1 4대강 사업에 대한 긍정, 부정적 논조 분류 성능 평가

Naïve Bayes를 사용하여 테스트 결과를 분류하였으며, 분류 결과를 살펴보아 해당 분야에 부정-긍정으로 나뉘어진 결과의 정확률과 재현률, F-value를 구하여 분류 성능을 평가해 보았다. 4대강 사업에 대한 모든 기사에 대해 긍정, 부정성을 나눈 후, 분류기의 성능을 평가하기 위해 학습을 시킨 긍정 및 부정적인 문단은 각 신문별로 150개씩이고, 테스트를 한 문단의 수는 각 신문사 당 50개씩으로 통일시켜 실험하였다. 즉 학습 문단을 전체 문단의 75%, 테스트 문단을 전체 문단의 25%로 하였다.

2009년 12월부터 한 달 간의 기간에 있는 세 신문사의 모든 신문기사들을 모아, 학습 데이터를 모으고 분류기법의 평가를 위해 신문기사를 직접 분류해보았을 때, 한겨레와 경향신문은 4대강 사업에 대해 대부분 부정적으로 인식하고 있었으며, 동아일보는 4대강의 긍정적인 부분을 부각시키고 있었다. 이를 분류하였을 때, 3개의 신문사에 관한 정확률과 재현률, F-value는 모두 0.6이상을 넘겨, 평균 0.7 정도로 다소 신뢰성이 있게 분류되었다고 판단할 수 있다.

<Table 16> Precisions, Recall, F-value of Three Newspapers in Positive-Negative Evaluation in the Issue of 4-major-river Project

	정확률	재현률	F-value
긍정	0.672414	0.78	0.722222
부정	0.62	0.738095	0.673913

3.6.2 4대강 사업에 대한 긍정, 부정적 논조 분류 실험

위의 정확률과 재현률, F-value를 확인해본 결과, 긍정적인 것에 대해 올바로 분류가 되는 확률은 F-value 상 72% 정도이고, 부정적으로 분류될 확률은 67% 정도여서 어느 정도 논조나 내용의 차이점 및 각 신문사의 성향을 보기에 의미가 있는 수치라고 보인다. 따라서 이 분류기에 대해 70% 정도의 신뢰도를 가지고 다음과 같은 실험을 해보았다.

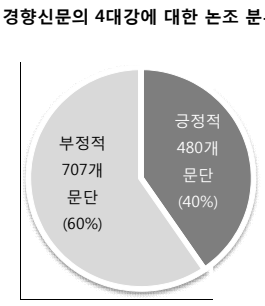
2012년 4월 8일부터 7월 8일까지 3개월 간의 4

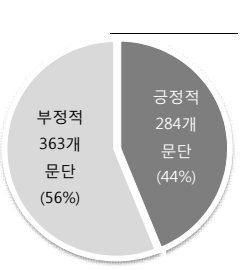
대강 사업에 대한 세 신문사의 신문기사들을 모두 수집하여 분류 결과를 확인해보고자 하였다.

<Table 17> Retrieved Date, Keywords and the Number of News in 4-major-river Project

검색 일자	2012.0 4. 08~2012. 07. 08
검색 내용	키워드 : 4대강 사업
신문기사 수	경향 : 97개/한겨레 : 85개/동아일보 : 47개

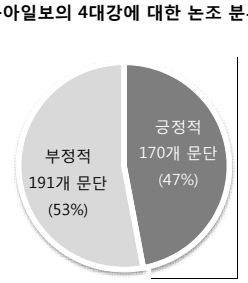
신문기사의 수는 경향신문 97개, 한겨레 85개로 두 신문사가 같은 시기에 비슷한 수준으로 신문기사를 게재했지만, 동아일보의 경우 같은 기간 내에 총 47개인 것으로 확인해 두 신문사에 비해서 상대적으로 적은 빈도로 신문기사를 게재함을 알 수 있었다.





<Figure 14> Positive-Negative Classification of 4-major-river Project in Kyunghyang and HanKyoreh

<Figure 14>와 같이 경향신문과 한겨레의 4대강에 대한 논조 분류는 확연하게 부정적인 논조가 많았음을 확인할 수 있다. 부정적인 문단이 전체 55% 이상을 차지하여 두 신문사 모두 전체적인 논조는 4대강 사업에 대해 부정적인 성향을 띠고 있음을 파악할 수 있다.



<Figure 15> Positive-Negative Classification of 4-major-river Project in Dong-A

<Figure 14>와는 달리, <Figure 15>의 동아일보 분류 결과를 살펴보면 앞의 두 신문사에 비해 긍정-부정의 분류가 달라져 있음을 알 수 있다. 즉 4대강 사업에 대해 70%의 확률로 동아일보는 경향신문에 비해 7% 더 긍정적인 성향을 보이며, 한겨레에 비해서는 3% 정도 더 긍정적으로 언급하는 경향을 보인다고 분석할 수 있다.

4. 결론 및 제언

본 연구에서 텍스트 마이닝으로 3개의 신문사 6가지 분야를 분석한 결과, 각 신문사별로 해당 이슈에 대해 자주 언급하는 특정 단어가 있음을 알 수 있었다. 특히 경제분야 전반이나 사회분야에서의 4대강 사업 관련 신문기사들에서 경향신문과 한겨레는 특정 단어들이 비슷한 순위에 위치하고 있었다면, 동아일보의 경우는 두 신문사와는 다르

게 구성되어 있는 경우가 많아 현상을 다르게 바라보고 있다고 해석할 수 있었다. 신문기사의 키워드들을 Clustering하였을 때 다른 연구에서는 주로 다루지 않았던 문화 전반에 관한 기사나 정치 분야에서의 통합진보당 관련 기사에서 세 신문사들 간의 내용 상 차이가 있음을 발견하였다는 점에서 의의가 있다. 또한 본 연구를 통해 신문기사의 키워드들 간 관계를 네트워크 상으로 시각화하여, 실제로도 경향신문과 한겨레, 동아일보가 다른 내용으로 구성되어 있다는 사실을 객관화된 데이터로 살펴볼 수 있었다. 논조의 부정 및 긍정 정도를 평가하기 위한 Classification에서는, 약 70%의 확률로 긍정과 부정적 논조가 분류되었으며 이를 활용해 살펴본 결과 4대강 사업에 대해 2012년 4월~7월 3달간 경향신문과 한겨레는 부정적인 논조를 많이 사용했으나 동아일보는 두 신문에 비해서는 좀 더 긍정적인 논조를 사용했음을 알 수 있었다.

본 연구는 지금까지 연구되어왔던 한글 신문기사의 코딩 및 담화분석 방법에서 벗어나, 텍스트 마이닝 기법을 활용하여 다량의 데이터를 분석하였다는 데에 의의가 있다. 한글 신문기사 3,000여 개를 분석하였고, 개별 신문기사를 단위로 하여 분석한 것이 아니라 신문기사 내의 문단을 단위로 분석을 시도하여 Clustering 및 Classification을 했다는 점이 여느 연구들과 다른 점이다. 또한 3개월간에 거친 많은 분량의 신문기사들을 모두 직접 분류한 것이 아니라 텍스트 마이닝 기법 중 하나인 classification 방법을 통해 과거 신문기사들을 학습시킨 후, 실제 테스트 데이터들의 긍정과 부정적인 논조의 정도를 분류하여 이를 객관적인 데이터로 보여준 점도 다른 연구들과의 차이점이다.

본 연구의 한계점은, 4대강 사업에 대한 내용 외에도 신문사별로 논조상 긍정성과 부정성에 차이를 보이는 분야를 더 발견해내지 못했다는 점이다. 또한 긍정-부정의 두 분류로만 나누는 것이 아니라 중립의 입장도 고려했어야 하지만 단순히 긍정-부정의 두 개의 분류로만 접근했기에 결과 분석이 다소 간결했다는 아쉬움이 있다. Classification 분석에서는 4대강 사업에 대한 논조 분류 결과 0.7 정도의 분류 성능을 가져왔기에 높은 수준의 분류 성능을 만족하지는 못했다. Naïve Bayes 분류기만을 사용했지만, 향후 연구에서는 그 외의 분류기들을 다양하게 결합하여 사용해보아 분류 성능을 높일 수 있는 방안을 찾아 나가도록 해야 할 것이다.

본 연구에서는 신문기사 상에 등장하는 단어들의 단순빈도를 활용하여 분석을 하였는데, 단순하게 빈도만을 활용한 방법 대신 가중치를 부여한 TFIDF, Maximum Entropy 방법 등을 활용하여 연구한다면 현재의 분석과는 또 다른 결과가 나올 수 있으리라 본다. 추후 연구 진행 시 신문 내용을 분석함에 있어서 가중치를 고려한 새로운 접근방법을 택하여, 신문사들의 특성을 보다 효과적으로 밝혀낼 수 있는 방법을 지속적으로 발견해나가야 할 것이다.

Classification을 진행할 시, uni-gram을 채택하여 한 단어들로만 이루어진 결과 분석을 하였다. 하지만 텍스트 마이닝의 여러 방법 중에서는 bi-gram이나 tri-gram 등 여러 단어를 합성한 분석결과를 보여주기도 한다. 본 연구에서 uni-gram을 활용한 이유는 영어와는 달리 한글의 특성상 bi-gram 이상으로 자르게 되면 같은 의미를 가지고 있어도 다른 의미로 분류되는 경우가 많기 때문이었다. 특히 조사 등에 의해 같은 내용이 다르게 분류되었을 경우 이를 분간해내기 어려운 경우가 많다. 추후 이와 같은 기술적인 한계를 극복한다면 bi-gram 이상으로 분석을 통해 현재와는 또 다른 새로운 분석이 가능하리라 보며, 향후 연구를 진행 시 이를 고려할 수 있다.

본 연구의 결과를 활용하여 후속연구로 텍스트 마이닝을 이용한 자동으로 신문의 논조를 분석하는 시스템을 개발할 것이다. 이와 같은 시스템은 사람들이 온라인에서 신문기사를 접할 때 특정 논조 성향에 대해 우선적으로 파악한 후 객관성을 유지한 채 정보에 접근할 수 있도록 도와주는 신뢰성 있는 툴이 될 수 있으리라 본다. 예를 들어 온라인 신문기사의 경우, 신문기사를 접하는 독자들이 기사 내용에 대한 객관성을 유지할 수 있도록 신문기사의 윗부분에 해당 신문사가 가지고 있는 특정 이슈에 대한 긍정성 및 부정성의 정도를 표시해줄 수 있을 것이다. 이러한 시스템의 실현은 본 연구에서 앞서 제시한 것과 같이, 사전에 특정 이슈에 대한 긍정성과 부정성을 나타내는 문단을 분류기에 학습시켜 놓은 후 분류하고자 하는 신문기사를 문단 단위로 끊어 테스트함으로써 해당 신문기사의 긍정성, 부정성 정도가 파악 가능해진다. 이를 위해 각 이슈 별로 학습된 데이터들이 필요할 것이다. 학습된 데이터를 통해 테스트 데이터의 긍정성과 부정성 정도를 <Figure 14>이나 <Figure 15>와 같이 나타내 보여줄 수 있을 것이며, 그 신뢰도 수준은 본 연구에 따르자면 72% 정도로 제시해줄 수 있지만 향후 연구에서의 노력에 따라 분류기의 신뢰도를 더욱 높일 수도 있을 것이다. 더 나아가 분류 결과에 따른 신문기사 논조의 스팩트럼을 작성하여 그래프로 제공하는 형식을 사용할 수도 있다. 이러한 스팩트럼은, 독자가 접하는 신문사가 다른 신문사들과는 어떤 논조의 차이를 보이고 있는지, 다른 신문사들과 비슷하게 언급하는 점은 무엇인지를 보여줄 수 있는 다차원의 그래프가 될 것이다. 즉 하나의 주제에 대한 긍정성 및 부정성에 대한 수치를 색상으로 제공하여 시각적으로 제시할 수 있고, 다른 신문사에 비해서는 어느 정도의 긍정성 혹은 부정성을 가지는지를 그래프 상에서 상대적인 위치를 표시해줌으로써 나타낼 수 있을 것이다. 이러한 결과 제시는 해당 주제들에 따라 학습 데이터들을 다수 확보했을 경우 가능해진다. 신문사별로 비교 시에 결과의 객관성을 유지하기 위해서는 신문기사 주된 이슈의 범위 및 이슈화된 기간, 혹은 검색이 이루어지는 기간을 통일시켜 제시해야 할 것이다. 이와 같이 본 연구를 활용한다면, 독자들이 자유롭게 객관성을 유지한 채 신문기사 및 사회 현상을 바라볼 수 있도록 돕는 새롭고 획기적인 매스미디어 툴을 만들 수 있으리라 기대한다.

참고문헌

Baek, Y. K. and Y. M. Seo, “A Study on Automatic Classification System of Hangul Internet News Articles”, Annual Fall symposium of 2003 of The Korea Society of Management Information Systems, *The Korea Society of Management Information Systems*, (2003), 574~580.

Balahur, A. and R. Steinberger, “Rethinking sentiment analysis in the news : From theory to practice and back”, *In Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis*, Satellite to CAEPIA 2009.

Carlos, H. Caldas, and L. Soibelman, “Automating hierarchical document classification for construction management information systems”, *Journal of Automation in Construction*, Vol. 12(2003), 395~406.

Choi, H. J., *A study on diversity of opinion in news market and Report Characteristics of Major Newspapers,* Korean journal of journalism and communication studies, 2010.

Choi, J. H., H. Kim, and N. Im, “Keyword Network

Analysis for Technology Forecasting", *Journal of Intelligence and Information Systems*, Vol.17, No.4(2011), 227~240.

Chung, J. C., "Korean Press and Discourse of Ideology", *Korean journal of journalism and communication studies*, Vol.46, No.4(2002), 314~348.

Chung, Y. M., *Research in Information Retrieval*, Ku-mi, Seoul, 2005.

Kim, J. A. and B. Chae, "The Political Attitude of Newspapers and the Coverage of Political Scandal", *Journal of communication and information*, Vol.41(2008), 232~267.

Kim, N. W. and J. Park, "Personal Information Detection by Using Naive Bayes Methodology", *Journal of Intelligence and Information Systems*, Vol.18, No.1(2012b), 91~107.

Kim, Y. S., N. Kim, and S. R. Jeong, "Stock-Index Invest Model Using News Big Data Opinion Mining", *Journal of Intelligence and Information Systems*, Vol.18, No.2(2012a), 143~156.

Lee, J. Y., "Centrality Measures for Bibliometric Network Analysis", *Korean Society for Library and Information Science*, Vol.40, No.3(2006a), 191~214.

Lee, J. Y., "A novel clustering method for examining and analyzing the intellectual structure of a scholarly field", *Korean Society for information Management*, Vol.23, No.4(2006b), 215~231.

Lee, M. K. and S. J. Kim, "A Comparative Analysis over News Framing of the Abolition of the Family Headship (Hoju) System : Examining Three Major Korean Dailies : Chosun, Kukmin, Hankyoreh", *Journal of communication and information*, Vol.34(2006), 132~162.

Media Today, "News Papers Report National Inspection Results with Party Biased view", Accessed 2012. 04. 12, <http://www.mediatoday.co.kr/news/articleView.html?idxno=91565>.

Pollak, S., R. Coesemans, W. Daelemans, and N. Lavrač, "Detecting Contrast Patterns in Newspaper Articles by Combining Discourse Analysis and Text Mining", *Pragmatics*, Vol.21, No.4(2011), 647~683.

Abstract

A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis

Miah Kam* · Min Song**

This study analyses the difference of contents and tones of arguments among three Korean major newspapers, the Kyunghyang Shinmoon, the HanKyoreh, and the Dong-A Ilbo. It is commonly accepted that newspapers in Korea explicitly deliver their own tone of arguments when they talk about some sensitive issues and topics. It could be controversial if readers of newspapers read the news without being aware of the type of tones of arguments because the contents and the tones of arguments can affect readers easily. Thus it is very desirable to have a new tool that can inform the readers of what tone of argument a newspaper has.

This study presents the results of clustering and classification techniques as part of text mining analysis. We focus on six main subjects such as Culture, Politics, International, Editorial-opinion, Eco-business and National issues in newspapers, and attempt to identify differences and similarities among the newspapers. The basic unit of text mining analysis is a paragraph of news articles. This study uses a keyword-network analysis tool and visualizes relationships among keywords to make it easier to see the differences.

Newspaper articles were gathered from KINDS, the Korean integrated news database system. KINDS preserves news articles of the Kyunghyang Shinmun, the HanKyoreh and the Dong-A Ilbo and these are open to the public. This study used these three Korean major newspapers from KINDS. About 3,030 articles from 2008 to 2012 were used. International, national issues and politics sections were gathered with some specific issues. The International section was collected with the keyword of 'Nuclear weapon of North Korea.' The National issues section was collected with the keyword of '4-major-river.' The Politics section was collected with the keyword of 'Tonghap-Jinbo Dang.' All of the articles from April 2012 to May 2012 of Eco-business, Culture and Editorial-opinion sections were also collected. All of the collected data were handled and edited into paragraphs. We got rid of stop-words using the Lucene Korean Module.

* Master's Course, Graduate school of Library and Information Science, Yonsei University
** Corresponding Author: Min Song
Department of Library and Information Science, Yonsei University, 50, Yonseiro, Seodaemun-gu, Seoul 120-749, Korea
Tel: +82-2-2123-2416, Fax: +82-2-393-8348, E-mail: min.song@yonsei.ac.kr

We calculated keyword co-occurrence counts from the paired co-occurrence list of keywords in a paragraph. We made a co-occurrence matrix from the list. Once the co-occurrence matrix was built, we used the Cosine coefficient matrix as input for PFNet(Pathfinder Network). In order to analyze these three newspapers and find out the significant keywords in each paper, we analyzed the list of 10 highest frequency keywords and keyword-networks of 20 highest ranking frequency keywords to closely examine the relationships and show the detailed network map among keywords. We used NodeXL software to visualize the PFNet. After drawing all the networks, we compared the results with the classification results. Classification was firstly handled to identify how the tone of argument of a newspaper is different from others. Then, to analyze tones of arguments, all the paragraphs were divided into two types of tones, Positive tone and Negative tone. To identify and classify all of the tones of paragraphs and articles we had collected, supervised learning technique was used. The Naïve Bayesian classifier algorithm provided in the MALLET package was used to classify all the paragraphs in articles. After classification, Precision, Recall and F-value were used to evaluate the results of classification.

Based on the results of this study, three subjects such as Culture, Eco-business and Politics showed some differences in contents and tones of arguments among these three newspapers. In addition, for the National issues, tones of arguments on 4-major-rivers project were different from each other. It seems three newspapers have their own specific tone of argument in those sections. And keyword-networks showed different shapes with each other in the same period in the same section. It means that frequently appeared keywords in articles are different and their contents are comprised with different keywords. And the Positive-Negative classification showed the possibility of classifying newspapers' tones of arguments compared to others. These results indicate that the approach in this study is promising to be extended as a new tool to identify the different tones of arguments of newspapers.

Key Words : Newspapers, Tone of Arguments, Contents, Text Mining, Clustering, Classification, The Kyunghyang Shinmun, The Hankyoreh, The Dong-A Ilbo, International Section, Politics Sections, National Issues Section, Eco-Business Section, Culture Section, Editorial-Opinion, Positive-Negative Classification

저 자 소 개

감미아

성균관대학교 사회복지학 및 심리학 학사와 연세대학교 문헌정보학 학사를 했으며 연세대학교 문헌정보학과 대학원에 재학 중이다. 관심분야는 텍스트 마이닝을 통한 감정정보 분석과 오피니언 마이닝이며, 그 외 계량정보학이나 도서관 경영과 도서관 서비스 및 이용자 연구에 관심을 가지고 있다. 정보학 분야의 연구로는 계량정보학을 활용한 천문학 분야의 키워드와 시소러스의 연계성을 분석하여 정보학분야의 국제 컨퍼런스에 게재된 바 있으며, 도서관 분야에서는 도서관 경영과 개인의 인식에 관련된 연구를 하여 학술대회 논문집에 게재되었다.

송 민

연세대학교 도서관학 학사와 Indiana University 문헌정보학 석사를 했고 Drexel University의 School of Information Science and Technology에서 박사학위를 마쳤다. 현재 연세대학교 문헌정보학과에 부교수로 재직 중이며 연세대학교로 부임하기 전에는 필라델피아에 있는 Thomson Reuters사에 Senior Software Engineer로 1999년부터 2005년까지 근무했으며 그 후 2006년부터 2012년 2월까지 뉴저지 공대(New Jersey Institute of Technology)에 정보시스템과에 부교수로 근무했다. 전공 분야는 Text Mining이며 지금까지 SCI급 논문 20편과 국제학술지 50편을 게재했고 Text Mining분야에서 활발한 학술활동을 펼치고 있다.