ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2013 September: 19(3): 1~23 http://dx.doi.org/10.13088/jiis.2013.19.3.001

# 사회문제 해결형 기술수요 발굴을 위한 키워드 추출 시스템 제안

정다미
서울대학교 융합과학기술대학원
디지털정보융합전공
(tclickdami@gmail.com)

김재석
서울대학교 융합과학기술대학원
디지털정보융합전공
(jck@snu.ac.kr)

김기남
아주대학교 일반대학원 미디어학과
(gnkim@ajou.ac.kr)

허종욱
한국과학기술원 일반대학원 웹사이언스학과
(juheo@mmc.kaist.ac.kr)

온병원
서울대학교 차세대융합기술연구원
(bwon@snu.ac.kr)

강미정
서울대학교 차세대융합기술연구원
(vchm162@snu.ac.kr)

융합 R&D가 추구해야 할 바람직한 방향은 이종 기술 간의 결합에 의한 맹목적인 신기술 창출이 아니라, 당면한 주요 문제를 해결함으로써 사회적 니즈를 충족시킬 수 있는 기술을 개발하는 것이다. 이와 같은 사회문제 해결형 기술 R&D를 촉진하기 위해서는 우선 우리 사회에서 주요 쟁점이 되고 있는 문제들을 선별해야 한다. 그런데 우선적이고 중요한 사회문제를 분별하기 위해 전문가 설문조사나 여론조사 등 기존의 사회과학 방법론을 사용하는 것은 참여자의 선입견이 개입될 수 있고 비용이 많이 소요된다는 한계를 지닌다.

기존의 사회과학 방법론이 지닌 문제점을 보완하기 위하여 본 논문에서는 사회적 이슈를 다루고 있는 대용량의 뉴스 기사를 수집하고 통계적인 기법을 통하여 사회문제를 나타내는 키워드를 추출하는 시스템의 개발을 제안한다. 2009년부터 최근까지 3년 동안 10개 주요 언론사에서 생산한 약 백 30만 건의 뉴스기사에서 사회문제를 다루는 기사를 식별하고, 한글 형태소 분석, 확률기반의 토픽 모델링을 통해 사회문제 키워드를 추출한다. 또한 키워드만으로는 정확한 사회문제를 파악하기 쉽지 않기 때문에 사회문제와 연관된 키워드와 문장을 찾아서 연결하는 매칭 알고리즘을 제안하다. 마지막으로 사회문제 키워드 비주얼라이제이션 시스템을 통해 시계열에 따른 사회문제 키워드를 일목요연하게 보여줌으로써 사회문제를 쉽게 파악할 수 있도록 하였다.

특히 본 논문에서는 생성확률모델 기반의 새로운 매칭 알고리즘을 제안한다. 대용량 뉴스기사로부터 Latent Dirichlet Allocation(LDA)와 같은 토픽 모델 방법론을 사용하여 자동으로 토픽 클러스터 세트를 추출할 수 있다. 각 토픽 클러스터는 연관성 있는 단어들과 확률값으로 구성된다. 그리고 도메인 전문가는 토픽 클러스터를 분석하여, 각 토픽 클러스터의 레이블을 결정하게 된다. 이를 테면, 토픽 1 = {(실업, 0.4), (해고, 0.3), (회사, 0.3)}에서 토픽 단어들은 실업문제와 관련 있으며, 도메인 전문가는 토픽 1을 실업문제로 레이블링 하게 되고, 이러한 토픽 레이블은 사회문제 키워드로 정의한다. 그러나 이와 같이 자동으로 생성된 사회문제 키워드를 분석하여 현재 우리 사회에서 어떤 문제가 발생하고 있고, 시급히 해결해야 될 문제가 무엇인지를 파악하기란 쉽지 않다. 따라서 제안된 매칭 알고리즘을 사용하여 사회문제 키워드를 요약(summarization)하는 방법론을 제시한다. 우선, 각 뉴스기사를 문단(paragraph) 단위로 세그먼트 하여 뉴스기사 대신에 문단 세트(A set of paragraphs)를 가지게 된다. 매칭 알고리즘은 각 토픽 클러스터에 대한 각 문단의 확률값을 측정하게 된다. 이때 토픽 클러스터의 단어들과 확률값을 이용하여 토픽과 문단이 얼마나 연관성이 있는지를 계산하게 된다. 이러한 과정을 통해 각 토픽은 가장 연관성이 있는 문단들을 매칭할 수 있게 된다. 이러한 매칭 프로세스를 통해 사회문제 키워드와 연관된 문단들을 검토함으로써 실제 우리 사회에서 해당 사회문제 키워드와 관련해서 구체적으로 어떤 사건과 이슈가 발생하는 지를 쉽게 파악할 수 있게 된다. 또한 매칭 프로세스와 더불어 사회문제 키워드 가시화를 통해 사회문제 수요를 파악하려는 전문가들은 웹 브라우저를 통해 편리하게 특정 시간에 발생한 사회문제가 무엇이며, 구체적인 내용은 무엇인지를 파악할 수 있으며, 시간 순서에 따른 사회이슈의 변동 추이와 그 원인을 알 수 있게 된다.

개발된 시스템을 통해 최근 3년 동안 국내에서 발생했던 다양한 사회문제들을 파악하였고 개발된 알고리즘에 대한 평가를 수행하였다(본 논문에서 제안한 프로토타입 시스템은 http://dslab.snu.ac.kr/demo.html에서 이용 가능함. 단, 구글크롬, IE8.0 이상 웹 브라우저 사용 권장).

논문접수일 : 2013년 05월 10일 논문수정일 : 2013년 07월 03일 게재확정일 : 2013년 08월 06일
투고유형 : 국문급행 교신저자 : 온병원

* 이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 일반연구자지원사업 지원을 받아 수행된 것임 (No. 2013012524).

## 1. 서론

오늘날 융합기술은 미래 기술의 혁신을 주도할 신 성장동력으로 주목 받고 있다. 사회 전반에 걸쳐 막대한 변화를 초래할 것으로 기대되는 융합기술은 차세대 신 산업 창출은 물론이고, 의료 및 건강, 안전, 에너지와 환경 등과 관련된 주요 사회문제의 해결에 기여할 것으로 전망된다. 그러나 IT, NT, BT와 같이 인접한 과학기술 사이에서 추구되어온 기존의 융합기술 R&D는, 선진국으로 발돋움하며 변화하고 있는 우리 사회의 요구를 충족시킬 수 없을뿐더러, 인간적 삶의 가치와는 동떨어진 기술 그 자체를 위한 기술에 머물 수밖에 없다는 한계가 지적되고 있다.

이와 같은 반성과 함께 최근에는 융합 R&D가 인문적 가치와 사회적 니즈를 반영하는 방향으로 나아가야 한다는 필요성이 대두되고 있다. 말하자면 우리 사회 구성원들의 가치관을 고려하고 그들의 사회적, 문화적 요구를 해결하지 않는다면 아무리 첨단의 혁신 기술이라 하더라도 그저 실험실 안에서만 통용되는 것에 그칠 것이라는 것이다. 요컨대 현 시점에 필요한 R&D는 맹목적 기술 개발 논리에 함몰되지 않고 사회적 니즈를 충족시키고 더 나아가 사회적으로 문제시되는 이슈들을 해결하는데 기여하는, “사회문제 해결형” 기술의 연구와 개발이라고 할 수 있다.

사회문제 해결에 기여하는 기술의 연구 및 개발을 위해서는 이에 앞서 오늘날 사회 각 분야에서 쟁점이 되고 있는 문제가 무엇이며, 사회구성원들이 요구하는 기술이 무엇인지 밝혀내는 기술수요에 대한 정확한 예측이 필요하다. 건강 및 의료, 환경 및 에너지, 교육, 복지의 각 분야별로 다양하게 나타나는 사회문제와 이에 따른 니즈를 조사하고 분석하는 일은 전통적으로 사회학, 경제학, 정치학, 심리학 등 사회과학에서 담당해온 것이다.

주지하듯이 사회과학 조사방법론에는 정량적 방법과 정성적 방법이 있다. 가령 인구분포도처럼 해당 분야의 통계자료를 활용한 사실 확인은 정량적 방법을 통해 가능하다. 반면 인구분포도에서 확인된 노령인구 증가 같은 사회적 쟁점의 원인 분석은 전문가 인터뷰나 델파이 기법 등 전문가 대상 설문조사방법을 통해, 즉 정성적 방법을 통해 가능해진다. 이 연구에서 주목하고 있는 “사회문제 해결형 기술수요 발굴”을 위해서는 정량적 방법과 정성적 방법이 상보적으로 사용되어야 한다.

우리 사회에서 문제시되고 있는 쟁점들을 카테고라이징 하고, 한 카테고리 내의 여러 쟁점들 중에서도 어떤 것이 우선적으로 해결되어야 하는가는 문헌연구, 여론조사, 설문조사 또는 전문가 인터뷰 같은 질적인 방법으로 판단되어야 한다. 사회문제에 대한 정성적 접근에 앞서 우선 정량적 접근이 요구되는데, 이는 사회현상의 변화와 우리 사회에서 쟁점화되고 있는 문제들에 대한 계량적 사실들이 파악되어야 하기 때문이다.

최근 빅데이터 마이닝은 기존의 사회과학적 방법들의 한계를 보완하고 더 나아가 대체할 방안으로 떠오르고 있다. 특히 여론조사나 설문조사 등 정성적 방법으로 정확하게 확인하기 힘든 사회현상의 변화나 현재 주요 쟁점이 되고 있는 사회문제를 빅데이터 분석을 통해 파악할 수 있게 되었다. 다시 말해 데이터마이닝은 기존의 정량적 방법을 대체하는 사회과학 방법론으로 부상하고 있는 것이다. 또한 델파이 조사나 전문가 인터뷰, 또는 일반인 대상 여론 조사와 통계처리 같은 정성적 방법들은 참여자의 선입견을 배제하기가 어렵다는 일정한 한계를 지니는데, 이러한 정성적 방법의 주관주의적 한계를 빅데이터 분석을 통해 보완할 수 있다. 더욱이 델파이 조사 등을 실시하려면 준비에서 조사결과 처리에 이르기까지 상당한 시간과 비용이 소요된다

는 점을 고려한다면, 빅데이터 마이닝은 여러 측면에서 유용하다고 할 수 있다.

이상에서 설명한 것처럼 본 연구는 기존의 사회과학적 방법의 문제와 한계를 보완하고 극복하는 것을 궁극적 목적으로 삼고 있다. 다시 말해 우리 사회의 당면한 문제들을 선별함에 있어서 전문가 설문조사나 일반인 여론조사가 초래하는 주관성의 한계를 인식하고, 이를 보완하기 위한 객관적인 방법을 개발하기 위해 본 연구에서 진행하게 된 것이다. 이와 같은 목적을 가지고 본 논문에서는 대용량의 뉴스 기사를 수집하고 통계적인 기법을 사용하여 주요 사회문제를 제시하는 키워드들을 추출하는 시스템을 제안하고자 한다.

본 논문에서는 일차적으로 한국언론진흥재단의 KINDS(Korean Integrated News Database System) 웹사이트(www.kinds.or.kr)로부터 2009년 6월 이후 최근 3년간의 약 백 30만 건의 뉴스기사 데이터에서 사회문제를 다루는 기사를 식별해 냈고, 한글 형태소 분석, 확률 기반의 토픽 모델링을 통해 주요 사회문제에 관한 키워드들을 추출하였다. 더 나아가, 시급성과 중요성에 있어서 더 우선적인 해결을 요구하는 사회문제를 파악하고 정확도를 높이기 위해, 앞서 추출한 사회적 이슈들과 연관된 키워드와 문장을 찾아 연결시키는 매칭 알고리즘을 제안하였다. 마지막으로, 본 연구진은 사회문제 키워드를 한 눈에 볼 수 있게 해주는 비주얼라이제이션 시스템을 활용하여, 최종적으로 도출된 주요 사회문제들이 일목요연하게 파악될 수 있도록 하였다. 다음 제 2장에서는 본 연구와 관련된 기존 연구에 대해 다루고자 한다.

## 2. 관련연구

Blei et al.(2003)는 확률 기법을 기반으로 문서의 토픽을 파악할 수 있는 토픽 모델링 알고리즘 중 하나인 LDA(Latent Dirichlet Allocation)를 제안하였다. 각 문서에는 여러 토픽들이 섞여있고 서로 다른 확률로 분포되어있다는 가정하에 LDA는 결합확률분포와 조건부분포를 계산하여 각 토픽 별로 해당 키워드들을 추출한다. 이러한 LDA는 더욱 복잡한 문제를 풀 수 있는 모델의 모듈로 쓰일 수 있다(Blei, 2012). 단어들과 문서들 사이의 순서를 고려하지 않는 LDA의 이러한 통계상의 가정들을 수정하거나 기존 LDA에 메타데이터를 포함하는 방식으로 더욱 정교한 토픽 모델링 기법들이 개발되었다. Blei et al.(2006)는 기존의 LDA방식에서 문서들의 순서를 고려하여 시간의 흐름에 따른 토픽의 변화를 추적하는 연구를 수행하였다. LDA에 메타데이터를 포함하는 연구로서 Rosen-Zvi et al.(2004)는 LDA를 통해 다중 저자를 가진 논문에서 어떤 부분을 어떤 저자가 담당하는지에 대해 추정하는 연구를 진행하였다.

또한 LDA는 텍스트로 이루어진 문서 컬렉션에서 적용할 수 있을 뿐 아니라 이미지, 유전 정보, 통계수치 등 여러 데이터에서도 패턴(topic)을 추출할 수 있다. 이와 관련된 연구로(Fei-Fie et al., 2005)는 카테고리화 되어있는 대량의 이미지 데이터의 패턴(topic)을 LDA를 통해 모델링 하여 자동으로 이미지를 분류하는 방법을 제안하였다. 따라서 앞서 기존 연구사례에서와 같이 컴퓨터공학과 이공계열 학문, 더 나아가 역사학, 사회학, 언어학, 정치학 등 인문사회계열 학문과의 학제간 연구에서도 LDA가 유용하게 쓰일 수 있다(Blei, 2012).

최근 국내에서는 텍스트 마이닝 기법(Liu, 2012; Aggarwal, 2012)을 이용한 다양한 적용 사례들이 나타나고 있다. 다음소프트에서는 비정형 텍스트 데이터인 소셜미디어 데이터를 자연 언어 처리 기술인 텍스트 마이닝 기법 등을 활용하여 분석하였다.

한국전자통신연구원이 개발 중인 소셜미디어 이슈 탐지 및 모니터링 플랫폼 WISDOM은 소셜미디어 데이터를 수집/저장하고, 텍스트 마이닝을 비롯한 차별화된 심층 언어분석을 기반으로 추출한 정보를 이용하여 이슈를 탐지하고 모니터링하는 기능을 제공한다(Lee et al., 2013). 해외의 사례로는 Recorded Future가 있다. Recorded Future는 웹 사이트, 블로그, 소셜미디어 등의 구조화되지 않은 대규모 텍스트 데이터를 대상으로 정보를 추출, 분석 및 시각화 서비스를 제공하고 있다(Recorded Future, 2012).

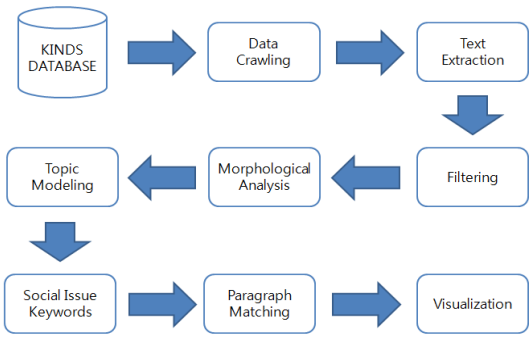
아울러 국내 주요 언론사의 논조 차이점을 텍스트 마이닝 기법을 통해 알아내는 연구가 수행되고 있다(Kam et al., 2012). 예를 들면 동아일보, 한겨레신문, 경향일보 등의 언론사에서 특정 사건과 관련 있는 키워드의 빈도수를 계산하거나 코사인 유사도를 사용하여 키워드 간의 상관관계를 나타내는 네트워크를 만들고 클러스터링 기법을 통해 분류하였다. 또한 신문사의 논조를 정확히 파악하기 위해 기계학습 기법을 적용하였다.

특정 토픽과 그에 대응되는 문서 혹은 문단(Segmentation)의 매칭은 여러 논문에서 연구되어 왔다. (Sun et al., 2007)은 T(Term)과 S(Segmentation)의 상호 의존 정도를 알 수 있는 상호 정보량(Mutual Information)과 MI에 가중치를 부여한 상호 정보량(Weighted Mutual Information)을 기반으로 여러 문서들의 토픽을 찾아내고 문서들을 토픽 별로 분절화(Segmentation)시켜 비슷한 세그멘테이션 끼리 연결시키는 연구를 수행하였다. 최적의 클러스터링은 원본의 MI(혹은 WMI) I(T;S)와 세그멘테이션 이후의 MI(혹은 WMI) I(T';S') 차가 최소가 되도록 만들어야 하는데 이를 위해 I(T';S')의 최대값을 구하기 위해 다이내믹 프로그래밍을 사용하였다. 이 모델은 기존의 여러 Topic Segmentation 방법들보다 더 좋은 성능을 보여주었는데, 특히 여러 문서의 세그멘테이션에서는 WMI가 MI보다 더 효율적이었다. 또 다른 텍스트 세그멘테이션 방법으로 (Misra et al., 2009)는 숨겨져 있는 토픽 구조로부터 실제로 관찰 가능한 문서 내 단어들이 발생하였다고 보는 생성확률모델링(Generative Probabilistic Modeling)인 LDA(Latent Dirichlet Allocation)를 이용하였다. 이 연구는 토픽과 세그멘테이션의 결합분포를 계산하여 각 세그멘테이션의 주제 내용에 대한 정보를 모으고 이것을 토픽을 알아내는 데에 반복적으로 사용하는 모델을 제안하였다. 실험결과 이러한 접근법이 일반적인 Baseline 방법보다 좀 더 우수한 성능을 보였다. Dalvi et al.(2009)는 야후의 지역 데이터베이스에 있는 식당과 Yelp의 리뷰 데이터를 연결시키는 생성확률모델을 제안하였다. 이 연구에서는 특정 식당 e에 대한 리뷰 r은 어떤 단어들은 그 식당을 가리키는 단어 text(e)에서 나왔고 다른 단어들은 e와 무관한 Generic Review Language(RLM)에서 나왔다고 가정하고 이렇게 RLM이 주어졌을 때 어떤 식당 e가 매칭되어야 할지 그 확률을 계산하였다. 그 결과 식당을 질의로, 리뷰를 문서로 처리하여 TF/IDF 값을 구하는 기존의 매칭 알고리즘보다 높은 정확도를 보였다. 이제 제 3장에서 사회문제 키워드 추출 시스템을 제안하는 본 연구에 대해 자세히 살펴보도록 하겠다.

## 3. 사회문제 키워드 추출 시스템

<Figure 1>에서 보는 바와 같이 사회문제 키워드 도출 시스템은 6개의 소프트웨어 컴포넌트들로 구성된다.

먼저 데이터수집 소프트웨어 에이전트는 국내 주요 언론사에서 발간한 뉴스기사 데이터를 아카이브하고 있는 한국언론진흥재단의 KINDS(Korean




<Figure 1> Diagram of Social Keyword System

Integrated News Database System) 웹사이트(www.kinds.or.kr)로부터 최근 3년간의 뉴스기사 데이터를 다운로드 받는다(KINDS, 2012).

일반적으로 다운로드 받은 각 뉴스기사는 언론사 ID, 뉴스ID, 날짜, 제목, 본문(텍스트), 이미지 파일 등으로 구성되며, 텍스트 추출 소프트웨어 에이전트는 사진 및 그림 파일과 동영상 파일 등을 제거한다. 그리고 추출된 뉴스기사에 대해 필터링 프로세스가 수행된다. 본 연구의 목적이 뉴스기사 데이터로부터 사회문제와 관련된 키워드를 추출하는 것에 초점을 맞추고 있기 때문에, 연예, 스포츠와 같이 일반적으로 사회문제와 관련 없는 뉴스기사들은 필터링 프로세스에 의해서 제거된다. 필터링 과정을 거쳐 사회문제를 다룰 것으로 예상되는 뉴스기사 데이터들은 한글형태소분석기를 통해 명사 식별 및 띄어쓰기 교정 등의 과정을 거친다. 그리고 3년간의 뉴스기사 데이터를 월별로 그룹핑하고 각 월별 뉴스기사 데이터를 입력하는 토픽 모델링 알고리즘을 사용하여, 그 달에 이슈가 되었던 사회문제 키워드 리스트를 추출한다. 그러나 키워드 리스트만으로 월별 사회 이슈를 정확히 파악하기 어렵기 때문에, 본 연구에서는 매칭 알고리즘을 제안한다. 매칭 알고리즘의 입력은 토픽 모델링 알고리즘을 사용하여 추출된 '키워드 리스트'와 월별의 모든 뉴스기사 텍스트들을 문단으로 세그먼트하여 만든 '문단 리스트'로 구성된다. 매칭 알고리즘은 문단 리스트 중에서 키워드 리스트와 가장 관련이 있는 상위 10개의 문단을 찾아내어 키워드 리스트와 연결한다.

마지막으로 월별 키워드와 키워드를 잘 설명하는 문단들을 시각화하여 웹 페이지에서 보여준다. 이를 통해 사회문제 이슈를 조사하는 전문가들은 최근 3년 동안 뉴스기사에서 이슈화되었던 키워드들이 무엇인지를 알 수 있고 시간이 지날수록 특정 키워드의 순위가 증가하는지 감소하는 지를 눈으로 직접 확인할 수 있다.

다음 장에서는 사회문제 키워드 도출 시스템의 각 소프트웨어 컴포넌트에 대해 자세히 살펴본다.

### 3.1 데이터 수집 및 텍스트 추출

국내 주요 언론사에서 생산한 뉴스기사 데이터를 아카이브하고 있는 한국언론진흥재단의 KINDS (Korean Integrated News Database System) 웹사이트(www.kinds.or.kr)로부터 최근 3년간의 뉴스기사 데이터를 수집하였다. KINDS 웹사이트는 조선일보를 제외한 국내 주요 언론사의 뉴스기사들을 저장하는 데이터베이스를 유지한다. 1990년부터 지금까지 10대 종합일간지(국민일보, 경향신문, 한국일보, 문화일보, 아시아투데이, 세계일보, 동아일보, 한겨레, 내일신문, 서울신문)과 미디어뉴스인 연합뉴스와 중앙일보, 그리고 지역종합일간신문 등에서 발행된 총 22백만 건의 뉴스기사들이 저장되어 있다. 본 연구를 위해 2009년 6월부터 2012년 7월까지 3년 동안 총 1,300,000건의 뉴스기사 데이터를 수집하였다. KINDS 웹사이트는 데이터베이스 서버에 뉴스기사들이 저장되어 있고 클라이언트의 웹 브라우저를 통해 뉴스기사를 검색할 수 있다. 구체적으로 javascript.viewKindsNews('언론사ID.날짜+뉴스

기사ID')와 같은 JavaScript API를 통해 해당 뉴스기사를 데이터베이스 서버로부터 웹 브라우저로 가져오게 된다. 우선 위의 API를 사용하여 3년 동안의 뉴스기사 데이터를 다운로드하기 전에 입력 파라미터인 언론사ID와 뉴스기사ID를 먼저 수집하였다. 그리고 입력 파라미터를 위의 API에 대입하여 뉴스기사들을 자동으로 다운로드 받았다. 또한 각 뉴스기사를 다운로드 받는 동시에 언론사ID, 뉴스ID, 날짜, 제목, 본문(텍스트) 등을 추출하였고 dslab.snu.ac.kr 데이터베이스 서버에 저장하였다. 각 뉴스기사는 제목과 본문 이외에 사진 및 그림과 같은 이미지 파일, 동영상 파일, 뉴스기사 타입(정치, 경제, 사설 등) 등 다양한 정보를 포함하고 있지만, 본 연구의 목적에 필요하지 않아 제거하였다.

## 3.2 필터링

일반적으로 많은 뉴스기사들은 사회문제와 상관없는 주제들을 다루는 경향이 있다. 예를 들면, 상품소개, 연예 뉴스, 스포츠 결과 등을 다루는 뉴스기사들은 사회문제 이슈와는 전혀 관련 없는 기사들이므로 필터링 모듈을 사용하여 그러한 뉴스기사를 제거하는 것이 필요하다. 본 연구에서 필터링 모듈은 1차와 2차 필터링으로 구성된다.

### 3.2.1 1차 필터링

1차 필터링은 '글자 수'와 '인용부호 유무'을 통해 뉴스기사를 제거한다. 임의로 샘플링된 만 건 이상의 뉴스기사를 조사한 바에 따르면, 글자 수가 500자 미만으로 작성된 뉴스기사는 주로 속보, 인사, 부고, 요약, 포토, 영상 관련 뉴스에 해당하였다. 또한 글자 수가 3,000자 이상으로 작성된 뉴스기사는 주로 리뷰, 칼럼, 광고, 인터뷰 위주의 기사였다. 따라서 각 뉴스기사에서 사용된 단어들의 수를 조사하여 500자 미만이고 3,000자 이상인 경우, 사회문제를 다루지 않는 기사로 판단하여 제거하였다. 또한 언론사에 재직 중인 기자들의 인터뷰를 통해, 뉴스기사에서 인용부호는 중요한 역할을 한다는 사실을 파악하였다. 즉 각 뉴스기사에는 핵심 주제가 있고 주제와 부합되면서 사실 관계를 확인한다든지 어떤 주장에 대한 지지를 얻기 위해, 관련 분야 전문가의 의견을 인용한다. 따라서 인용부호가 전무한 뉴스기사들은 중요한 기사가 아닌 것으로 판단되어 제거하였다.

### 3.2.2 2차 필터링

1차 필터링을 수행하고 남아있는 뉴스기사들을 대상으로 2차 필터링을 수행한다. 2차 필터링은 이미 구축한 사전을 통해 필터링을 수행한다. 백 30만 건의 뉴스기사 중에서 만 건의 뉴스기사를 임의로 샘플링 한 다음, 매뉴얼하게 조사하여 사회문제와 관련 없는 단어들을 선정하였다. 사전에는 문화, 연예, 예능, 영화, 야구, 리뷰, 소설, 걸 그룹, 신기록 등 사회문제와 직접 관련이 없는 50여 개의 단어들이 포함된다. 각 뉴스기사에서 사전(D)에 포함된 단어가 등장하는지를 조사하여 아래와 같은 수식을 사용하여 각 뉴스기사의 필터링 점수를 계산한다.

$$f(n_k) = \left(1 - \frac{1}{(\sum_{i=1}^{|n_k|}\sum_{j=1}^{|D|}E(w_i,\ w_j)+1)}\right)\times 100 \quad (1)$$

위의 수식에서 $n_k$는 입력으로 주어진 하나의 뉴스기사를 의미하고 $f(n_k)$는 $n_k$의 필터링 점수를 나타낸다. 또한 $|n_k|$는 $n_k$에 있는 총 단어의 개수를 말하며 $|D|$는 사전에 있는 단어의 개수를 나타낸다. $E(w_i, w_j)$ 함수는 단어 $w_i$와 $w_j$와 동일하면 1을 리턴 하고, 그렇지 않으면 0을 리턴 한다. $n_k$에 있는 모든 단어가 사전에 포함되지 않으면($\forall w_i \ni D$),

$f(n_k) = 0$이 된다. $f(n_k)$의 점수가 높을수록 $n_k$는 사회문제를 다루지 않는 뉴스기사일 확률이 높다는 것을 의미한다. 각 뉴스기사의 필터링 점수를 계산하여 상위 25%의 필터링 점수를 가지는 뉴스기사를 제거한다. 실험 결과를 바탕으로 25%의 임계 값이 가장 적당하여 뉴스기사를 필터링 하는 파라미터로 사용하였다.

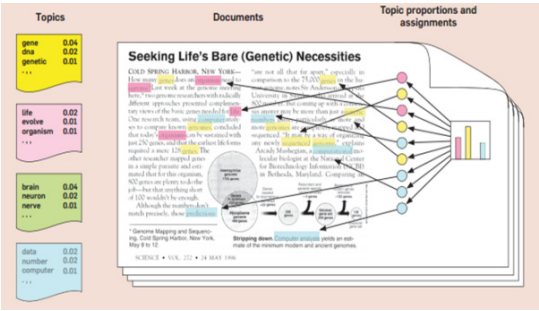
### 3.3 한글 형태소 분석

뉴스기사 원문 텍스트는 토픽 모델링 알고리즘을 바로 적용하여 사회문제 키워드를 추출하기 위해서는 정제되어 있지 않다. 예를 들어 '위해서' 또는 '일제히' 등과 같은 단어들은 사회문제 이슈에서 중요한 단어들이 아니다. 또한 '마찬가지였다', '마찬가지였어' 등의 단어들은 '마찬가지'라는 명사로 통일될 수 있지만 이를 위해서는 각 동사의 조사를 제거해주어야 한다. 이와 같이 뉴스기사의 원문 텍스트가 입력으로 주어지면, 한글 형태소 분석기는 단어의 조사를 제거한다. 또한 '야권연대'와 같은 복합명사를 '야권', '연대'와 같은 단일 명사로 나누어준다. 또한 '왕재산'과 같은 사전에 없는 미등록어를 식별하고 문장에서 띄어쓰기를 교정해 준다. 이러한 과정을 통해, 뉴스기사의 원문 텍스트는 단일 명사들의 집합이 되고, 토픽 모델링을 수행하기 위해 잘 정리된 텍스트로 재구성된다. 이를테면 한글 형태소 분석기를 사용하여 "최근 빅데이터라는 이름으로 이슈화된 데이터 처리의 중요성을 공공 데이터의 입장에서 검토한다"의 문장은 "최근 빅데이터 이름 이슈화 데이터 처리 중요성 공공 데이터 입장 검토"등의 문장으로 재구성된다. 토픽 모델링 알고리즘은 이러한 재구성된 텍스트를 입력으로 하여 사회문제 키워드를 추출하게 된다. 본 연구에서는 한글 형태소 분석기로 국민대의 한국어 형태소 분석 라이브러리 오픈 소스(Kang, 2012)를 사용하여, 각 뉴스기사의 원문 텍스트를 토픽 모델링에 사용할 수 있는 텍스트로 변환하였다.

### 3.4 토픽 모델링

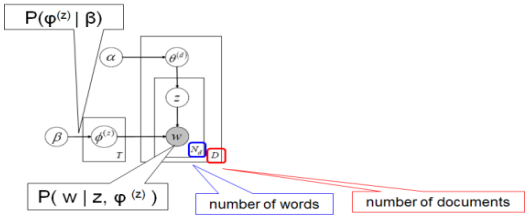
토픽 모델링 알고리즘은 문서의 주제를 알기 위해 원본 텍스트 내의 단어를 분석하는 통계적인 방법이다. 부가적인 레이블링이나 주석을 달 필요가 없어 대량의 문서 컬렉션을 기계적으로 빠르고 정확하게 처리할 수 있다. LDA(Latent Dirichlet Allocation)는 이러한 확률 그래프 모델 중 하나로 Dirichlet 분포를 이용하여 텍스트 문서 내의 단어들이 어떤 특정 토픽에 포함될 확률을 계산하는 모델이다(Blei et al., 2003). LDA는 텍스트 문서에는 여러 토픽들이 혼합되어있다는 가정에서부터 출발한다. 문서 내의 단어들은 사람이 관찰할 수 있는 변수(Observed Variable)이다. 반면 토픽의 개수, 문서별 토픽 분포도, 문서 내 단어들의 토픽 지정 등 문서 컬렉션의 토픽 구조는 숨겨져 있는 변수(Hidden Variable)이다. LDA는 숨겨져 있는 토픽 구조를 미리 가정하고 현재 관찰 가능한 문서 내의 단어들은 이로부터 생성된 것으로 다루어지는 생성 확률 모델(Generative Probabilistic Model)이다. 즉 어떤 데이터(텍스트 문서)도 생성되기 전에 이미 토픽 구조가 존재해있다고 가정하는 것이다. 문서에 나타나있는 단어들은 다음 두 단계에 걸쳐 생성된다.

1단계에서는 <Figure 2>의 오른쪽 히스토그램과 같이 문서 내의 토픽 분포도를 랜덤 하게 결정한다. 2단계에는 두 가지 과정이 수행되어야 한다. 그림의 오른쪽 부분에 나타나있는 원 도형은 토픽을 의미하는데 일단 이 중에서 토픽 하나를 선택하고 문서 내 단어들 중에서 랜덤 하게 선택하여 연결시킨다. 이 2단계 작업은 텍스트 문서 내의 각각의 단어에 대해 수행된다.




<Figure 2> Topic Distribution Chart, Derived from Document that has been Selected from Left Histogram. Circle Located in Center Mean Topic, other Shapes Aligned with Left Side are Words, Grouped by Topic. Connecting Words and Topics within Document at Random and Classifying Topics(Blei, 2012)

최종 결과는 <Figure 2>의 왼쪽과 같이 토픽 별로 단어들이 그룹핑 되어 나타난다. LDA 생성 프로세스(Generative Process)는 숨겨진 변수(Hidden Variable)인 문서 D의 토픽 분포($\theta_d$), 문서 D의 각 단어들의 특정 토픽 배정 확률($z_d$), 토픽 z의 용어 분포($\phi_z$)와 관찰되는 변수(Observed Variable)인 문서 내 단어($w_d$)에 대하여 결합확률분포(Joint Probability Distribution)을 정의한다.




<Figure 3> Graphical Model of LDA

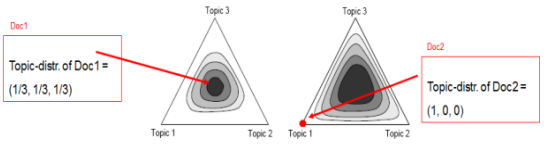
위의 <Figure 3>과 연결시켜 설명하면 $\theta_d$는 그림의 히스토그램, $z_d$는 원 도형 별 단어 매칭, $\phi_z$는 그림 오른쪽의 토픽 별 용어 분포를 의미한다. 이것을 이용하여 관찰되는 변수(Observed Variable)가 주어졌을 때 숨겨져 있는 토픽 구조가 어떠한지를 알아내기 위해 조건부 분포(Conditional Distribution)를 계산하는 것이 LDA의 핵심이다. 여기에서 조건부 분포는 Posterior Distribution이라고도 불린다. 이를 수식으로 표현하면 다음과 같다.

$$P(w, z, \phi, \theta|\alpha, \beta) \tag{2}$$
$$= \Pi_{m=1}^{M} \Pi_{n=1}^{N} P(w_{m,n}, z_{m,n}, \phi, \theta_m|\alpha, \beta)$$
$$= \Pi_{m=1}^{M} P(\theta_m|\alpha) \Pi_{n=1}^{N} P(z_{m,n}|\theta_m) P(w_{m,n}|z_{m,n}, \phi)$$

위의 수식에서 나타나는 $\alpha$와 $\beta$는 사용자가 직접 입력해야 하는 값이다. 이 전 과정을 도식화하고 좀 더 자세한 수식으로 표현하면 다음과 같다.

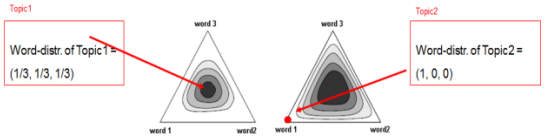
$$P(d, w) = P(d)P(\theta_d|\alpha)\sum_z P(\phi_z|\beta) \tag{3}$$
$$P(w|z, \phi_z)P(z|\theta_d)$$

$\alpha$는 문서 내 토픽 분포에 영향을 끼치는데 $\alpha$가 커질수록 한 문서 내에서 각각의 토픽들이 고르게 분포되도록 계산된다. 반면 $\alpha$가 작아질수록 어느 특정 토픽에 쏠리게 된다. 그림으로 표현하면 다음과 같이 나타낼 수 있다(Wagner, 2012).




<Figure 4> Topic Distribution in Accordance with $\alpha$

한편 $\beta$는 토픽 내의 단어가 어떤 확률로 분포되어 있는지를 나타낸다. 아래 그림에서 보는 것처럼 $\alpha$와 비슷하게 $\beta$ 역시 값이 커질수록 토픽에 속하는 각 단어들의 확률이 균등하고, 값이 작을수록 특정 단어의 확률이 커진다.




<Figure 5> Word Distribution in Accordance with $\beta$

$\alpha$와 $\beta$값은 수식에서 문서의 개수와 함께 쓰이기 때문에 본 연구에서는 매달 기사 개수에 따라 $\alpha$와 $\beta$값을 달리 주었다. 평균 $\alpha$값은 0.004~0.005, $\beta$값은 0.16~0.2로 설정하였다.

w가 주어졌을 때 가능한 토픽 구조의 개수는 기하급수적으로 늘어난다. $\alpha$와 $\beta$, 관찰되는 변수(Observed Variable)인 문서 내 단어가 주어졌을 때 토픽 구조를 확률적으로 알기 위해서는 정확한 값을 계산해내는 것이 아니라 값을 추정해야 한다. 문서 내의 각 단어들이 어떤 토픽에 연결되는지를 나타내는 z는 매우 고차원의 랜덤 변수이기 때문에 P(Z|W)는 추정할 수밖에 없다. 예를 들어 토픽 수가 50개이고 단어 개수가 1,000개인 경우 501,000번의 P(z)을 모두 계산해야 하는데 이는 거의 불가능한 수치이다. 추정하는 방식에는 직접적인 것과 간접적인 방식 두 가지가 있다. 이 중에서 간접적인 방식인 깁스 샘플링(Gibbs Sampling)이 많이 사용된다. 본 연구에서도 깁스 샘플링을 이용해 작성된 LDA 오픈 소스를 사용하였다(JGibbLDA, 2012).

깁스 샘플링은 두 개 혹은 그 이상의 변수들의 결합확률분포(Joint Probability Distribution)으로부터 연속적인 표본을 채취(Sampling)하는 것이다. 결합분포가 명확히 알려져 있지 않으나, 각 변수의 조건부 분포는 알려져 있을 경우 적용이 가능하다. 추정하고자 하는 변수의 나머지 변수에 대한 조건부 확률분포에 의존하여 교대로 표본을 채취하는 방법으로 구현된다. 수식으로 표현하면 다음과 같다. $c^{WT}$는 토픽 별 용어 분포를, $c^{DT}$는 문서 별 토픽 분포를 의미한다.

$$P(z_i = j | z_{-i}, w_i, d_i \cdot) \propto \frac{C^{WT}_{w,j} + \beta}{\sum_{w=1}^{W} C^{WT}_{w,j} + W\beta} \frac{C^{DT}_{d,j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d,t} + T\alpha} \quad (4)$$

본 연구에서는 위에서 밝혔듯이 평균 $\alpha$값은 0.004~0.005, $\beta$값은 0.16~0.2, 토픽 수는 70개, 샘플링 iteration은 100으로 설정하였다. 출력 파일로는 토픽 별로 그룹핑 된 단어와 그 확률 값을 나타내는 TWORD파일, 문서 별 토픽 분포 확률을 나타내는 THETA파일 등이 생성된다(JGibbLDA, 2012).

### 3.5 매칭

제 2.4절에서 설명한 토픽 모델링 알고리즘을 사용하여 월별 뉴스기사들에 대해 토픽들을 추출하였다. 각 토픽은 키워드 리스트와 각 키워드의 확률 값으로 이루어진다. 토픽에 속한 키워드 리스트를 살펴보고, 매뉴얼하게 토픽을 레이블링 하였다. <Table 1>은 토픽 모델링의 한 예로 '은행', '저축', '대출' 등 상위 20개의 키워드들과 해당 키워드들의 확률 값을 나타내고, 이러한 키워드들은 '저축은행 부실경영'으로 매뉴얼하게 레이블링 한다.

그러나 <Table 1>에서 보는 것처럼 토픽과 연관 키워드들만으로 저축은행 부실경영과 관련된 사회 이슈를 온전히 파악하기란 사실상 불가능하다. 따라서 본 연구에서는 토픽의 키워드들과 가장 잘 매칭 되는 문장들을 자동으로 찾아서 제공하는 알고리즘을 제안한다. 우선 월별 뉴스기사 데이터, 이를테면 2011년 9월의 11,242건의 각 뉴스기사 텍스트를 문단 단위로 세그먼트 한다. 이러한 프로세스는 월별 뉴스기사 텍스트들을 문단들의 집합으로 재구성한다. 다음, 입력으로 주어진 토픽 키워드들과 각 문단의 연관성을 수치로 계산한다. 예를 들면, 토픽 t와 문단 $s \in S$(s : 문단 집합)가 주어지고, 그 연관성을 $s^*$라고 하면, 각 $s \in S$은 $s^*$에 의해 내림차순으

<Table 1> Topic of "저축은행 부실경영"(2011. 09)

| Rank | Keyword | Probability |
|---|---|---|
| 1 | 은행 | 0.020001 |
| 2 | 저축 | 0.017927 |
| 3 | 저축은행 | 0.011776 |
| 4 | 대출 | 0.005572 |
| 5 | 영업 | 0.004901 |
| 6 | 정지 | 0.004474 |
| 7 | 금융 | 0.0041 |
| 8 | 영업정지 | 0.004021 |
| 9 | 예금 | 0.003098 |
| 10 | 제일 | 0.002096 |
| 11 | 토마토 | 0.002008 |
| 12 | 부실 | 0.001878 |
| 13 | 경영 | 0.001433 |
| 14 | 대주주 | 0.001329 |
| 15 | 예금자 | 0.001233 |
| 16 | 당국 | 0.001215 |
| 17 | 7개 | 0.001189 |
| 18 | BIS | 0.001163 |
| 19 | 비율 | 0.001094 |
| 20 | 매각 | 0.00105 |

로 정렬될 수 있다. 그리고 $s^*$값이 높은 상위 10개의 문단을 리턴 한다. 이러한 상위 10개의 문단은 토픽 키워드와 가장 잘 매칭되는 문장들로 토픽을 잘 설명하는 정보원으로 볼 수 있다. 먼저 $s^*$을 계산하기 위해, 다음 가설을 정의한다.

가설 : 토픽 모델링을 통해 얻어진 토픽 키워드들과 확률 값들을 사용하여, 주어진 토픽에 대한 각 문단의 확률 값을 계산한다.

위의 가설을 바탕으로 다음과 같이 생성확률모델을 제안한다.

$$P(t|s) = Z(t)\prod_{w\in t}P(w|s) = Z(t)\prod_{w\in t}P_s(w) \quad (5)$$

위의 수식에서 $P(t|s)$는 문단 s에 대한 토픽 t의 확률 값을 나타낸다. t는 토픽과 연관된 키워드들의 집합을 의미한다. 그리고 Z(t)는 t의 총 키워드 개수로 노멀라이즈 하기 위한 파라미터로 사용된다. 각 토픽 키워드 당 s에 대한 토픽 키워드 w의 확률은 독립확률변수이기 때문에 각 w의 확률 값의 곱으로 계산된다. 그리고 각 w의 확률 값 $P(w|s)$는 다음과 같이 근사치를 계산할 수 있다.

$$P_s(W) = \frac{g(w)}{\sum_{w'\in t\cap text(s)} g(w')} \quad (6)$$

즉, <Table 1>과 같이 토픽이 주어진다고 가정하자. 각 토픽 키워드 w와 그 확률 값 $P(w)$라고 정의하면, 토픽의 모든 확률 값을 더해서 각 확률 값으로 나누어준다. 예를 들면, 20개의 토픽 키워드의 전체 확률 값은 0.091558이고, '은행'의 확률 값은 0.020001로 전체 확률 값에서 나누면 $P_s$('은행') = $\frac{0.020001}{0.091558} = 0.2185$이다. 마찬가지로 $P_s$('저축') = $\frac{0.091558}{0.017927} = 0.1958$로 계산된다. 토픽 t와 문단 s의 연관성을 수치($s^*$)로 계산하기 위해, t에 대한 $P(s|t)$의 조건부 확률 값을 계산할 수 있다.

$$s^* = \mathrm{argmax}_s P(s|t) \quad (7)$$

$$= \mathrm{argmax}_s \frac{P(t|s)P(s)}{P(t)} \quad (8)$$

$$= \mathrm{argmax}_s \frac{P(s)}{P(t)} P(t|s) \quad (9)$$

$$= \mathrm{argmax}_s P(t|s) \quad (10)$$

$$= \mathrm{argmax}_s Z(t)\prod_{w\in t}P(w|s) \quad (11)$$

$$= \mathrm{argmax}_s Z(t)\prod_{w\in t}P_s(w) \quad (12)$$

$$= \mathrm{argmax}_s \prod_{w\in t}P(w) \quad (13)$$

$$= \mathrm{argmax}_s \prod_{w\in t}\log P_s(w) \quad (14)$$

$P(s|t)$는 Bayes' theorem에 의해 식 (8)로 치환될 수 있다. 식 (9)에서 P(s)는 알려지지 않은 값이기 때문에 t에 대해 모든 P(s)들을 Uniform Distribution으로 가정하여 $\frac{P(S)}{P(t)}$을 생략할 수 있다. 식 (10)에서 $P(s|t)$는 앞에서 제안한 생성확률모델을 기반으로 하여 식 (5)를 사용해서 식 (12)로 치환할 수 있다. 또한 Z(t)는 s에 독립변수이기 때문에 생략한다. 식 (13)에서 확률분포함수가 곱셈 꼴일 때 미분 계산의 편의를 위해 Log Likelihood 함수로 치환하여 계산한다. 그 이유는 로그 함수는 Monotone Increasing하기 때문에 Likelihood에서 극값을 가지는 위치와 Log Likelihood에서 극값을 가지는 위치가 같기 때문이다. 따라서 토픽 t가 주어지면, 문단 s의 연관성은

$$s^* = \mathrm{argmax}_s \sum_{w \in t} \log P_s(w) \qquad (15)$$

으로 구할 수 있다. 이것은 문단 s에 포함되는 토픽 키워드들의 확률 값에 로그를 취한 총합을 최대로 만드는 s를 구하는 것이다.

### 3.6 사회문제 키워드 비주얼라이제이션

사회문제 키워드 도출 시스템은 정량적으로 사회문제에 대한 수요 조사를 수행할 때 유용한 자료를 제공한다. 또한 전문가들에게 사회문제에 대한 기초 자료를 제공하여 사회문제를 구체화하는 데 꼭 필요한 도구이다. 그러나 사회문제 키워드를 시각화하지 않으면 방대한 자료를 검토하고 사회현상을 이해하는 데 큰 어려움과 함께 제대로 활용되지 않을 가능성이 높다. 따라서 본 연구에서는 전문가들이 직관적으로 쉽게 사회문제를 이해할 수 있도록 비주얼라이제이션 웹 페이지를 설계하고 구현한다.

## 4. 시스템 성능 평가

### 4.1 필터링 결과 분석

먼저 글자 수와 인용부호가 필터링에서 실제로 중요한지를 확인하기 위해 백 30만 건의 뉴스기사 중에서 임의로 32,790건의 뉴스기사를 샘플링 하였다. 1차 필터링을 수행한 결과, 18,939건의 뉴스기사들이 사회문제를 다루는 기사로 분류되었다. 1차 필터링에서 제거된 뉴스기사의 수는 샘플링한 뉴스기사에서 42%을 차지한다. <Table 2>는 1차 필터링에서 제거된 뉴스기사의 제목의 일부를 보여준다. Table에 속하는 대부분의 뉴스기사들이 사회문제를 다루지 않는 뉴스기사임을 알 수 있다.

<Table 2> Examples of News Headlines Removed in Primary Filtering

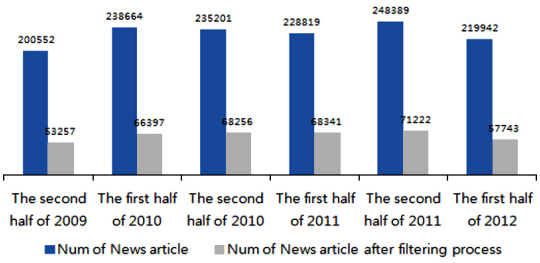
| Num | News headline |
|---|---|
| 1 | 거식증 극복하고 연기자로 새 출발하는 '포동이' |
| 2 | 비포 사이즈 모델 겸 연기자 |
| 3 | 빙판을 가르던 갈라의 여왕에서 코치 겸 안무가로 |
| 4 | 살풀이 속풀이와 걷기 전도사, 한의사 |
| 5 | 시누이들은 얄밉기만 하다고?! |
| 6 | 전 재산 29만원, 전두환 전 대통령 손녀딸 결혼하던 날 |
| 7 | 에어컨 없이도 시원하게! 청량한 침실 꾸밈 아이디어 |
| 8 | 공간 디자이너 권순복의 프로방스 스케치③ |
| 9 | 아스토리아호텔 경영대표이사 이경수 |
| 10 | 멱살을 낚아채듯 건져 올린 시의 짜릿한 매력 |

또한 2차 필터링의 성능 평가를 위해 1차 필터링 후에 남겨진 18,939건의 뉴스기사를 사용하였다. 2차 필터링 후에 전체 뉴스기사(32,790) 중에서 27%에 해당하는 8,939건의 뉴스기사가 제거되었고, 총 10,000건의 뉴스기사가 사회문제를 다루는 기사로 남았다. <Table 3>은 2차 필터링 후에 남겨진 뉴스기사 제목의 일부이다.

<Table 3> Examples of News Headlines Removed in Secondary Filtering

| Num | News headline |
|---|---|
| 1 | 국내 기업 최고경영진 100명 중 여성은 불과 2명 꼴, 여성 고위직 비율 아시아에서 꼴찌 |
| 2 | R&D 조세지원 1% 줄이면 … 국가경제 10년간 4조원 손실 |
| 3 | 서울시내 보도면적, 차도의 10% 이하 수준 |
| 4 | 서울 6월 평균기온, 기상 관측 이래 가장 높아 |
| 5 | 서울시 자동차 공회전 특별점검 실시 |
| 6 | 한국의 실질 최저 임금 … 프랑스의 30%도 안돼 |
| 7 | 청계천 복원 효과 … 서울 미세먼지 농도 줄었다. |
| 8 | 청주-청원 통합시 명칭 '청주시' 논란 … 청원군 주민들 반발 |
| 9 | 상반기 인터넷 쇼핑 이슈 살펴보 … 반값제품 등 '실속'이 대세 |
| 10 | "무상급식 이후 교육시설 예산 절반 깎여" … 교총, 서울교육재정분석 |

<Table 2>와 비교했을 때, <Table 3>에서는 연예 및 스포츠 뉴스기사와 같이 직접적으로 사회문제를 다루지 않는 뉴스기사들이 줄어들었음을 알 수 있다.




<Figure 6> Number of News Headlines that Handling Social Issue after Primary and Secondary Filtering

<Figure 6>은 2009년 6월부터 2012년 7월까지 총 3년 동안의 뉴스 데이터를 월별로 1, 2차 필터링을 수행한 후 얻은 통계 데이터를 보여준다. 그림에서 보는 것처럼, 필터링 후에 뉴스 기사 수는 전체 뉴스기사 데이터의 24%~30% 정도이고 사회문제를 다루는 뉴스기사는 전체 $\frac{1}{3}$에 해당하는 것을 알 수 있다.

## 4.2 토픽 모델링 결과 분석

월별 뉴스기사 데이터가 입력으로 주어지면 토픽 모델링의 결과는 토픽 별로 그룹핑 된 키워드들과 그 확률 값들이 출력된다. <Figure 7>에서 보는 것처럼, 62번째 토픽은 '인구', '고령', '고령화', '저출산' 등 20여 개의 키워드들과 그 확률 값으로 구성된다. 그리고 '인구', '고령화', '저출산', '2050년'등은 이 토픽에서 가장 높은 확률 값을 가지는 키워드들이다. 이러한 키워드들을 통해 현재 국내에 고령화와 저출산에 대한 우려가 커지고 있음을 알 수 있다. 이러한 키워드들을 통해 저출산으로 2050년에 인구가 급격히 줄어들거나 군인수가 부족하여 국방에 심각한 위기가 나타날 수 있고 동남아에서 이주민들이 증가할 것이라는 연관 키워드들도 포함되어 있음을 알 수 있다. 이러한 월별 토픽 키워드들을 조사하여, 전체 백만 건이 넘는 뉴스기사에서 다루고 있는 국내의 사회문제에 대해서 각계각층의 다양한 전문가들과 함께 국내의 사회문제를 <Table 4>에서 정리하였다.

```
Topic 62th:
        인구 0.005318458448827496
        고령 0.003759446340712233
        고령화 0.0028367657052970775
        저출산 0.0016595524808018793
        2050년 0.001277753597181815
        실탄 0.0010550375817367777
        지출 0.0010232210081017722
        복지 9.91404434466767E-4
        사격 9.595878608317616E-4
        연금 9.277712871967563E-4
        원권 9.277712871967563E-4
        베이비 8.323215662917402E-4
        우리나라 8.005049926567349E-4
        국방부 8.005049926567349E-4
        2020년 8.005049926567349E-4
        베이비붐 8.005049926567349E-4
        붐 8.005049926567349E-4
        세대 7.686884190217294E-4
        서바이벌 7.050552717517188E-4
        이주 6.41422124481708E-4
Topic 63th:
        출연 0.0025046204305970736
        방송 0.0021811020565180713
        영화 0.001916405204998888
        배우 0.0016811191147596137
        MBC 0.0016517083534797045
```

<Figure 7> Example of Topic Modeling

<Table 4> Example of Extracted Social Issue using Topic Modeling

| Volume Number | Topic | | Keywords | Social Issue | Num of News Articles |
|---|---|---|---|---|---|
| | Category | Division | | | |
| 1 | 정치, 입법 | 선거 | 의원, 공천, 총선, 새누리당, 민주당, 통합, 대선, 통합진보당 부정경선, 안철수 | 민주주의와 선거(2010 지방선거, 2011 서울시 장재보궐선거, 총선) | 2,535 |
| | | 미디어법 | 직권상정, 본회의, 미디어법, 방통위, 주파수, 종편 | 대기업 및 신문사의 방송진출 관련 법 | 419 |
| | | 무상급식 | 서울시, 주민투표 | 서울시 초등학교 전면 무상급식 실시 투표 | 2,259 |
| | | 지역개발 사업 | 4대강, 세종시 | 4대강 사업과 세종시 개발사업의 절차적 합리성 문제 | 5,704 |
| | | 한미FTA | 국회, 비준동의안, 번역오류, 한나라당, 국회의장 | 한미FTA 비준동의안 한나라당 단독처리 | 4,425 |
| | | | | | 15,342 |
| 2 | 국방, 외교 | 북한체제 | 김정은, 김정일 | 2011. 12. 김정일 사망 이후 김정은 체제로 전환 | 1,557 |
| | | 한일 관계 갈등 | 위안부, 왕실의궤, 정신대, 역사교과서, 동해, 표기, 간도 | 독도, 동해, 간도, 역사교과서, 위안부 문제로 인한 한일 관계 갈등 | 408 |
| | | 남북관계 | 회담, 6자회담, 도발, 해군, 미사일, 연평도, 서해교전 | 서해교전, 천안함 사태, 연평도 포격 등 북한과의 군사적 충돌 | 10,029 |
| | | 해적 피랍 | 석선장, 소말리아, 삼호드림호, 한진해운, 해적선, 청해부대 | 소말리아 해역에서 해적에 의한 피랍사건 증가 | 488 |
| | | 대북사업 | 이산가족, 개성공단사업, 금강산 관광, 남북적십자회담 | 개성공단사업, 금강산 관광사업, 이산가족상봉 등 대북사업 | 1,132 |
| | | | | | 13,614 |
| 3 | 국제 | 글로벌 금융위기 | 금융위기, 그리스, 부채, 유럽, 스페인, 구제금융 | 유럽발 재정금융위기 확산(그리스 부채 문제) | 3,857 |
| | | 테러위협의 증가 | 테러, 탈레반, 이슬람무장단체, 파키스탄, 자폭테러, 알카에다 | 중동지역 내전과 테러로 인한 민간인 사망 증가 | 276 |
| | | 국제질서의 다극화 | 국제회담, G20, 정상회담 | 글로벌 거버넌스의 확대(UN, G20 등) | 2,531 |
| | | 민족갈등 | 위구르, 인종차별, 이주노동자인권, 가자지구, 남아공, 흑백갈등, 티베트, 중동 | 민족갈등 심화(위구르-한족, 남아공 흑백갈등, 이스라엘-팔레스타인 등) | 553 |
| | | 반정부시위 민주화운동 | 태국, 방콕, 유혈사태, 키르키즈, 리비아, 카다피, 이집트, 무바라크, 미얀마, 아웅산수치, 군부 | 독재국가의 반정부 민주화시위 확산(태국, 리비아, 이집트, 미얀마) | 2,507 |
| | | 핵확산금지 | 비핵화, 핵안보, 오바마, 유엔안보리정상회의, 이란, 6자회담 | 국제사회의 핵무기 개발 제재 강화(북한, 이란 등) | 4,552 |
| | | | | | 14,276 |
| 4 | 경제, 금융 | 부동산 | 종부세, 주택사업, 보금자리, 분양, 임대, 용산, 재개발, 재건축 | 불안정한 부동산 경기(서울시 전세값 폭등, 보금자리사업 무산) | 1,711 |
| | | 금융비리 | 저축은행, 영업정지, 구속, 부실경영, 대출, 펀드, 투자 | 저축은행 부실경영가계대출에서 제2금융권 비중 증가 | 1,217 |
| | | 양극화 | 중소기업, 협력, 대기업, 상생, 포스코, 삼성전자, 대형마트, 슈퍼마켓, SSM | 대기업-중소기업(대형마트-동네슈퍼/재래시장)의 격차 심화 | 1,223 |
| | | 세제 개편 | 부자, 과세, 감세, 전세담보보증금과세, 직접세, 조세형평성, 탈세 | 부자 감세, 조세형평성 문제 | 171 |
| | | FTA 체결 | 한미FTA, 한-EU FTA, 자유무역협정, 세계시장 | 미국, 유럽연합과의 자유무역협정 체결 및 한-중-일 무역공동체 추진모색 | 1,958 |
| | | | | | 6,280 |
| 5 | 사회, 복지 | 노사갈등 | 한진중공업, 민주노총, 쌍용차, 노사갈등 | 노사갈등과 파업사태(한진중공업, 쌍용차, 화물연대, 버스노조 등) | 2,465 |
| | | 일자리 및 근로정책 | 고용유발예산 확대, 취업정보센터, 공공근로사업, 일자리사업, 비정규직, 취업률저하, 청년인턴, 비정규직법, 최저임금 | 정규직 취업률 저하, 비정규직 및 인턴 취업률 증가 | 1,174 |

| | | | | | |
|---|---|---|---|---|---|
| 5 | 사회, 복지 | 여성복지 신장 | 여성일자리증가, 육아휴직제도, 기업어린이집 | 저출산 극복을 위해 일과 가정을 모두 지원하는 국가적, 기업적 방안 모색 | 557 |
| | | 저출산/ 고령화 | 저출산, 가족계획, 학령인구, 양육, 복지정책, 육아, 보육시설, 어린이집 | 출산률 저하와 평균수명 연장에 따른 노동시장 고령화, 노인일자리확충사업 수요 | 188 |
| | | 성범죄 | 아동대상성범죄 증가, 김길태, DNA, 조두순 | 성범죄 및 아동성범죄 증가 | 600 |
| | | 인권문제 | 단속, 이주노동자, 출입국체류자, 네팔, 탈북자, 장애인, 용산, 철거민, 세입자, 재개발사업 | 사회적 약자의 인권문제(탈북자, 장애인, 철거민, 이주노동자, 영화 도가니, 용산참사) | 221 |
| | | 자살률 증가 | 카이스트, 자살, 학교폭력 | 10, 20대의 자살문제 | 266 |
| | | | | | 5,471 |
| 6 | 교육 | 대학입시 | 대입, 사교육, 교과부, 교육청, 수능, 전국연합학력평가, EBS, 문제지, 사전유출, 논술, 면접, 학생부(0908), 지역균형선발, 수시, 정시, 외국어, 특수학급, | 교육 양극화 문제와 대학입시개선 노력(입학사정관제확대, 수능제도개선, 공교육 강화) | 739 |
| | | 학교폭력 | 학교폭력, 상담, 경찰, 가해, 피해, 상담, 안전, 자살, 신고, 조사, 교사폭행 | 학생 간 폭행, 학생의 교사 폭행 증가 | 654 |
| | | 교육정책 | 교육, 사업, 지원, 조례, 인권, 대학등록금, 특성화고, 고교선택제, 교육감, 교육계, 공모제, 교육비리, 교사, 체벌, 금지, 대학정보 공개, 학생인권조례 | 서울시 학생인권조례, 체벌금지 논란 | 238 |
| | | 사교육 | 사교육, 학원, 대학, 지원, 고교, 입학, 평가, EBS, 영어유치원, 빈부격차, 특목고, 공교육 | 지나치게 높은 사교육 의존도로 빈부격차와 사회갈등 유발, 가계지출에서 사교육비가 차지하는 비율이 높아 소비 위축 | 310 |
| | | 대학 등록금 | 학자금, 총장, 전형료, 동결, 입시, 전공, 총장, 지원 | 대학생 학자금 대출 증가, 2011. 06 반값등록금 시위 | 1,464 |
| | | | | | 3,405 |
| 7 | 과학기술 | 스마트폰 기술 | 삼성전자, 애플, 특허, 소송, 스마트폰 | 삼성전자와 애플의 글로벌 소송, 국내 스마트폰 기술 해외 진출 | 150 |
| | | 사이버 테러 | 농협, 현대캐피털, 해킹, 디도스, 선관위, 북한발, 중국발 | 사이버테러에 의한 국가경제 손실 및 개인정보 대량유출(네이트, 농협, 현대캐피털) | 465 |
| | | 바이오 테크놀로지 | 임상시험, 관절염, 치료제, 배아줄기세포, 노화, 임상, 불임연구, 성체줄기세포, 각막 | 줄기세포 치료제 개발로 실명, 뇌질환, 불임 등 치료 예상 | 196 |
| | | 우주과학 | 나로호, 우주선, 인공위성, 러시아 | 나로호, 한국형 발사체, 나로호 발사실패 | 323 |
| | | | | | 1,134 |
| 8 | 환경, 에너지 | 원자력 발전소 | 후쿠시마, 원자력, 방사능, 핵폐기물, 세슘, 고리원전 | 원자력발전의 위험성(후쿠시마원전폭발로 인한 일본 및 인근 환경오염, 인체유해, 각종질병 유발/고리원전 안전성 의심) | 1,229 |
| | | 에너지 자원부족 | 정전, 수요, 예비, 전력, 전기, 가격, 인상, 석유 | 예비전력 문제(2011년 9월 15일 정전사태), 에너지자원 부족문제(원유수입, 석유의존율) | 321 |
| | | 자연재해 | 쓰나미, 지진, 인도네시아, 칠레, 일본, 화산, 아이티 | 대규모 지진, 쓰나미로 인한 인명, 각국 피해 속출(인도네시아, 아이티, 칠레, 일본) | 12,031 |
| | | 지구온난화 | 온실가스감축, 녹색성장, 탄소, 배출량, 에너지, LED | 온실가스배출 감축 노력(LED 기술, 의류 소재 혁신 등)/지구온난화로 인한 이상기후 | 701 |
| | | 환경오염 | 수질오염, 급수, BOD, 4대강, 낙동강, 기름유출, 해양오염, 대기중미세먼지, 토양오염, 석면, 구제역침출수 | 환경오염의 증가(4대강 수질오염, 토양오염, 기름유출로 해양오염) | 1,982 |
| | | 생태계의 변화 | 반달가슴곰, 지리산, 여우, 서울동물원, 멸종위기, 산양, 야생동물, 거래다랑어, 서식 | 야생생물종의 감소와 멸종위기(여우, 산양, 반달곰 등) | 78 |
| | | 친환경 대체에너지 개발 | 우뭇가사리양식, 우뭇가사리에탄올추출기술, 친환경바이오에탄올, 신재생에너지, 태양열, 수소연료전지, 태양광, 풍력발전사업, 전기차 | 대기업의 녹색에너지사업으로 온실가스감축, 태양광, 풍력, 연료전지 등 신재생에너지를 주축으로 한 대체 에너지 개발 노력 | 667 |
| | | | | | 17,009 |
| 9 | 건강, 의료 | 신종플루 | 신종플루, 호흡기확산, 보건, 백신, 타미플루, 인플루엔자, 바이러스, 사망자 | 신종바이러스 발병 | 1,118 |
| | | 전염병 | 구제역, 농가, 방역, 아이티, 콜레라, AI, 조류인플루엔자 | 구제역과 조류인플루엔자 등 가축전염병 유행 | 2,502 |
| | | 고령화 | 건강, 노인, 보험, 예산, 적자, 의료, 노인질환 | 노인의료비 급증으로 인한 건강보험예산 적자 | 28 |
| | | | | | 3,648 |

### 4.3 매칭 알고리즘 결과

실제 본 연구에서 제안한 방안이 각 문단과 토픽 간의 연관성을 정확하게 나타내는 지를 평가하기 위해, <Table 1>의 토픽을 사용하여 상위 10개의 관련 있는 문단을 <Table 5>에서 정리하고 분석하였다. <Table 5>는 저축은행 부실경영 토픽과 관련된 상위 10개의 문단을 보여준다. 이러한 문단들은 토픽 모델링 통해 얻은 토픽 키워드들을 잘 설명해 주고 있다. 특히 토픽과 관련하여 구체적으로 정보를 제공하는 문장은 밑줄로 표현하였는데, 사실 관계 확인과 함께 문제점이 무엇인지를 잘 설명하고 있다. 예를 들면, 저축은행 부실경영 토픽과 가장 관련이 깊은 문단은 다음과 같다.

> 토마토**저축은행** 계열사인 토마토2**저축은행**(부산)은 BIS 비율이 6.26%로 **영업정지** 조치를 면했다. 7개 **저축은행** 외에 6개 **저축은행**은 BIS 비율이 5%에 미달하거나 부채가 자산보다 많은 것으로 드러났지만 **영업정지** 조치는 피했다. **금융위**는 "**대주주** 증자와 자산 **매각** 등 **경영**개선계획의 실현 가능성을 인정해 최대 1년까지 자체 정상화를 추진토록 결정했다"고 밝혔다.

위의 문장에서 토픽 키워드들은 볼드체로 표시하였다. 이러한 문단을 통해 사용자는 사회문제 키워드들을 좀더 잘 이해할 수 있게 된다.

### 4.4 사회문제 키워드 비주얼라이제이션

사회문제 키워드 시각화를 위해 먼저 데이터베이스 서버에 토픽 키워드와 연관된 문단을 저장한다. 그리고 PHP을 사용하여 클라이언트가 요청한 데이터를 가져오고 웹 브라우저에서 비주얼라이제이션하게 된다. 사회문제 키워드 비주얼라이제이션은 자바스크립트와 HTML5 캔버스를 사용하여 구현하였다(Fulton et al., 2012).


<Figure 8> Visualization of Social Issue Keywords in Main Webpage

<Figure 8>는 사회문제 키워드 비주얼라이제이션 메인 웹 페이지를 보여주고 있다. 웹 페이지 하단에 있는 버튼을 사용하여 연도와 달을 선택할 수 있다. 예를 들면, 2011년 8월을 선택하면 2011년 11월까지 약 4개월 동안에 이슈가 되었던 토픽들의 순위가 나타난다. 본 시스템은 가독성을 높이기 위해 상위 15위까지의 토픽들만 제공한다. 토픽의 순위는 토픽에 포함되는 뉴스기사의 개수를 사용하여 정하게 된다. 이것은 이슈화가 많이 되었던 토픽은 여러 언론사에서 많은 뉴스기사를 작성하였기 때문이다. 이러한 비주얼라이제이션을 통해 전문가들은 특정 토픽의 이슈성이 시간에 따라 증가 또는 감소하는 지를 쉽게 파악할 수 있다. 또한 특정 시점에 이슈화된 사회문제들을 파악하기에도 용이하다. 또한 <Figure 8>의 메인 웹 페이지에서 특정 토픽을 클릭하게 되면 해당 토픽에 대한 키워드들과 연관된 문단들이 나타난다. 예를 들면, 2010년 8월의 13위에 랭크 된 토픽을 클릭하면 <Figure 9>와 같은 토픽 키워드들과 문단들이 나타나게 된다.

<Table 5> Example of Paragraph Matching to "저축은행 부실경영"

| Rank | $s^*$ | Paragraph |
|---|---|---|
| 1 | 47.1947 | 토마토저축은행 계열사인 토마토2저축은행(부산)은 BIS 비율이 6.26%로 영업정지 조치를 면했다. 7개 저축은행 외에 6개 저축은행은 BIS 비율이 5%에 미달하거나 부채가 자산보다 많은 것으로 드러났지만 영업정지 조치는 피했다. 금융위는 "대주주 증자와 자산매각 등 경영개선계획의 실현 가능성을 인정해 최대 1년까지 자체 정상화를 추진토록 결정했다"고 밝혔다. |
| 2 | 46.703 | [사설] 저축은행 혼란 더 없도록 철저한 마무리를, 저축은행 구조조정의 뚜껑이 열렸다. 금융위원회는 어제 임시회의에서 토마토 등 7개 부실 저축은행에 대해 6개월간 영업정지 결정을 내렸다. 금융당국이 전면 경영진단에 들어간 지 두 달여 만의 일이다. 퇴출 대상에는 국제결제은행(BIS) 기준 자기자본비율이 1% 미만이고 부채가 자산을 초과해 정상 영업이 어려운 은행 등이 포함됐다. |
| 3 | 42.383 | <사설> 저축은행 '구조조정 일단락' 믿을 수 있나","토마토저축은행 등 7개 부실 저축은행이 영업 정지됐다. 금융위원회는 어제 경영진단 결과 국제결제은행(BIS) 기준 자기자본비율이 5% 미만이거나 부채가 자산보다 많은 13개 저축은행 가운데 7개 저축은행에 영업정지를 명령하고 나머지 6개 저축은행은 자체 경영정상화를 추진토록 했다고 밝혔다. 금융위의 이번 조치는 올 상반기 8개 저축은행을 영업 정지시킨 데 이은 2차 구조조정이다. 85개 저축은행을 대상으로 경영진단을 실시해 13곳으로부터 경영정상화 계획을 받았는데 7개 저축은행은 정상화 계획이 퇴짜를 맞아 영업정지를 당한 것이다. |
| 4 | 42.2723 | 자산 2조 원이 넘는 대형 금융회사인 토마토저축은행과 제일저축은행 등 7개 저축은행의 영업이 정지됐다. 국제결제은행(BIS) 기준 자기자본비율이 1% 미만이거나 부채가 자산보다 많아 정상적인 경영이 힘들다는 판단에 따른 조치다. 계열 자산 기준으로 저축은행업계 1위인 부산에 이어 독립법인 자산 기준 2위인 토마토와 3위인 제일이 사실상 퇴출의 길을 걷게 됐다. 이로써 올 들어 영업정지 대상에 오른 저축은행은 모두 16개사로 늘었다. |
| 5 | 41.7929 | 19일 금융당국에 따르면 동일인 대출한도 위반은 이번에 영업 정지된 7개 저축은행뿐 아니라 국제결제은행(BIS) 자기자본비율 5% 이상의 정상 저축은행에서도 공공연하게 이뤄진 것으로 나타났다. 금융감독원은 이러한 위반이 경영진단과정에서 적발된 불법행위중 대부분을 차지한다고 설명했다. |
| 6 | 39.4272 | 하지만 저축은행들은 마치 이 규정이 존재하지 않는 듯이 이를 앞 다퉈 위반했다. 금융권 관계자는 "금감원이 저축은행에 대해 지속적으로 감독을 수행했음에도 불구하고 눈 뜬 장님처럼 이를 적발하지 못한 것은 문제가 있다"며 "결국 금융당국의 솜방망이 처벌과 온정주의적 감독 행태가 저축은행의 부실을 키웠다고밖에 할 수 없다"고 말했다. 대주주 대출도 일부 적발됐다. 금감원 고위 관계자는 "부산저축은행처럼 대주주가 운영하는 사업장에 거액을 몰아주거나 차명계좌와 특수목적법인(SPC)을 대놓고 불법 영업을 한 사실은 발견되지 않았다"며 "하지만 우회적으로 대주주와 연관된 사업장에 불법 대출한 경우가 있었다"고 말했다. 대주주 대출과 한도위반 대출은 손실가능성이 큰 것으로 간주해 충당금을 더 쌓아야 하고 충당금 적립액만큼 자기자본은 감소한다. 금감원 경영진단 이후 BIS 자기자본비율이 급전직하한 것은 바로 이 때문이다. |
| 7 | 37.6578 | [사설]저축은행 비리, 수사로 다 도려내라","올해 2월 부산저축은행과 대전저축은행의 영업이 정지되면서 부실저축은행에 대한 1차 퇴출작업이 시작됐다. 3월 금융당국은 불법행위를 한 저축은행 대주주에게 최고 10년의 징역형을 내리겠다는 감독강화 방안까지 내놨다. 그러나 업계 2위인 토마토저축은행은 공시지가 12억 원짜리 땅을 담보로 978억 원을 토마토저축은행 회장의 고교 후배에게 대출해준 것으로 뒤늦게 밝혀졌다. 자산보다 부채가 4419억 원이나 많은 이 은행은 올해 7월 직원들에게 보너스 잔치를 벌였다. 토마토저축은행을 포함해 18일 영업 정지된 7개 저축은행은 한사람에게 자기자본의 20% 이상을 대출할 수 없다는 저축은행법을 밥 먹듯 어겼다. |
| 8 | 37.4676 | '불법대출' 이용준 제일저축은행장 체포, 영업 정지된 7개 저축은행의 비리 혐의에 대한 수사가 속도를 내고 있다. 수사 착수 후 처음으로 저축은행 고위급 임원 2명이 체포된 가운데 앞으로도 경영진과 대주주 중에서 구속수사 대상자가 속출할 것으로 보인다. 당분간은 동일인 대출한도를 초과한대출과 대주주신용공여, 부실대출 등이 수사의 핵심 '타깃'이 될 전망이다. |
| 9 | 37.1173 | 이번에 영업 정지된 에이스저축은행과 토마토저축은행은 BIS 비율이 1년 사이에 8.51%와 9.45%에서 -51.10%와 -11.47%로 약 60% 포인트와 20% 포인트씩 급락했다. 나머지 영업정지 저축은행들도 BIS 비율이 10% 포인트 넘게 하락해 마이너스로 떨어졌다. 이밖에 몇몇 저축은행은 불법으로 경비를 유용한 사실도 드러났다. 금융당국은 불법대출에 대해선 법률검토 등을 거쳐 조속히 검찰에 수사 의뢰할 예정이다. |
| 10 | 34.2189 | 무엇보다도 경영정상화 계획을 요구 받은 대상이 13곳에 그친 것은 그 동안 알려진 저축은행의 부실 상황에 비춰볼 때 너무 적다. 이번 조치에 앞서 금융위가 저축은행에 대한 국제회계기준 적용을 5년간 유예하고, 부동산 프로젝트 파이낸싱 채권 1조 7000억 원어치를 매입해줘 BIS 비율이 높아진 덕분이라면 정부가 구조조정 대상 축소를 위해 사전 정지작업을 했다는 얘기가 된다. 퇴출 수를 줄이기 위해 부실 저축은행에 시간을 벌어주고 부실을 떠안아주는 정책이 효과가 없다는 것은 과거 저축은행 구조조정 사례가 잘 말해준다. BIS 비율 5% 이상인 저축은행에 공적 자금을 투입해 자본확충을 지원키로 한 것도 부실 문제의 근본적 해결에 얼마나 도움이 될지 의문이다. 6개월~1년 뒤에 영업정지 조치가 다시 나오지 않을까 우려된다. |


<Figure 9> Visualization of “줄기세포” topic

<Figure 9>는 줄기세포 토픽과 연관된 키워드들과 문단들을 보여줌으로써 전문가들이 쉽게 사회문제 이슈를 파악할 수 있도록 하는 기능을 제공한다. 이러한 비주얼라이제이션 웹 페이지는 PC 버전과 스마트패드 버전으로 제공된다. 사용자는 http://dslab.snu.ac.kr/demo.html 웹 페이지를 방문하여 사회문제 키워드 비주얼라이제이션을 사용할 수 있다.

### 4.5 사회문제 키워드 분석

본 논문에서 제안한 사회문제 키워드 추출 시스템을 통해 2009년 6월부터 2012년 7월까지 총 3년 동안의 사회 이슈를 살펴보았다. 먼저 가장 관심이 높은 토픽을 찾기 위해 토픽 모델링을 사용한 후 토픽에 어사인 된 뉴스기사의 개수를 조사하였다. 토픽에 포함된 뉴스기사 수가 많을수록 관심이 높은 토픽으로 간주하였다. 상위 10개의 토픽으로는 일본 대지진, 천안함, 한미FTA, 남북관계, 세종시, 연평도, G20경제 토론, 4대강, 무상급식, 글로벌 금융위기 등이었다. 상식적으로 이러한 토픽들이 최근 3년 동안 큰 화제거리였음을 비추어볼 때 실제 제안된 시스템의 결과도 크게 다르지 않았다. 그리고 이러한 토픽들은 크게 두 가지로 분류할 수 있다. 일본 대지진이나 천안함처럼 특정 시점에 집중적으로 이슈화가 되거나 아니면 FTA나 남북관계처럼 오랜 시간에 걸쳐 지속적으로 이슈화가 되는 토픽들임을 알 수 있었다. 반면에 관심이 적은 토픽으로는 치매신약개발, 종량제, 희귀성분, 피임약을 일반약으로 분류하는 것에 대한 논란, 석면 위험성 등이었으며 이러한 토픽들은 특정 시점의 짧은 시간 동안 이슈화가 되는 것이 큰 특징이었다.

이와 같이 시간 순서에 따라 월별 토픽들에 속해 있는 뉴스기사의 수를 측정함으로써 일본 대지진 또는 삼호 주얼리호 피랍 같은 이벤트를 쉽게 찾을 수 있었다. 예를 들면 일본 대지진의 경우에는 2011년 3월부터 6월까지 집중적으로 이슈화가 되었다. 한편 주기적으로 발생하는 이벤트도 발견되었다. 예를 들면, 수능시험 같은 이벤트는 매년 11월~12월에 이슈화가 되고 있음을 결과 분석을 통해 알 수 있었다.

반면에 지속적으로 관심도가 증가하거나 감소하는 토픽을 찾기 어려웠다. 그 이유는 특정 시점에서 이슈가 등장하여 짧은 시간 내에 정점까지 올라가고 그 이후로는 급속도로 관심도가 줄어드는 형태가 일반적이었다. 예를 들면, 일본 대지진의 경우에는 2011년 3월에 11,602건의 뉴스기사들이 나타났고 4월부터 6월까지는 매달 수백 개 안팎의 뉴스기사가 등장했고 7월 이후에는 뉴스기사 수가 미미한 수준으로 관측되었다. 한편 남북관계와 같은 이슈는 지속적으로 관심을 받고 있지만 시간에 따라 관심도의 증가 내지 감소폭의 변동이 심해서 이러한 이슈가 지속적으로 증가하는지 또는 감소하는지 파악하기 힘든 것으로 조사되었다. 이러한 특성은 이

벤트 위주의 보도에 치중하는 뉴스기사의 속성을 잘 반영한다고 할 수 있다.

또한 하나의 토픽의 발생이 다른 토픽을 유도하는 경우도 종종 발견되었다. 예를 들면, 2011년 3월부터 2011년 6월까지 일본 대지진이 이슈화가 되었고 동년 9월부터 3개월 동안 일본 원전 방사능이 크게 이슈화되었다. 이처럼 이벤트 간에 순차적인 관계를 고려하여 두 토픽간의 관계를 파악할 수 있다.

끝으로 시계열에 따른 기사빈도수를 조사했을 때, 금융위기와 같이 고도로 복잡한 문제는 특정기간 동안 다수의 관련 키워드들이 나타난다. 예를 들면, 금융위기는 평균적으로 매달 약 150개 정도의 금융위기관련 기사들이 나타났다. 반면에 장기적으로 진행되는 사회문제는 빈도수가 낮고 연속적으로 나타나지는 않지만 장기적으로 꾸준히 나타난다. 대표적인 예가 기후변화에 관한 키워드들이다. 마지막으로 예측 불가능한 대형 사건은 매우 짧은 기간에 매우 많은 기사가 게재된다. 예를 들면, 아이티 대지진, 일본 대지진 등의 기사들이며 사건 직후 게재된 뒤 그 이후에는 빈도가 크게 줄었다.

## 5. 결론 및 제언

앞서 서론에서 설명한 것처럼 본 연구는 우리 사회에서 주요 이슈로 부상하고 있는 문제들을 선별하기 위한 객관적인 방법을 마련하기 위한 것이다. 이러한 목적을 위해 이 논문에서는 사회문제 키워드 도출 시스템이 개발될 수 있다는 것을 일련의 프로토타입 시스템을 통해 보여주었고, 이 시스템이 얼마나 효용성이 있는지 평가하였다. 이 연구에서 개발된 프로토타입 시스템은 주요 사회문제 추출을 위한 정량적인 방법론을 제시하고 이를 반자동화함으로써 사회문제에 관한 조사 · 연구의 시간과 비용을 최소화할 수 있는 하나의 객관적인 방안을 도출했다는 의의를 지닌다. 그러나 향후에 현재 개발된 시스템을 확장하고 성능을 보완한다면, 프로토타입 버전을 벗어나 실제 업무에 활용될 수 있을 것으로 예상된다.

향후 시스템을 확장할 수 있는 여러 요소 가운데, 첫째로 시스템 자동화 및 실시간 처리 기술이 필요하다. 뉴스 데이터는 실시간으로 업데이트되기 때문에 신속한 수요 조사를 위해서는 토픽 모델링이 실시간으로 처리되는 것이 필요하며, 뉴스 데이터의 수집부터 비주얼라이제이션까지 사람의 개입 없이 자동화되는 것이 필요하다. 현재는 토픽 모델링의 경우 토픽 키워드들이 추출되고 매뉴얼 하게 토픽이 레이블링 되는데, 이 점은 꼭 보완해야 될 사항이다. 또한 뉴스 데이터와 같은 대용량 데이터를 빠르게 처리하기 위해 하둡 기반의 빅데이터 처리 플랫폼의 구축이 필요하다. 이를 바탕으로 현재 국내 뉴스 데이터뿐만 아니라 전 세계 뉴스 데이터를 광범위하게 수집하고 분석함으로써 글로벌 한 이슈를 쉽게 파악하고 선제적으로 대응할 수 있을 것이다. 또한 키워드 연관관계와 같은 추가적인 비주얼라이제이션 기능 구현, 사회문제 예측, 특정 토픽에 대한 뉴스기사의 긍정 · 부정을 판단하는 평판 분석, 토픽 모델링의 정확성 향상 등의 연구가 필요하다.

## 참고문헌

Aggarwal, C. and C. Zhai, *Mining Text Data*, Springer, 2012.

Blei, D., A. Ng, M. Jordan, and J. Lafferty, “Latent Dirichlet Allocations,” *Journal of Machine Learning Research*, Vol.3, No.4-5(2003), 993~1022.

Blei, D. and J. Lafferty, "Dynamic topic models," *International Conference on Machine Learning*, (2006), 113~120.

Blei, D., "Probabilistic Topic Models," *Communications of the ACM*, Vol.55, No.4(2012), 77~84.

Dalvi, N., R. Kumar, B. Pang, and A. Tomkins, "Matching Reviews to Objects using a Language Model," *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2009), 609~618.

Fei-Fie, L. and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," *IEEE Computer Vision and Pattern Recognition*, Vol.2(2005), 524~531.

Fulton, S. and J. Fulton, *HTML5 Canvas*, O'Reilly Media, Inc., The first edition, 2012.

JGibbLDA-*A Java Implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling for Parameter Estimation and Inference*. Available at http://jgibblda.sourceforge.net (Accessed 13 September, 2013).

Kang, S., *Korean Lexical Analysis*. Available at http://nlp.kookmin.ac.kr/HAM/kor/ham-intr.html(Accessed 13 September, 2013).

Kam, M. and M. Song, "A Study on Differences of Contents and Tones of Arguments among Newspapers using Text Mining Analysis," *Journal of Intelligence and Information Systems*, Vol.18, No.3(2012), 53~77.

Korean Integrated News Database Systems(KINDS). Available at http://www.kinds.or.kr(Accessed 13 September, 2013).

Lee, C., J. Hur, H. Oh, H. J Kim, P. Ryu, and H. K. Kim, "Technology Trends of Issue Detection and Predictive Analysis on Social Big Data," *Electronics and Telecommunications Research Institute*, Vol.28, No.1(2013), 62~71.

Liu, B., *Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)*, Morgan and Claypool Publishers, 2012.

Misra, H., F. Yvon, J. Jose, and O. Cappe, "Text Segmentation via Topic Modeling : An Analytical Study," *Proceedings of International Conference on Information and Knowledge Management(CIKM)*, (2009), 1553~1556.

Recorded Future, *Web Intelligence for Business Decisions*. Available at https://www.recorded-future.com (Accessed 13 September, 2013).

Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smith., "The author-topic model for authors and documents," *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, (2004), 487~494.

Sun, B., P. Mitra, H. Zha, C. Giles, and J. Yen, "Topic Segmentation with Shared Topic Detection and Alignment of Multiple Documents," *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, (2007), 199~206.

Wagner, C., Topic Models, *DIGITAL-Institute of Information and Communication Technologies*. Available at http://www.slideshare.net/clauwa/topic-models-5274169(Accessed 13 September, 2013).

Abstract

# A Proposal of a Keyword Extraction System for Detecting Social Issues

Dami Jeong* · Jaeseok Kim* · Gi-nam Kim**
Jong-Uk Heo*** · Byung-Won On**** · Mijung Kang*****

To discover significant social issues such as unemployment, economy crisis, social welfare etc. that are urgent issues to be solved in a modern society, in the existing approach, researchers usually collect opinions from professional experts and scholars through either online or offline surveys. However, such a method does not seem to be effective from time to time. As usual, due to the problem of expense, a large number of survey replies are seldom gathered. In some cases, it is also hard to find out professional persons dealing with specific social issues. Thus, the sample set is often small and may have some bias. Furthermore, regarding a social issue, several experts may make totally different conclusions because each expert has his subjective point of view and different background. In this case, it is considerably hard to figure out what current social issues are and which social issues are really important.

To surmount the shortcomings of the current approach, in this paper, we develop a prototype system that semi-automatically detects social issue keywords representing social issues and problems from about 1.3 million news articles issued by about 10 major domestic presses in Korea from June 2009 until July 2012. Our proposed system consists of (1) collecting and extracting texts from the collected news articles, (2) identifying only news articles related to social issues, (3) analyzing the lexical items of Korean sentences, (4) finding a set of topics regarding social keywords over time based on probabilistic topic modeling, (5) matching relevant paragraphs to a given topic, and (6) visualizing social keywords for easy understanding.

In particular, we propose a novel matching algorithm relying on generative models. The goal

* Graduate School of Convergence Science and Technology, Seoul National University
** Department of Digital Media, Ajou University
*** Department of Web Science, Korea Advanced Institute of Science and Technology
**** Corresponding Author: Byung-Won On
Advanced Institutes of Convergence Technology, Seoul National University
C307, 864-1 Iui-dong, Yeongtong-gu, Suwon-si, Gyeonggi-do 443-270, Korea
Tel: +82-31-888-9048, Fax: +82-31-888-9040, E-mail: bwon@snu.ac.kr
***** Advanced Institutes of Convergence Technology, Seoul National University

of our proposed matching algorithm is to best match paragraphs to each topic. Technically, using a topic model such as Latent Dirichlet Allocation (LDA), we can obtain a set of topics, each of which has relevant terms and their probability values. In our problem, given a set of text documents (e.g., news articles), LDA shows a set of topic clusters, and then each topic cluster is labeled by human annotators, where each topic label stands for a social keyword. For example, suppose there is a topic (e.g., Topic1 = {(unemployment, 0.4), (layoff, 0.3), (business, 0.3)}) and then a human annotator labels "Unemployment Problem" on Topic1. In this example, it is non-trivial to understand what happened to the unemployment problem in our society. In other words, taking a look at only social keywords, we have no idea of the detailed events occurring in our society. To tackle this matter, we develop the matching algorithm that computes the probability value of a paragraph given a topic, relying on (i) topic terms and (ii) their probability values. For instance, given a set of text documents, we segment each text document to paragraphs. In the meantime, using LDA, we can extract a set of topics from the text documents. Based on our matching process, each paragraph is assigned to a topic, indicating that the paragraph best matches the topic. Finally, each topic has several best matched paragraphs. Furthermore, assuming there are a topic (e.g., Unemployment Problem) and the best matched paragraph (e.g., Up to 300 workers lost their jobs in XXX company at Seoul). In this case, we can grasp the detailed information of the social keyword such as "300 workers", "unemployment", "XXX company", and "Seoul." In addition, our system visualizes social keywords over time. Therefore, through our matching process and keyword visualization, most researchers will be able to detect social issues easily and quickly.

Through this prototype system, we have detected various social issues appearing in our society and also showed effectiveness of our proposed methods according to our experimental results. Note that you can also use our proof-of-concept system in http://dslab.snu.ac.kr/demo.html.

**Key Words** : Topic Modeling, Generative Model, Matching, Text Mining, Social Issue Keywords, Social Issue Filtering, News Articles, Time Series Keyword Visualization

# 저 자 소 개

**정다미**

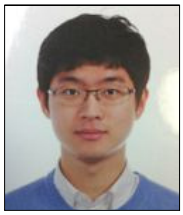
이화여자대학교 문헌정보학 학사를 마쳤고, 현재 서울대학교 융합과학기술대학원 디지털정보융합전공 석사과정에 재학 중이다. 관심분야는 텍스트마이닝, 정보 검색, 사용자경험(UX) 등이다.

**김재석**

서강대학교 경영학과와 컴퓨터공학 학사를 마쳤고, 서울대학교 융합과학기술대학원에서 디지털정보융합전공 석사과정에 재학 중이다. 관심분야는 시멘틱 웹, 정보 설계, Human-Computer Interaction이다.

**김기남**

경기대학교 컴퓨터과학과에서 학사를 마쳤고, 현재 아주대학교 미디어학과에서 석사과정에 재학 중이다. 관심분야는 데이터마이닝과 정보시각화, HTML5를 비롯한 웹 기술, HCI(Human Computer Interaction) 등이다.

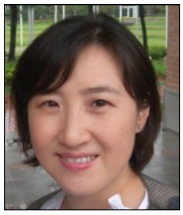


**허종욱**

아주대학교 정보컴퓨터공학과에서 학사를 마쳤고, 한국과학기술원 웹사이언스 공학과에서 석사과정에 재학 중이다. 관심분야로는 영상 처리 및 보안 이며, 그 외 멀티미디어 검색 알고리즘 및 데이터베이스에 관심을 갖고 있다.

**온병원**

미국 펜실베이니아주립대학교 컴퓨터공학과에서 박사학위를 취득하였고, 캐나다 브리티시컬럼비아대학교 컴퓨터과학과에서 포스닥 연구원으로 재직하였으며, 미국 일리노이대학교의 Advanced Digital Sciences Center에서 선임 연구원으로 근무하였다. 현재는 서울대학교 차세대융합기술연구원에서 연구교수와 공공데이터 연구센터장을 역임하면서, 빅데이터 기술 및 분석에 관한 다수의 정부 프로젝트를 수행하고 있다. 주요 연구분야로는 데이터마이닝, 데이터베이스, 기계학습, 비주얼라이제이션, 소셜 네트워크 마이닝, 텍스트 마이닝 등이다.



**강미정**

서울대학교 미학과에서 학사를 마쳤고, 같은 대학교 대학원 미학과에서 시각문화 이론과 기호학으로 석사 및 박사학위를 취득했다. 대구 시청에서 미술관준비팀 수석큐레이터로 근무했으며, 현재는 서울대학교 차세대융합기술연구원 연구교수로 재직하면서 감성학습, 신경디자인 등 인문 및 예술 중심 융합연구를 하고 있다. 저서 『퍼스의 기호학과 미술사』(2011) 외에 기타 다수 논문을 출판했다.