ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2014 June: 20(2): 109~122 http://dx.doi.org/10.13088/jiis.2014.20.2.109

# 토픽 모델링을 이용한 트위터 이슈 트래킹 시스템

**배정환**
연세대학교 문헌정보학과 대학원
(*hng88@naver.com*)

**한남기**
연세대학교 문헌정보학과 대학원
(*haruaki.pn@gmail.com*)

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

현재 우리는 소셜 네트워크 서비스(Social Network Service, 이하 SNS) 상에서 수많은 데이터를 만들어 내고 있다. 특히, 모바일 기기와 SNS의 결합은 과거와는 비교할 수 없는 대량의 데이터를 생성하면서 사회적으로도 큰 영향을 미치고 있다. 이렇게 방대한 SNS 데이터 안에서 사람들이 많이 이야기하는 이슈를 찾아낼 수 있다면 이 정보는 사회 전반에 걸쳐 새로운 가치 창출을 위한 중요한 원천으로 활용될 수 있다. 본 연구는 이러한 SNS 빅데이터 분석에 대한 요구에 부응하기 위해, 트위터 데이터를 활용하여 트위터 상에서 어떤 이슈가 있었는지 추출하고 이를 웹 상에서 시각화 하는 트위터 이슈 트래킹 시스템 TITS(Twitter Issue Tracking System)를 설계하고 구축 하였다. TITS는 1) 일별 순위에 따른 토픽 키워드 집합 제공 2) 토픽의 한달 간 일별 시계열 그래프 시각화 3) 토픽으로서의 중요도를 점수와 빈도수에 따라 Treemap으로 제공 4) 키워드 검색을 통한 키워드의 한달 간 일별 시계열 그래프 시각화의 기능을 갖는다. 본 연구는 SNS 상에서 실시간으로 발생하는 빅데이터를 Open Source인 Hadoop과 MongoDB를 활용하여 분석하였고, 이는 빅데이터의 실시간 처리가 점점 중요해지고 있는 현재 매우 주요한 방법론을 제시한다. 둘째, 문헌정보학 분야뿐만 아니라 다양한 연구 영역에서 사용하고 있는 토픽 모델링 기법을 실제 트위터 데이터에 적용하여 스토리텔링과 시계열 분석 측면에서 유용성을 확인할 수 있었다. 셋째, 연구 실험을 바탕으로 시각화와 웹 시스템 구축을 통해 실제 사용 가능한 시스템으로 구현하였다. 이를 통해 소셜미디어에서 생성되는 사회적 트렌드를 마이닝하여 데이터 분석을 통한 의미 있는 정보를 제공하는 실제적인 방법을 제시할 수 있었다는 점에서 주요한 의의를 갖는다. 본 연구는 JSON(JavaScript Object Notation) 파일 포맷의 1억 5천만개 가량의 2013년 3월 한국어 트위터 데이터를 실험 대상으로 한다.

**주제어** : 데이터 마이닝, 소셜네트워크 분석, 이슈 클러스터링, 토픽 분석

논문접수일 : 2014년 6월 15일 논문수정일 : 2014년 6월 19일 게재확정일 : 2014년 6월 21일
투고유형 : 학술대회 우수논문 교신저자 : 송민

## 1. Introduction

현재 우리는 트위터, 페이스북 등 소셜 네트워크 서비스(Social Network Service, 이하 SNS) 상에서 친구들과 일상을 공유하고, 다양한 사람들과 연결하여 서로 소통하면서 수많은 데이터를 만들어 내고 있다. EMC 보고서(2012)에 따르면 2012 년 매일 4 억개의 트윗이 개제되었고, 페이스북 이용자 수는 10 억명에 달하며, 20 조 분량의 Text message 가 생성되었다. 특히, 모바일 기기의 발전과 SNS 의 결합으로 인해 과거와는 비교할 수 없는 대량의 데이터가 생성되면서 사회적으로도 큰 영향을 미치고 있다. 사람들간의 관계, 시간, 장소, 취향 등 실제 세상의 모든 것을 담고 있는 방대한 SNS 데이터 안에서 우리가 결국 알고 싶은 것은 '사람들이 누구와 언제 어디서 무엇을 어떻게 왜 이야기하는가?'일 것이다. 이처럼 사람들이 많이 이야기하는 이슈를 찾아낼 수 있다면, 이 정보는 정치, 경제, 문화 등 사회 전반에 걸쳐 새로운 가치 창출을 위한 중요한

원천으로 활용될 수 있다. IDC(2014)가 최근 발간한 '전세계 빅데이터 기술 및 서비스 전망 보고서'에 따르면, 글로벌 빅데이터 기술 및 서비스 시장이 연평균(CAGR) 27%로 성장해 2017 년 324 억 달러 규모에 이를 것으로 전망된다.

본 연구는 이러한 SNS 빅데이터 분석에 대한 요구에 부응하기 위해, 트위터 데이터를 활용하여 트위터 상에서 어떤 이슈가 있었는지 추출하고 이를 웹 상에서 시각화 하는 트위터 이슈 트래킹 시스템 TITS(Twitter Issue Tracking System)를 설계하고 구축 하였다. TITS 는 먼저 Hadoop 과 MongoDB 를 사용하여 빅데이터를 실시간으로 처리 및 저장하고, 둘째로 토픽 모델링을 통해 키워드 빈도수를 기반으로 하는 기존의 이슈 트래킹과 차별점을 두었다. 마지막으로 분석 결과를 시각화 하여 웹 페이지를 통해 이용자 GUI 환경을 구축했다. 특히 Open Source 라이브러리를 활용하여, 부트스트랩으로 반응형 웹을 구현하였고 d3.js 로 그래프를 시각화하여 직관적인 컨텐츠 배치를 통한 이용자 편의를 고려하였다.

본 연구의 나머지 부분은 다음과 같다. 2 장에서는 SNS 데이터 분석을 위한 토픽 모델링 방법론 및 소셜미디어 분석에 대한 선행 연구들을 살펴본다. 3 장에서는 2013 년 3 월 트위터 상에서 나타난 이슈 분석을 위한 실험 데이터 및 시스템 설계에 대해 설명한다. 4 장에서는 시스템 구현 결과에 대해 보고하며, 연구에 대한 결론과 후속 연구 제안으로 5 장을 맺는다.

## 2. Related Works

이 장에서는 SNS 데이터 분석을 위한 토픽 모델링 방법론 및 소셜미디어 분석에 대한 선행 연구들을 살펴본다.

### 2.1 Topic Modeling

토픽 모델링은 Blei et al.(2003)의 LDA(Latent Dirichlet Allocation) 알고리즘을 기반으로 한 절차적 확률 분포 모델로, 텍스트 마이닝 영역에서 사용하는 연구 방법론이다. 어떤 주제들의 집합이라고 가정된 한 문헌을 구성하는 단어들을 확률적으로 계산하여, 이 결과 값을 토픽 주제어들의 집합으로 추출하는 알고리즘이다. Ryu et al.(2013)의 연구에서는 토픽 모델링 기법을 사용해 트위터 트렌드를 분석한 결과 키워드 빈도수 기반의 방법에 비하여 보다 효과가 우수함을 확인하였다. Jin et al.(2013)은 토픽 모델링 기법을 사용하여 특정 키워드 중심의 네트워크를 연결하고 시계열에 따른 토픽 변화를 추적함으로써 토픽 모델링 기법이 빠르게 변화하는 소셜미디어 상의 토픽을 추적하는데 효과적임을 제안하였다. Bae et al.(2013)은 2012 년 대선 당시의 트위터 데이터를 수집 후, 토픽 모델링 기법으로 각 후보 별 이슈를 분석하는 연구를 수행하였다. Kang et al.(2013)은 토픽 모델링 기법을 신문 데이터에 적용하여 오피니언 마이닝을 수행하였는데, 이를 통해 토픽 모델링 기법이 트위터 데이터 이외에도 일반적인 기사 데이터 분석에 사용할 수 있음을 보였다. 이상의 연구들을 통해, 토픽 모델링 기법이 SNS 빅데이터 환경에서 토픽을 분석하고 주제어를 도출하는데 유용함을 확인할 수 있었고, 본 연구에서도 3 월 한달 간 일별로 변화하는 트위터 상 이슈를 추출함에 있어 토픽 모델링 기법을 사용하여 정확한 토픽의 변화 양상을 추적하였다.

### 2.2 SocialMedia Analysis

소셜미디어(http://ko.wikipedia.org/wiki/소셜미디어)는 개방, 참여, 공유의 가치로 요약되는 웹 2.0시대의 도래에 따라 개인의 생각이나 의견, 경험, 정보 등을 서로 공유하고 타인과의 관계를 생성 또는 확장시킬 수 있는 개방화된 온라인 플랫폼을 의미한다. 소셜미디어는 양방향성을 활용하여 이용자들이 자발적으로 참여하고 정보를 공유하며 컨텐츠를 만들어 나가는 특성이 있고, 일반적으로 사람과 사람, 또는 사람과 정보를 연결하고 상호 작용할 수 있는 서비스를 제공한다. 특히, 최근 모바일 기기와의 결합으로 인해 대량의 소셜미디어 데이터가 발생하고 있고, 이를 분석하기 위한 다양한 연구가 이루어지고 있다.

우선 소셜미디어 데이터의 유용성에 대한 연구들이 있다. Sohn et al.(2012)의 연구에서는 트위터를 분석해서 사회 현상 및 동향을 분석하는데 트위터가 유용한 매체임을 확인하였다. Kim et al.(2012)은 마케팅 전략적인 관점에서 트위터의 잠재성을 이해하고자 하는 목적으로 트위터 상의 정보 전달체계를 중심으로 신제품 프로모션에 대한 소셜 네트워크의 구전 효과를 분석 하였다. Nam et al.(2011)은 소셜미디어에서의 정보 확산에 대한 기존의 학문적 지식을 증대 시키고 실무적으로 기업이 고객과 소통하는 데 있어 소셜미디어를 어떻게 전략적으로 활용할 수 있는가에 대한 실질적인 방향을 제시했다. 본 연구에서도 소셜미디어 중 사회 변화에 가장 빠르게 반응하는 트위터 데이터를 활용하여 사회적인 이슈를 추출하는데 사용하였다.

또한, 소셜미디어 데이터를 효과적으로 처리하기 위한 기법적인 측면의 연구들도 활발한데 Byeon et al.(2011)은 트위터에서 이슈 키워드를 도출하는 연구를 수행하면서 형태소 분석을 통해 기본적으로 명사만 남긴 후 그 빈도수를 기준으로 추출하였다. 위키피디아를 활용한 연구들 중에서, Bae and Ko(2009)는 한국 위키피디아 데이터에서 개체를 추출하여 한국어 개체명 사전을 구축하는 연구를 수행하였다. Han(2009)의 연구는 위키피디아에서 추출한 단어가 시소러스를 작성하는데 사용해도 될 정도로 학술적임을 증명하였다. Kim and Chung(2012)의 연구에서는 위키피디아 데이터를 사용하여 검색어에 대한 대체어(annotation) 기능을 넣어 검색의 성능 향상을 이루었다. 본 연구에서도 위키피디아 데이터에서 개체명 항목을 추출하여 이를 이슈 키워드에 가중치로 부여하였다.

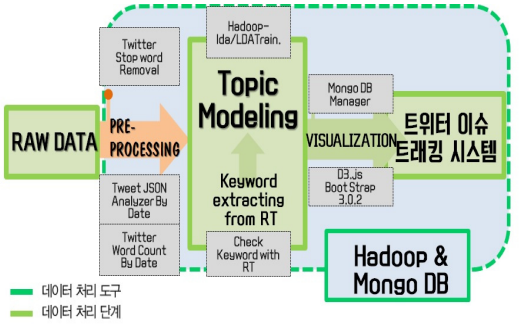
Kim(2013)은 트위터 상에서 리트윗 행위를 하는 이유에 대해 연구하였는데, 그 결과 리트윗의 이유로 두드러지는 것은 공감과 공유를 통한 자기 만족의 가치였고, 이를 위해서 화제성과 정보성이 포함된 트윗 메시지들을 리트윗하는 경향이 많다는 것을 밝혔다. 본 연구에서도 이슈 키워드를 선정하는데 있어 리트윗 횟수를 반영하여 가중치를 두었다.

## 3. Methodology

이 장에서는 본 연구에서 개발한 트위터 이슈 트래킹 시스템에 대해 소개하고, 이를 위한 실험 데이터와 시스템 설계에 대해 설명한다.

### 3.1 Twitter Issue Tracking System(TITS)

본 연구에서는 트위터에서 생성되는 방대한 양의 텍스트 데이터를 효과적으로 실시간 처리




〈Figure 1〉 Overview of Twitter Issue Tracking System

하고, 토픽 모델링을 통한 이슈 추출과 이의 시각화 기능을 특징으로 하는 TITS 를 개발하였다. 본 연구를 위한 실험 데이터는 JSON(JavaScript Object Notation) 파일 포맷의 1 억 5 천만개 가량의 2013 년 3 월 한국어 트위터 데이터를 대상으로 한다. <Figure 1>은 전체 시스템을 도식화해서 보여주는데, 30GB 분량의 원 데이터(Raw Data)를 Tweet JSON Analyzer By Date 모듈을 사용하여 일별로 정리하고 이를 해쉬태그, 멘션을 주고 받은 이용자, rt, url, 본문 5 종류로 parsing 하여 55GB 데이터를 MongoDB(http://www.mongodb.org/)에 저장한다. 다음 단계로 Twitter Stop word Removal 모듈로 트위터 본문에 대한 전처리(Pre-Processing) 과정을 수행하고, Twitter Word Count By Date 모듈로 일별 출현 단어의 빈도수를 계산한다. 이와 같은 전처리 과정을 거친 데이터에 LDA 토픽 모델링 알고리즘을 적용하여 이슈 토픽을 추출하고 RT 와 위키피디아에서 추출한 키워드와 비교하여 일치하는 키워드에 가중치를 부여하여 순위를 매긴다. 최종적으로 분석한 결과를 웹 시스템으로 구축, 시각화해서 보여준다. 이 처리는 모두 Hadoop 환경하에 Java 프로그래밍으로 수행되었다.

### 3.2 Big Data Processing Tools

SNS 데이터는 빅데이터의 조건으로 정의되는 데이터의 양(volume), 데이터 입출력 속도(velocity), 데이터 종류의 다양성(variety)을 만족하는데, 정제 되지 않은 다양한 형태의 비정형 데이터를 처리하기 위한 자연어 처리, 불용어 제거, 명사어 추출 등의 텍스트 마이닝 기법들과 많은 양의 데이터를 빠르게 실시간으로 처리하기 위한 Hadoop 과 같은 분산처리 시스템, 기존의 관계형 데이터베이스를 보완하는 NoSQL 과 같은 최신 빅데이터 기술이 필요하다.

Hadoop 은 Map-Reduce 방식을 통해 빅데이터의 분산 처리를 가능하게 하는 Apache 재단의 Open Source Framework 이다. Hadoop 은 기존의 Single-Node Computing 으로는 처리할 수 없는 방대한 양의 데이터를 처리할 수 있도록 설계되었기 때문에, 빅데이터 연구와 활용에 가장 많이 사용되고 있다. 본 연구에서는 트위터 빅데이터의 전처리 과정부터 토픽 모델링 알고리즘을 통한 이슈 추출에까지 전 과정에서 사용되었다.

MongoDB 는 기존 관계형 DB 와는 달리 테이블과 스키마가 없이 데이터 처리 성능과 데이터 접근성을 가장 중요한 목표로 하는 확장성 있는 새로운 NoSQL 방식의 Open Source 데이터베이스이다. MongoDB 는 데이터의 삽입, 삭제, 쿼리 등을 고속으로 처리할 수 있어, 기존의 RDB 보다 빅데이터 처리에 적합하여 트위터 빅데이터의 처리 결과 저장 및 최종 시스템 구축에 이용하였다. 또한, MongoDB 는 JSON 문서를 지원하는 문서 기반 데이터베이스로 실험 문헌 파일의 형태가 JSON 포맷이었기 때문에 MongoDB 를 보다 효과적으로 활용할 수 있었다.

빅데이터가 이슈가 되면서 방대한 데이터를 분석하는 것뿐만 아니라 분석한 데이터를 이용자에게 알기 쉽게 보여주기 위한 시각화의 중요성도 점점 커지고 있다. 본 시스템에서는 시각화를 위한 도구로 d3.js(http://d3js.org/) 라이브러리를 사용하였다. d3.js 는 임의의 데이터를 문서 객체 모델(DOM)과 결합시켜 데이터 기반 문서(Data-Driven Documents)를 작성하기 위한 목적으로 만들어져서, 데이터간의 상호작용이 용이하고 데이터의 흐름을 부드러운 애니메이션으로 처리 하여 실시간 데이터를 다루는데 유용하다. 이 라이브러리를 사용하여 복잡한 데이터를 빠르게 이해하기 위한 데이터 시각화 부분을 CSS3, HTML5, SVG 라는 웹 표준을 통해 쉽게 할 수 있었다.

마지막으로 TITS 웹 시스템 구축을 위해 사용한 부트스트랩(http://getbootstrap.com/2.3.2/)은 미리 만들어진 스타일시트와 자바스크립트 플러그인으로 구성된 라이브러리이다. HTML 로 웹 페이지의 뼈대를 만들고 스타일시트에서 레이아웃을 만들기 위한 CSS 속성과 값을 입력하는 대신 부트스트랩에 미리 정의된 클래스 선택자를 HTML 코드에 삽입만 하면 레이아웃과 각종 요소를 만들 수 있다. 부트스트랩 스타일시트는 웹 페이지를 만드는 데 필요한 거의 모든 요소를 정의해놓았기 때문에 손쉽게 웹 페이지를 만들 수 있는 프레임워크(Framework )이다.

### 3.3 Big Data processing

이 장에서는 데이터 통계적 기법을 적용하여 수행한 빅데이터 처리 방법에 대해서 설명한다.

#### 3.3.1 Raw Data Pre-Processing

원 데이터에서 가치 있는 정보를 추출하기 위해서는 사전 데이터 정제 작업이 필요하기 때문에 트위터 데이터 정제를 위한 다음의 전처리 과정을 수행하였다.

1) 트윗들의 날짜 별 정리
2) 트윗 문장의 Tokenize 를 통한 단어 정보 추출
3) 단어의 빈도수, 품사, 길이를 통한 불용어 검출

먼저 실험 데이터가 GMT 0 시를 기준으로 수집되었기 때문에 날짜 별 이슈 추출을 위해서 한국 시간인 GMT 9 시를 적용하여 일별로 정리했고, 트윗 본문에서 space 를 기준으로 single token 으로 분리하여, 각 단어들의 단순 출현 빈도와 한 트윗에서의 동시 출현 빈도를 계산하였다. 다음으로 날짜 별로 출현 빈도가 1 인 단어는 불용어로 취급하여 삭제하였고, Open Source 인 Lucene Arirang 형태소 분석기(http://sourceforge.net/projects/lucenekorean/)를 사용하여 명사/명사추정(고유명사)으로 판정되지 않는 단어들을 전부 삭제하였다. 그 이유는 이슈가 될 수 있는 주제어나 키워드는 기본적으로 그 품사가 명사라고 가정하였기 때문이다.

#### 3.3.2 Extract Issue Keyword from RT

기본적인 전처리 과정을 거친 후, RT 된 횟수가 많으면 많을수록 해당 날짜의 이슈와 연관될 가능성이 높기 때문에 날짜 별로 RT 트윗에서 사용한 키워드를 따로 추출하여 출현 빈도 순으로 정렬 하였다. 이후 중요도를 처리하기 위하여 정리된 단어들에 RT 횟수를 기반으로 한 정규화된 가중치를 부여하였다. 또한 인명, 지명, 회사명 등의 고유명사에 가중치를 주기 위해 한국 위키피디아(http://dumps.wikimedia.org/kowiki/)의

항목명 데이터 44 만개 가량을 추출하여 일치하는 단어에는 가중치를 적용했다. 이를 사용한 이유는 선행연구에서 살펴보았듯, 위키피디아 항목명과 일치한 단어는 분명 고유명사 혹은 의미를 지닌 명사일 가능성이 높기 때문이다. 이 조건을 만족하는 단어들이 다른 단어들에 비해 더 많은 출현치를 획득할 수 있도록 가중치를 높게 설정하였고, 이를 통해 RT 출현 빈도가 높고, 위키피디아에 등재된 단어가 이슈 키워드 중 높은 순위에 올라올 수 있도록 보정하였다. 다음 <Table 1>은 RT 에서 추출한 상위 10 개 단어와 그 빈도수의 일부분이다.

〈Table 1〉 Sample RT Top 10 List

| 3/8 | | 3/10 | | 3/13 | |
|---|---|---|---|---|---|
| word | count | word | count | word | count |
| 트위터 | 31135 | 캬캬 | 37015 | 셀카 | 79200 |
| 인천공항 | 24325 | 트위터 | 25070 | 트위터 | 36215 |
| 이벤트 | 23863 | 려욱 | 22505 | 이벤트 | 27215 |
| 스타벅스 | 22480 | 예성 | 21630 | RT | 23266 |
| RT | 21499 | 태민 | 20765 | 스타벅스 | 16960 |
| 오픈 | 13585 | 스타벅스 | 19255 | 매일 | 15930 |
| 알바몬 | 11840 | RT | 17902 | 드립 | 12514 |
| 시티 | 11775 | 이벤트 | 17418 | 안철수 | 12405 |
| 드립 | 11007 | 뮤직뱅크 | 15530 | 몬스타 | 11040 |
| 자카르타 | 10510 | 오픈 | 10725 | 오픈 | 10725 |

### 3.3.3 Time-Series Analysis by Topic Modeling

토픽 모델링 알고리즘을 사용하면 데이터 주제를 단어 집합으로 추출하여, 이슈를 단순 키워드가 아닌 스토리텔링으로 표현 가능하다. 또한 추출된 토픽 간의 유사도를 계산하여, 토픽의 변화를 시계열로 나타낼 수 있다. 키워드의 빈도수를 기반으로 하는 기존 이슈 트래킹에서는, 토픽의 스토리텔링 및 시간에 따른 토픽의 변화 양상에 대한 분석이 어렵다는 한계가 있다. 따라서 본 연구에서는 Hadoop을 기반으로 토픽 모델링을 가능하게 한 Open Source, hadoop-lda Library에 TF*IDF(Term Frequency*Inverse Document Frequency) 가중치를 추가하여 적용하였다. 정제를 위한 전처리 과정을 거친 각 단어의 주제 확률 분포를 계산해 일별 50개의 토픽으로 단어들을 분류하고 이중 실제 이슈와 일치하는 토픽 10개를 선정하였다.

1) 일별 50개 토픽으로 단어들을 분류하여 결과가 실제 이슈와 일치하는 토픽 선정
2) 이때, RT 트윗과 위키피디아에서 추출한 결과와 토픽 모델링 결과가 일치하는 단어에 가중치 부여
3) 추출한 각 토픽 그룹 결과를 정치, 사회, 연예, 스포츠, 일상 5개 분야로 범주화

아래의 <Table 2>는 토픽 모델링 결과의 일부분으로 특정일의 토픽 그룹 및 각 토픽에 포함된 중요 키워드들과 그 점수를 보여준다.

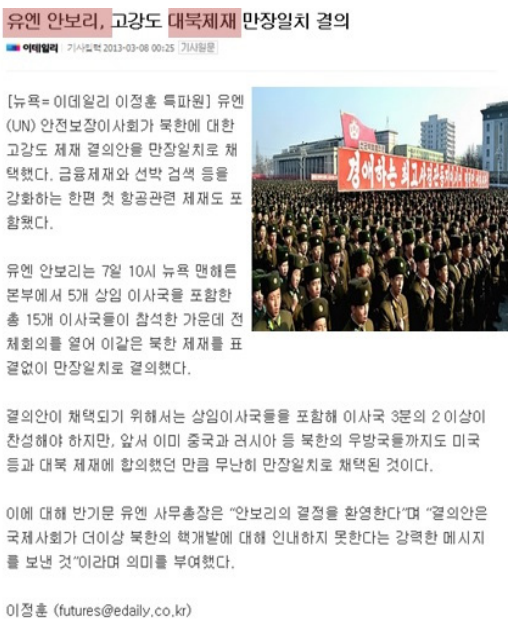
〈Table 2〉 Sample Topic Modeling Result

| Rank / Topic Group | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 4 (0.0254) | 대통령 (0.0459) | 박근혜 (0.0332) | 부정선거 (0.0094) | MB (0.0083) | 박원순 (0.0057) |
| 18 (0.0227) | 안철수 (0.0391) | 민주당 (0.0314) | 노원병 (0.0083) | 블랙야크 (0.0067) | 최초 (0.0046) |
| 32 (0.0223) | 북한 (0.0472) | 전쟁 (0.0162) | 미국 (0.0146) | 김정은 (0.0120) | 안보리 (0.0069) |
| 15 (0.0222) | 출국 (0.0714) | 자카르타 (0.0237) | TEEN (0.0114) | 태민 (0.0102) | 성종 (0.0101) |
| 36 (0.0211) | 김병관 (0.0358) | 장관 (0.0176) | ZZZzzz (0.0114) | 피고네 (0.0114) | 국방장관 (0.0079) |
| 6 (0.0209) | 눈물 (0.0137) | 추억 (0.0113) | 음악 (0.0077) | 슬픔 (0.0071) | 시절 (0.0063) |
| 22 (0.0206) | 대한문 (0.0291) | 분향소 (0.0206) | 시민 (0.0122) | 노동자 (0.0103) | 강제철거 (0.0060) |

이와 같이 3월 한달 간의 트위터 데이터를 대상으로 일별 토픽 모델링을 통해 추출한 이슈 키워드들을 그룹화하여 각 그룹의 주제에 가장 적합한 범주화 과정을 수행하였고, 이 결과가 유의미한지 확인해 보기 위하여 실제 뉴스 기사와 비교해 보았다. 예를 들어, <Table 3> 에서 3월 8일 북한, 전쟁 등의 이슈 키워드가 추출되어 이를 '사회'라는 주제로 범주화 시켰고, 해당일의 실제 뉴스 기사를 검색해본 결과 동일한 내용이라고 판단되는 기사가 있는지 확인하였다. 이를 통해 트위터 데이터에 대한 토픽 모델링 결과가 실제 현실의 이슈를 반영한다는 것을 확인할 수 있었고, 또한 기존의 빈도수 기반 키워드 순위 나열 방식에서 벗어나 토픽 별로 유사한 단어를 그룹화하여 보여줌으로써 이용자에게 사건을 하나의 이야기로 전달하는 스토리텔링의 가능성을 보여주었다. <Table 3>에 제시한 일별 토픽과 키워드의 스토리텔링을 실제 기사에서 확인한 예는 <Figure 2>, <Figure 3>, <Figure 4>에서 확인할 수 있다.

〈Table 3〉 Topic Labeling & Issue Keyword

| Date | Label | Issue Keyword |
|---|---|---|
| 3/8 | 사회 (Social Affairs) | 북한, 전쟁, 미국, 김정은, 안보리, 도발, 대북제재, WORLD, 선제타격, UN |
| 3/10 | 연예 (Entertainment) | 최고, 이순신, 아이유, OST, 투표, 안녕, 첫사랑, 지시자, 어머님, 조정치 |
| 3/13 | 정치 (Politics) | 안철수, 민주당, 노원병, 의원, 문재인, 새누리당, 국회의원, 출마, 선거, 충격 |



유엔 안보리, 고강도 대북제재 만장일치 결의

이데일리 기사입력 2013-03-08 00:25 기사원문

[뉴욕=이데일리 이정훈 특파원] 유엔(UN) 안전보장이사회가 북한에 대한 고강도 제재 결의안을 만장일치로 채택했다. 금융제재와 선박 검색 등을 강화하는 한편 첫 항공관련 제재도 포함했다.

유엔 안보리는 7일 10시 뉴욕 맨해튼 본부에서 5개 상임 이사국을 포함한 총 15개 이사국들이 참석한 가운데 전체회의를 열어 이같은 북한 제재를 표결없이 만장일치로 결의했다.

결의안이 채택되기 위해서는 상임이사국들을 포함해 이사국 3분의 2 이상이 찬성해야 하지만, 앞서 이미 중국과 러시아 등 북한의 우방국들까지도 미국 등과 대북 제재에 합의했던 만큼 무난히 만장일치로 채택된 것이다.

이에 대해 반기문 유엔 사무총장은 "안보리의 결정을 환영한다"며 "결의안은 국제사회가 더이상 북한의 핵개발에 대해 인내하지 못한다는 강력한 메시지를 보낸 것"이라며 의미를 부여했다.

이정훈 (futures@edaily.co.kr)

〈Figure 2〉 Label 'Social Affairs'

첫방 '최고다 이순신', 막내딸 아이유의 희망찾기

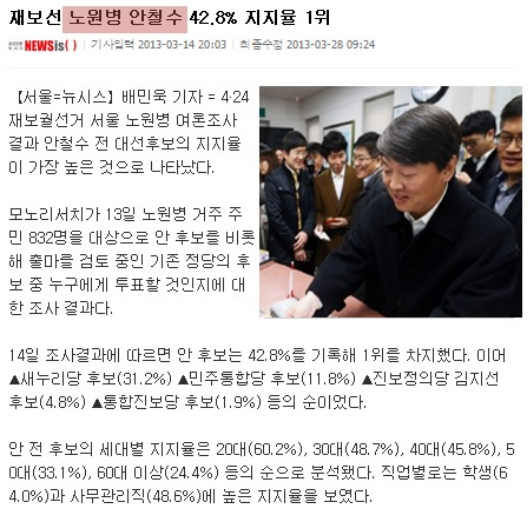
스타뉴스 기사입력 2013-03-09 21:06 최종수정 2013-03-09 21:15 기사원문

[스타뉴스 김성희 기자]

<사진캡쳐=KBS 2TV 최고다 이순신>

'국민여동생' 아이유가 특별한 능력은 없어도 밝고 당찬 막내딸로 변신해 안방극장을 찾았다.

〈Figure 3〉 Label 'Entertainment'



재보선 노원병 안철수 42.8% 지지율 1위

NEWSis( ) | 기사입력 2013-03-14 20:03 | 최종수정 2013-03-28 09:24

【서울=뉴시스】 배민욱 기자 = 4·24 재보궐선거 서울 노원병 여론조사 결과 안철수 전 대선후보의 지지율이 가장 높은 것으로 나타났다.

모노리서치가 13일 노원병 거주 주민 832명을 대상으로 안 후보를 비롯해 출마를 검토 중인 기존 정당의 후보 중 누구에게 투표할 것인지에 대한 조사 결과다.

14일 조사결과에 따르면 안 후보는 42.8%를 기록해 1위를 차지했다. 이어 ▲새누리당 후보(31.2%) ▲민주통합당 후보(11.8%) ▲진보정의당 김지선 후보(4.8%) ▲통합진보당 후보(1.9%) 등의 순이었다.

안 전 후보의 세대별 지지율은 20대(60.2%), 30대(48.7%), 40대(45.8%), 50대(33.1%), 60대 이상(24.4%) 등의 순으로 분석됐다. 직업별로는 학생(64.0%)과 사무관리직(48.6%)에 높은 지지율을 보였다.

<Figure 4> Label 'Politics'

#### 3.3.4 Visualization

본 TITS 시스템에서는 데이터 분석 결과를 웹 페이지를 통해 시각화하고 이용자 GUI 를 구현하기 위해서 Open Source 인 d3.js, Bootstrap 라이브러리를 이용해 반응형 웹 페이지를 구축하여 분석 결과를 나타내었다. 각 페이지에서는 토픽의 시계열 분석, 특정 날짜의 주요 토픽, 주요 토픽을 구성하는 이슈 키워드 등을 한눈에 제공하고 적절한 컨텐츠 배치를 통해 이용자에게 트위터 상의 이슈를 쉽고 직관적으로 파악할 수 있게 하였다. 각 기능별 상세 내용은 다음 장에서 살펴본다.

## 4. System Implementation

본 연구에서 개발한 TITS 시스템이 제공하는 기능은 다음과 같다.

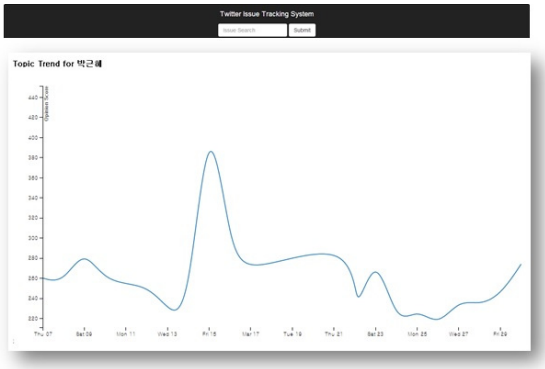
1) 토픽 모델링 결과를 일별 토픽 10개, 각 토픽당 이슈 키워드 10개로 제공.
2) 토픽 간의 유사도를 계산하여 일별 변화를 시계열 그래프로 시각화.
3) 키워드마다 계산된 토픽으로서의 중요도를 점수와 빈도수에 따른 Treemap으로 구현.
4) 마지막으로 검색 기능을 구현하여, 각 키워드 별 토픽 점수에 따른 일별 트랜드를 시계열 그래프로 시각화.

본 TITS 시스템이 실제로 구현된 모습은 인터넷의 http://informatics.yonsei.ac.kr/bigdata/d3.html 페이지에 등록되어 있으며, 전체 페이지에 대한 개괄적인 설명은 <Figure 5>와 같다.


〈Figure 5〉 Overview of Web Page

### 4.1. Keyword Search Graph

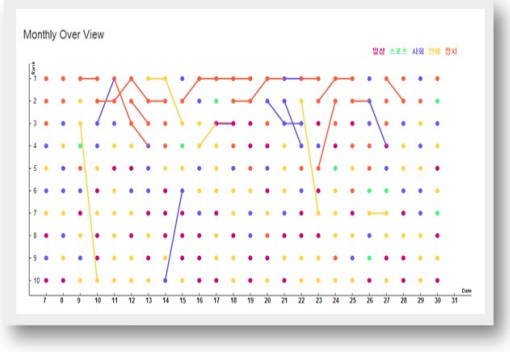
<Figure 6>은 실험을 통해 50개 토픽, 토픽 당 20개로 총 1000개 키워드를 저장해두고, 이용자가 질의어를 입력하면 Tomcat Servlet을 이용하여 MongoDB에 질의를 보내 검색하고, 그 결과 해당 키워드의 한달 간 일별 변화 양상을 Line 그래프로 구현하는 기능을 보여준다.




〈Figure 6〉 Keyword Search Graph

### 4.2 Daily Time-Series Graph by Topic

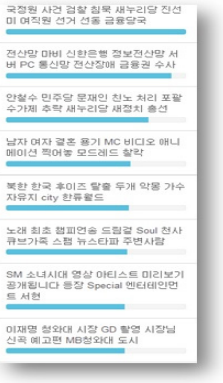
<Figure 7>은 토픽 모델링 수행 결과로 추출한 50개 토픽 중, 일별 상위 점수의 토픽 10개를 선정하고, 각 토픽 간의 유사도를 계산하여 한달간 주요 토픽의 변화 양상을 Line 그래프로 시각화하였다. 또한 각 토픽을 정치, 사회, 연예, 스포츠, 일상의 5개 카테고리로 범주화하고 색을 다르게 부여하여 변화 양상을 쉽게 확인할 수 있게 구현하였다. 특히 <Figure 7>에서는 특정 토픽이 다음 날짜의 다른 토픽과 선으로 연결되어 있는 것을 볼 수 있는데, 이런 Line 그래프를 통해 토픽의 시계열성을 파악할 수 있다. TITS 시스템의 토픽 간 연결은 토픽 모델링 알고리즘에서 제공하는 토픽 간의 유사도 점수를 통해 구현하였다.




〈Figure 7〉 Daily Time-Series Graph by Topic

### 4.3. Daily Issue Keyword

<Figure 8>은 웹 페이지 상단에 위치한 Monthly OverView의 각 날짜를 클릭할 시, 일별 상위 10개 토픽 그룹과 각 토픽당 10개 키워드, 그리고 토픽 점수를 Bar 그래프와 함께 제공하여 이용자에게 매일의 주요 키워드들과 그 순위를 한눈에 확인할 수 있도록 구현하였다.

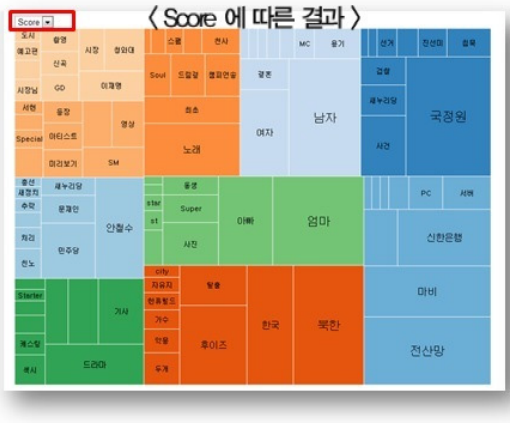


〈Figure 8〉 Daily Issue Keyword

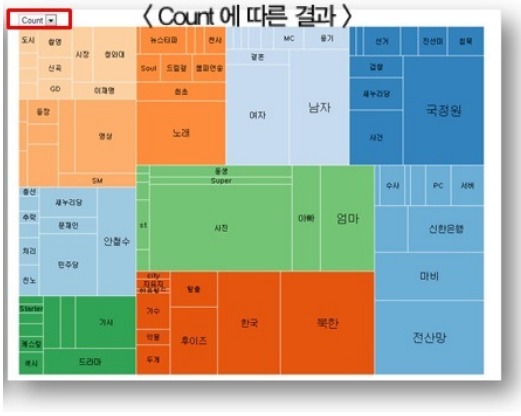
### 4.4. Treemap

d3.js의 treemap 라이브러리를 활용하여 매일의 이슈 키워드를 토픽 점수와 키워드 빈도수에 따라 두 종류로 시각화 하였다. 토픽 점수는 <Table 2>와 같이 토픽 모델링 알고리즘을 통한 단어 별 토픽점수에 3.3.2. 절에서 제시한 가중치를 적용하여 계산하였다. 이 treemap에서는 10개 토픽 그룹 별로 색깔을 다르게 하여 구분했고, 각 그룹에 속하는 10개 키워드 역시 토픽 점수와 키워드 빈도수에 비례하여 map의 크기가 달라지

도록 구현하였다. 이는 모두 이용자에게 키워드의 중요성을 한 눈에 파악할 수 있도록 정보 전달력을 높이기 위한 방법이다. 해당 treemap의 예시는 <Figure 9>, <Figure 10>과 같다.




〈Figure 9〉 Treemap_Score




〈Figure 10〉 Treemap_Count

## 5. CONCLUSION

본 연구에서는 Hadoop과 MongoDB를 기반으로 토픽 모델링을 수행하여 빅데이터를 빠른 속도로 분석할 수 있었다. 기존의 연구 환경에서는 빅데이터를 빠르게 처리하기 어려운데, 특히 Twitter 등 SNS 상에서 데이터가 항상 실시간으로 발생하기 때문에 기존 방법론으로는 처리하기 곤란한 수준의 데이터가 축적되고 있다. 그러나 Hadoop과 MongoDB는 막대한 하드웨어 비용 없이도 충분히 빅데이터를 처리할 수 있는 Framework를 제공하고 있고, 이는 데이터의 실시간 처리가 중요해지고 있는 현재 매우 의미가 크다 할 수 있다. 둘째, 문헌정보학 뿐 아니라 다양한 분야에서 사용하고 있는 토픽 모델링 기법을 실제 트위터 데이터에 적용하여 이의 유용성을 확인할 수 있었다. 이슈를 단순 키워드로 표현하는 기존 방식에 비하여, 토픽 모델링 기법은 이슈를 단어 집합으로 추출하여 사건에 대해서 하나의 연결된 이야기로 이해할 수 있게 한다. 또한, 추출된 토픽 간의 유사도를 계산하여 시간에 따른 토픽의 변화를 시계열로 추적할 수 있었다. 셋째, 연구 실험을 바탕으로 이를 실제 사용 가능한 시스템으로 구현하였다. 시각화와 웹 페이지 구축을 위하여 Open Source인 Tomcat Servlet, d3.js, 부트스트랩 라이브러리를 사용하였고, 이를 통해 이용자 GUI 환경을 구축하여 소셜미디어 데이터 분석을 통한 의미 있는 정보를 제공하는 실제적인 방법을 제시할 수 있었다. 추후 연구 과제로, 명사 추출의 정확성을 높이기 위한 NER(Named Entity Recognition) 기법 적용과 기계 학습을 통한 자동 범주화 기능에 관한 연구를 수행하고자 한다.

## Acknowledgments

본 연구는 미래창조과학부(한국정보화진흥원) 주관 2013 빅데이터 분석 경진대회 대상 수상작임.

본 연구는 2012년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2012-2012S1A3A2033291).

## 참고문헌 (References)

Bae, J. H., J. E. Son, and M. Song, "Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques," *Journal of Intelligence and Information Systems*, Vol.19, No.3(2013), 141~156.

Bae, S. J. and Y. J. Ko, "Automatic Construction of Korean Named Entity Dictionaries from Wikipedia," *Proceedings of Korea Computer Congress*, (2009), 78~79.

Blei, D., A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol.3(2003), 993~1022.

Byeon, J. H., J. M. Oh, and N. M. Moon, "A Study on Keyword Discovery Based on Social Network Service," *HCI*, (2011), 471~474.

Han, S. H., "Thesaurus Updating Using Collective Intelligence: Based on Wikipedia Encyclopedia," *Journal of the Korean Society for Information Management*, Vol.26, No.3(2009), 25~43.

Jin, S. A., G. E. Heo, Y. K. Jeong, and M. Song, "Topic-Network based Topic Shift Detection on Twitter," *Journal of the Korean Society for Information Management*, Vol.30, No.1(2013), 285~302.

Kang, B. I., M. Song, and W. S. Jho, "A Study on Opinion Mining of Newspaper Texts based on Topic Modeling," *Journal of the Korean Library and Information Science Society*, Vol.47, No.4(2013), 315~334.

Kim, H. D., "Message Attributes, Consequences, and Values in Retweet Behavior : Based on Laddering Method," *The Journal of the Korea Contents Association*, Vol.13, No.3(2013), 131~140.

Kim, H. j., I. S. Son, and D. W. Lee, "The Viral Effect of Online Social Network on New Products Promotion : Investigating Information Diffusion on Twitter," *Journal of Intelligence and Information Systems*, Vol.18, No.2(2012), 107~130.

Kim, Y. H. and Y. M. Chung, "An Experimental Study on Feature Selection Using Wikipedia for Text Categorization," *Journal of the Korean Society for Information Management*, Vol.29, No.2(2012), 155~171.

Nam, Y. W., I. S. Son, and D. W. Lee, "The Impact of Message Characteristics on Online Viral Diffusion in Online Social Media Services : The Case of Twitter," *Journal of Intelligence and Information Systems*, Vol.17, No.4(2011), 75~94.

Ryu, W. J., J. W. Ha, Md. Hijbul Alam, and S. K. Sang, "Extracting Trends from Twitter using a Topic Modeling Technique," *Proceedings of Korea Computer Congress,* (2013), 191~193.

Sohn, J. S., S. W. Cho, K. L. Kwon, and I. J. Chung, "Improved Social Network Analysis Method in SNS," *Journal of Intelligence and Information Systems*, Vol.18, No.4(2012), 117~127.

ALL IDC research, *Consumers and the Digital Universe,* EMC, 2014. Available at http://www.emc.com/infographics/digital-universe-consumer-infographic.htm.

IDC, *IDC, Big Data technologies and services worldwide market forecast $ 32.4 billion in 2017*, IDC, 2014. Available at http://www.idckorea.com/product/Getdoc.asp?idx=585&field=PressRelease.

Abstract

# Twitter Issue Tracking System by Topic Modeling Techniques

Jung-hwan Bae*, Nam-gi Han*, Min Song**

People are nowadays creating a tremendous amount of data on Social Network Service (SNS). In particular, the incorporation of SNS into mobile devices has resulted in massive amounts of data generation, thereby greatly influencing society. This is an unmatched phenomenon in history, and now we live in the Age of Big Data. SNS Data is defined as a condition of Big Data where the amount of data (volume), data input and output speeds (velocity), and the variety of data types (variety) are satisfied. If someone intends to discover the trend of an issue in SNS Big Data, this information can be used as a new important source for the creation of new values because this information covers the whole of society. In this study, a Twitter Issue Tracking System (TITS) is designed and established to meet the needs of analyzing SNS Big Data. TITS extracts issues from Twitter texts and visualizes them on the web. The proposed system provides the following four functions:

(1) Provide the topic keyword set that corresponds to daily ranking;
(2) Visualize the daily time series graph of a topic for the duration of a month;
(3) Provide the importance of a topic through a treemap based on the score system and frequency;
(4) Visualize the daily time-series graph of keywords by searching the keyword;

The present study analyzes the Big Data generated by SNS in real time. SNS Big Data analysis requires various natural language processing techniques, including the removal of stop words, and noun extraction for processing various unrefined forms of unstructured data. In addition, such analysis requires the latest big data technology to process rapidly a large amount of real-time data, such as the Hadoop distributed system or NoSQL, which is an alternative to relational database. We built TITS based on Hadoop to optimize the processing of big data because Hadoop is designed to scale

* Dept. of Library and Information Science, Yonsei University
** Corresponding Author: Min Song
Dept. of Library and Information Science, Yonsei University
120-749, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea
Tel: +82-2-2123-2405, Fax: +82-2-393-8348, E-mail: min.song@yonsei.ac.kr

up from single node computing to thousands of machines. Furthermore, we use MongoDB, which is classified as a NoSQL database. In addition, MongoDB is an open source platform, document-oriented database that provides high performance, high availability, and automatic scaling. Unlike existing relational database, there are no schema or tables with MongoDB, and its most important goal is that of data accessibility and data processing performance. In the Age of Big Data, the visualization of Big Data is more attractive to the Big Data community because it helps analysts to examine such data easily and clearly. Therefore, TITS uses the d3.js library as a visualization tool. This library is designed for the purpose of creating Data Driven Documents that bind document object model (DOM) and any data; the interaction between data is easy and useful for managing real-time data stream with smooth animation. In addition, TITS uses a bootstrap made of pre-configured plug-in style sheets and JavaScript libraries to build a web system. The TITS Graphical User Interface (GUI) is designed using these libraries, and it is capable of detecting issues on Twitter in an easy and intuitive manner. The proposed work demonstrates the superiority of our issue detection techniques by matching detected issues with corresponding online news articles.

The contributions of the present study are threefold. First, we suggest an alternative approach to real-time big data analysis, which has become an extremely important issue. Second, we apply a topic modeling technique that is used in various research areas, including Library and Information Science (LIS). Based on this, we can confirm the utility of storytelling and time series analysis. Third, we develop a web-based system, and make the system available for the real-time discovery of topics. The present study conducted experiments with nearly 150 million tweets in Korea during March 2013.

**Key Words** : Social Media Mining; Text Mining; Twitter Issue; Topic Modeling; Social Network Service; Big Data

Received: June 15, 2014 Revised: June 19, 2014 Accepted: June 21, 2014

# 저 자 소 개

**배 정 환**
Dept. of Library and Information Science, Yonsei University
Tel: +82-2-2123-2405, Fax: +82-2-393-8348, E-mail: haruaki.pn@gmail.com
연세대학교 문헌정보학 학사를 졸업 후 동 대학원 석사과정에 재학 중이다. 주요 관심분야는 텍스트 마이닝에 기반한 바이오와 소셜미디어 빅데이터 분석, 디지털 도서관 그리고 HCI 이다.

**한 남 기**
Dept. of Library and Information Science, Yonsei University
Tel: +82-2-2123-2405, Fax: +82-2-393-8348, E-mail: hngin@yonsei.ac.kr
연세대학교 문헌정보학 학사를 졸업 후 동 대학원 석사과정에 재학 중이다. 주요 관심분야는 텍스트 마이닝에 기반한 빅데이터 분석, 정보 공학 및 정보 검색 분야이다.

**송 민**
Associate Professor, Dept. of Library and Information Science, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea
Tel: +82-2-2123-2405, Fax: +82-2-393-8348, E-mail: min.song@yonsei.ac.kr
http://informatics.yonsei.ac.kr/tsmm/home.html
Prof. Song has a background in Text Mining, Bioinfomatics, Information Retrieval and Information Visualization. Prior to Yonsei, he was an Associate Professor with tenure in the Department of Information Systems at New Jersey Institute of Technology (NJIT). At NJIT, he received several grants from NSF and IMLS and published a number of papers in the Text Mining research area. Before joining NJIT, Professor Song worked at Thomson Scientific (now Thomson Reuters). At Thomson, the major responsibilities were to develop Knowledge Management tools, middleware components, and the search engine for citation database. His recent work in Text Mining addresses automatic database selection, entity and relation extraction, high speed document filtering, algorithms that learn a person's information needs from experience, automatic analysis of gathered information. He is also involved in a variety of information visualization projects. Prof. Song is also interested in information and knowledge management in large organizations. He is currently interested in applying Text Mining algorithms to Bioinformatics and Social Media.