텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석

Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques

저자 (Authors)
배정환, 손지은, 송민
Jung-hwan Bae, Ji-eun Son, Min Song

출처 (Source)
지능정보연구 19(3), 2013.9, 141-156(16 pages)
Journal of Intelligence and Information Systems 19(3), 2013.9, 141-156(16 pages)

발행처 (Publisher)
한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE02246458

APA Style
배정환, 손지은, 송민 (2013). 텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석. 지능정보연구, 19(3), 141-156

ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2013 September: 19(3): 141~156 http://dx.doi.org/10.13088/jiis.2013.19.3.141

텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석

배정환
연세대학교 문헌정보학과 대학원
(haruaki.pn@gmail.com)

손지은
연세대학교 문헌정보학과 대학원
(sekert@hanmail.net)

송민
연세대학교 문헌정보학과 부교수
(min.song@yonsei.ac.kr)

최근 소셜미디어는 전세계적 커뮤니케이션 도구로서 사용에 전문적인 지식이나 기술이 필요하지 않기 때문에 이용자들로 하여금 콘텐츠의 실시간 생산과 공유를 가능하게 하여 기존의 커뮤니케이션 양식을 새롭게 변화시키고 있다. 특히 새로운 소통매체로서 국내외의 사회적 이슈를 실시간으로 전파하면서 이용자들이 자신의 의견을 지인 및 대중과 소통하게 하여 크게는 사회적 변화의 가능성까지 야기하고 있다. 소셜미디어를 통한 정보주체의 변화로 인해 데이터는 더욱 방대해지고 '빅데이터'라 불리는 정보의 '초(超)범람'을 야기하였으며, 이러한 빅데이터는 사회적 실제를 이해하기 위한 새로운 기회이자 의미 있는 정보를 발굴해 내기 위한 새로운 연구분야로 각광받게 되었다. 빅데이터를 효율적으로 분석하기 위해 다양한 연구가 활발히 이루어지고 있다. 그러나 지금까지 소셜미디어를 대상으로 한 연구는 개괄적인 접근으로 제한된 분석에 국한되고 있다. 이를 적절히 해결하기 위해 본 연구에서는 트위터 상에서 실시간으로 방대하게 생성되는 빅스트림 데이터의 효율적 수집과 수집된 문헌의 다양한 분석을 통한 새로운 정보와 지식의 마이닝을 목표로 사회적 이슈를 포착하기 위한 실시간 트위터 트렌드 마이닝 시스템을 개발 하였다. 본 시스템은 단어의 동시출현 검색, 질의어에 의한 트위터 이용자 시각화, 두 이용자 사이의 유사도 계산, 트렌드 변화에 관한 토픽 모델링 그리고 멘션 기반 이용자 네트워크 분석의 기능들을 제공하고, 이를 통해 2012년 한국 대선을 대상으로 사례연구를 수행하였다. 본 연구를 위한 실험문헌은 2012년 10월 1일부터 2012년 10월 31일까지 약 3주간 1,737,969건의 트윗을 수집하여 구축되었다. 이 사례연구는 최신 기법을 사용하여 트위터에서 생성되는 사회적 트렌드를 마이닝 할 수 있게 했다는 점에서 주요한 의의가 있고, 이를 통해 트위터가 사회적 이슈의 변화를 효율적으로 추적하고 예측하기에 유용한 도구이며, 멘션 기반 네트워크는 트위터에서 발견할 수 있는 고유의 비가시적 네트워크로 이용자 네트워크의 또 다른 양상을 보여준다.

논문접수일 : 2013년 05월 25일 게재확정일 : 2013년 07월 23일
투고유형 : 학술대회우수논문 교신저자 : 송민

1. 서론

최근 페이스북이나 트위터와 같은 소셜미디어는 전세계적인 커뮤니케이션 도구로서 중요한 자리를 차지하고 있다. 특히 새로운 소통매체로서 소셜미디어는 국내외의 사회적 이슈를 실시간으로 전파하면서 이용자들이 자신의 의견을 지인 및 대중과 소통하게 하여 크게는 사회적 변화의 가능성까지 야기하고 있다.

소셜미디어는 사용에 전문적인 지식이나 기술

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

이 필요하지 않기 때문에 수동적인 콘텐츠 소비자(Consumer)였던 기존의 대중들은 지식과 정보의 생산과 소비를 동시에 주도하는 프로슈머(Prosumer) 즉, 생산적 소비자로 거듭나게 되었으며, 스마트폰과 같은 정보통신기기의 발달은 이용자들로 하여금 콘텐츠의 실시간 생산과 공유를 가능하게 하여 기존의 커뮤니케이션 양식을 새롭게 변화시켰다. 소셜미디어를 통한 정보주체의 변화로 인해 데이터는 더욱 방대해지고 '빅데이터'라 불리는 정보의 '초(超)범람'을 야기하였다. 이러한 빅데이터는 사회적 실제를 이해하기 위한 새로운 기회로서 각광받게 되었으며, 이를 효율적으로 요약, 정리 및 추론하여 가시화하기 위한 데이터마이닝, 네트워크 분석, 오피니언 마이닝 등 다양한 연구와(Kim et al., 2012) 소셜미디어를 활용한 상업화가 활발히 이루어지고 있다. 소셜미디어에 관한 초기 연구들은 마이크로블로깅(microblogging)으로서 소셜미디어의 이용과 커뮤니티의 구조를 이해하고자 하는 시도들이 주를 이루었다(Java et al., 2007; Huberman et al., 2008). 그 이후로는 트위터를 구두 마케팅(viral marketing)의 일환으로 분석하는 연구 Jansen et al.(2009)와 같이 응용 연구들이 있었다. 다른 한편, 마이크로블로그를 포함한 웹블로그를 정치적 토론과 논의가 이루어지는 일종의 포럼으로서의 잠재력을 평가한 연구(Jansen and Koop, 2005; Woodley, 2007)도 진행되어 왔다. 최근에는 정치적 특색을 지닌 마이크로블로깅이 그들의 실제 정치상황에 어떤 영향을 미치는 가에 대한 연구들(Drezner and Farrell, 2007; Williams and Gulati, 2008)뿐만 아니라 실제 선거의 결과를 예측하는 도구로서 소셜미디어를 활용하는 연구들(Tumasjan et al., 2010; Livne et al., 2011)이 활발하게 이루어져 왔다. 이처럼 소셜미디어는 실시간으로 생성되는 방대한 양의 텍스트 데이터 외에도 의제설정 및 여론형성 등 오피니언 리딩(opinion leading)을 위한 영향력 있는 채널로 사용되고 있다는 점에서, 그 출현과 함께 정보검색과 텍스트 마이닝의 주요한 연구대상이 되어왔다.

그러나 지금까지 소셜미디어를 대상으로 한 대부분의 연구는 토픽 모델링 기법을 이용한 개괄적인 수준에서의 이슈분석, 오피니언 마이닝 및 이용자 간 상호연계성(팔로우, 친구 맺기 등)에 기반을 둔 네트워크 분석 등에 국한되어 있는 실정이다. 따라서 본 연구에서는 기존 연구들의 한계점을 극복하고자 1) 실시간으로 트위터에서 발생하는 데이터를 수집하여 2) '대선'이라는 특정한 이슈를 중심으로 발생하는 콘텐츠, 즉 멘션을 기반으로 트위터 상의 사회적 이슈를 시계열로 추적하였다. 또한 3) 기존의 상호연결성에 기반한 네트워크 분석에서 벗어나 멘션을 기반으로 '대선'이라는 이슈를 중심으로 발생하는 사회적 네트워크의 특성을 규명하고자 하였다.

본 연구의 나머지 부분은 다음과 같다. 제 2장에서는 소셜미디어를 이용한 토픽모델링 및 네트워크 분석에 대한 선행 연구들을 살펴본다. 제 3장에서는 2012년 대선 관련 이슈와 이용자 네트워크 분석을 위한 실험집단 및 설계에 대해 설명한다. 제 4장에서는 실험결과에 대해 보고하며, 연구에 대한 총체적 결론과 후속연구 제안으로 제 5장을 맺는다.

2. 선행 연구

이 장에서는 소셜미디어를 이용한 토픽 모델링, 이용자 네트워크 분석, 대선 결과 예측 등 소셜미디어 마이닝과 관련된 선행 연구들을 살펴본다.

2.1 소셜미디어를 이용한 토픽 모델링과 네트워크 분석 연구

토픽 모델링은 그 출현과 함께 텍스트 마이닝 분야에서 큰 관심을 받아온 연구방법론이다. 특히

Blei(2003)가 제안한 LDA(Latent Dirichlet Allocation)는 토픽 모델링 연구에 있어 표준 도구로 자리잡은 알고리즘이다. 소셜미디어는 공통된 관심사 혹은 사회적인 이슈를 공유하는 이용자들의 연결로 이루어져 있다는 점과 연일 생산되는 방대한 양의 텍스트 데이터 때문에, 다양한 함의를 가질 수 있는 잠재적인 정보를 발굴하기 위한 연구의 장으로 각광받아 왔다. Chang et al.(2009)은 Wikipedia 등재된 엔티티들과 그들의 관계를 추론하기 위해 토픽 모델링을 사용하였으며, Zhao et al.(2011)은 트위터 상에 출현하는 토픽의 분포와 트윗을 수집한 같은 기간 내 New York Times 기사의 주제 분포를 비교하여 트위터의 뉴스 매체적인 특성을 확인하였다. Gam and Song(2012)은 텍스트 마이닝 기법을 활용하여 신문기사의 키워드 단순빈도 분석과 Clustering, Classification 결과를 분석하여 제시하였으며, 신문사 간 차이점을 분석하고자 하였다. 본 연구에서도 2012년 대선과 관련한 트윗에서 추출한 토픽과 그 변화 양상이 실제 사건 혹은 신문기사와 어떠한 관계를 갖는지 분석 하였다. Choi et al.(2011)는 마이크로블로그를 통한 단어의 동시 발생빈도를 이용하여 토픽 연관 단어를 추출하고 단어 간 연관 정도를 파악하였다. 단어의 동시출현 빈도를 통해 토픽을 추출하는 방법은 시간과 비용면에서 효율적이며, 빠르게 변화하는 소셜미디어의 토픽을 추적하는데 유효함을 확인할 수 있었다. 따라서 본 연구에서는 동시출현 단어 검색 기능을 트위터 마이닝 시스템에 추가하여 빈도순으로 2,000개까지 출력이 가능하도록 하였다.

한편, 정량적인 분석에 초점을 맞춘 소셜미디어 이용자 네트워크 분석에 관한 연구도 꾸준히 수행되었다. Cha et al.(2010)은 트위터에서 어떠한 이용자가 유력자가 될 수 있는지를 알아보기 위해 개별 이용자의 팔로워(follower) 수, 트윗이 리트윗(retweet)된 횟수, 멘션(mention)된 횟수 등 다양한 기준을 적용하여 이용자 네트워크를 분석하였으며, 트위터의 팔로어의 수가 영향력을 측정하는 절대적 지표일 수 없다는 결론을 제시했다. Kwak et al.(2010)의 연구에서는 약 4,170만 명의 트위터 이용자 프로파일과 1억 6백만 개의 트윗을 수집, 분석하여 트위터 유저의 팔로워 수와 페이지랭크(PageRank) 값, 리트윗의 개수가 트위터 내에서의 영향력과 어떤 관계가 있는지 밝혀내었다. Sohn et al.(2012)의 연구에서는 약 16만 명으로 구성된 SNS에서의 실제 데이터를 이용하여 매개 중심성 분석과 근접 중심성 분석을 수행하여 사용자 그래프에서 사용자간 최단거리를 빠르게 찾기 위한 휴리스틱 기반의 최단 경로 탐색 방법을 제안하였다.

소셜 네트워크 서비스의 이용자간 유사도나 친밀도를 측정하기 위한 연구도 진행되고 있다. Lee et al.(2010)의 연구에서는 기존의 연구가 친밀도에 영향을 미치는 다양한 요소들의 상호관계를 고려하지 않는다는 문제점을 지적하고, 이를 반영하기 위해 친밀도, 용어유사도, 관계유사도를 산정하여 합산해 상호관계를 측정하는 기법을 제안하였다. Seol et al.(2011)는 트위터에서 이용자들이 현실세계에서 직접적으로 친분이 없더라도 정보를 얻기 위해 유명인이나 회사를 상대로 일방적인 친구관계를 맺고 있다는 것에 착안하여, 공통의 친구 관계와 메시지 교환 정보를 이용한 이용자 간의 친밀도를 측정하고자 하였다. 한편, 이용자의 성향이나 이용자 간의 친밀도를 알 수 있으면 이용자의 영향력을 더 정확하게 파악할 수 있기 때문에 소셜 네트워크 서비스에서 영향력이 큰 이용자를 찾기 위한 연구들도 활발하게 진행되고 있다. 페이지랭크 알고리즘은 영향력이 큰 이용자를 찾기 위한 해결책의 하나로서 사용되고 있어 Song et al.(2007)의 연구에서는 영향력이 큰 블로거를 찾기 위해 페이지랭크 알

고리즘을 적용하였고, Weng et al.(2010)은 트위터에서 이용자간의 주제적 유사도와 링크 구조를 통해 영향력을 측정하기 위해 페이지랭크 알고리즘을 확장한 TwitterRank라는 알고리즘을 제안하였다.

Xu et al.(2007)은 기존의 네트워크 분석방법과는 달리 노드 군집과 군집 사이를 연결하는 중요한 노드인 '허브(hub)'를 탐지하는 동시에, 상호 연결 정도가 낮아 네트워크상에서 중요한 위치를 차지하지 못하는 아웃라이어(outlier)를 제거해내기 위한 SCAN (Structural Clustering Algorithm for Networks) 알고리즘을 제안하였다. Erjia Yan et al.(2012)의 연구에서는 토픽 모델링과 커뮤니티 디텍션(community detection)을 통합한 하이브리드 기법을 사용하여 연구 주제와 학술 커뮤니티간의 관계를 측정하고자 하였다. 정보검색 분야를 대상으로 데이터를 수집하였으며 연구 결과 공저자 네트워크의 topology를 통해 탐지된 커뮤니티와 토픽 모델을 통해 탐지된 커뮤니티들의 토픽이라는 두 개의 계층이 구성되어 있음을 발견하였다. 이 연구에서 사용된 하이브리드 접근법은 토픽과 커뮤니티간의 유동적인 상호작용을 분석하기에 적합하기 때문에, 본 연구에서 멘션 기반으로 구축된 이용자 네트워크의 특성을 밝히기 위해 커뮤니티 디텍션 기법을 사용하여 비가시적 커뮤니티(invisible community)를 추출하였다.

2.2 소셜미디어를 이용한 사회적 이슈 및 대선 예측 연구

페이스북이나 트위터와 같은 소셜미디어를 이용하여 선거의 결과를 예측하고자 하는 연구는 전 세계적으로 활발하게 진행되고 있다. Williams and Gulati(2008)에 의하면 특정 선거 후보자를 지지하는 페이스북 팬의 수는 실제로 그 후보에게 투표하는 투표자의 수에 영향을 미칠 수 있음을 밝히고, 소셜미디어의 지지는 일반 유권자 중에서도 특히 젊은 유권자들의 유력한 후보자를 예측할 수 있는 지표로 사용될 수 있다고 하였다. O'Connor et al.(2010)의 연구에서는 간단한 감성 분석 기법(sentiment analysis methods)을 사용하여 트위터가 대중의 의견을 자동적으로 측정할 수 있는 도구임을 증명하고자 하였다. 그러나 연구결과를 2008년 미 대선의 사전 여론조사와 대통령 지지도 지수(index of Presidential Job Approval), 소비자 감성 지수(index of Consumer sentiment)와 같은 다른 전통적인 측정 도구와 비교하였을 때, 감성 분석은 아직 불명확한 분야로 두드러질 정도의 의미는 갖지 못한다고 결론을 내렸다. Tumasjan et al.(2010)은 독일 연방선거를 대상으로 텍스트 분석 소프트웨어인 LIWC (Linguistic Inquiry and Word Count)을 이용하여 트위터가 현실세계의 정치적 관점을 보여줄 수 있는지에 대해 연구하였다. 약 100,000건의 트윗 메시지를 분석한 결과, 트위터의 정치적 감성(Political sentiment)은 실제 정당이나 정치인의 위치와 유사함을 보였으며 상당수의 멘션이 선거 결과를 반영하고 있음을 밝혀냈다. Livne et al.(2010)의 연구에서는 2010년에 치러진 상, 하의원과 주지사 선거를 대상으로 일반 대중이 아닌 선거 후보자가 작성한 트윗만을 사용하였는데, 88%의 정확률로 당선여부를 예측하는데 성공하였다. 그러나 위의 두 연구 모두 데이터로 사용된 모든 트윗 메시지를 하나의 문서로 보고 트윗이 생성된 날짜에 대해서는 전혀 고려하고 있지 않았으며, Livne et al.(2010)의 연구의 경우 LDA를 이용하여 토픽 분석을 시도하였으나 충분하지 않은 트윗의 길이로 높은 질의 토픽을 생성하지 못하였다고 밝혔다. 이에 본 연구는 트윗 멘션의 콘텐츠를 분석하기 위해 LDA 알고리즘을 이용한 토픽 모델링의 결과를 시각화하고 시간의 흐름에 따른 정치적 논의의 변화를 측정하고자 하였다.

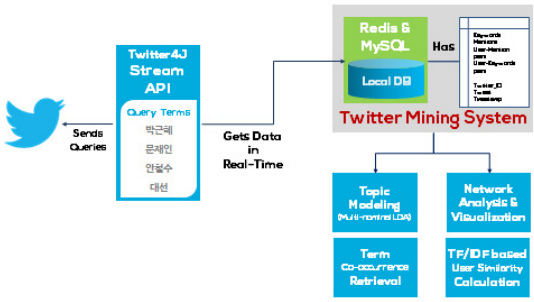
선거 및 정치를 대상으로 한 국내의 소셜미디어 연구동향으로는 Choi and Kim(2012)의 연구가 있는데, 이들은 2011년 정부와 여·야당, 시민이 대립했던 대학교 등록금 인하 문제와 관련하여 블로그와 트위터를 통해 의제가 설정되고 논의되는 방식을 연관어 분석과 내용 분석의 데이터 마이닝 기술을 이용하여 분석하였다. 그러나 이 연구는 블로그와 트위터의 내용이나 형식, 이슈가 확산되고 소멸되는 추이 등 두 매체의 차이를 밝혀내는데 그 초점이 맞추어져 있다. 또한 연관어 추출이나 날짜 별 생성 트윗의 통계를 측정하는 등 비교적 단순한 기법을 사용하는데 그쳐 텍스트 마이닝 분석 측면에서 부족한 점이 다소 있다. Kim(2011)는 트위터를 통해 선거와 정치에 대한 여론을 파악하고 방향을 예측할 수 있는 방법의 가능성에 대해 2011년 서울시장 재보선을 대상으로 네트워크 이론을 적용하여 분석하였다. 리트윗과 멘션으로 가시화한 네트워크의 구조를 통해 당시 서울시장 후보인 박원순 후보에 대한 의견은 보편적으로 유사한 반면 나경원 후보에 대한 의견은 다양하여 첨예한 대립이 있을 가능성이 있다고 밝혔다. 그러나 네트워크의 구조는 여론이 무엇을 의미하는지는 설명해 주지 못하기 때문에 페이지랭크를 이용하여 매개 중심성이 큰 행위자를 찾아내 주요한 여론의 성향을 파악하고자 하였다.

3. 연구 방법론

이 장에서는 본 연구에서 개발한 트위터 마이닝 시스템에 대해 소개하고, 이를 위한 실험문헌의 수집과 실험설계에 대해 설명한다.

3.1 트위터 마이닝 시스템과 데이터 수집

본 연구에서는 트위터 상에서 실시간으로 방대하



<Figure 1> Overview of Twitter Mining System

게 생성되는 데이터의 효율적 수집과 수집된 문헌의 다양한 분석을 통한 새로운 정보와 지식의 마이닝을 목표로 실시간 트윗의 수집과 토픽 모델링 및 이용자 네트워크 시각화 등의 기능을 특징으로 하는 트위터 마이닝 시스템을 개발하였다. 본 시스템은 트위터 한국어 홈페이지(http://www.twitter.com/)에 게재된 트윗과 이용자 ID, 그리고 해당 트윗이 게재된 시간을 Open Java Library인 Twitter4J의 Twitter Stream API를 이용하여 수집하고 키워드와 멘션 뿐만 아니라 멘션 셋과 이용자 쌍, 키워드 셋과 이용자 셋을 Open Source DB인 Redis(http://redis.io)로 구축한 로컬 데이터베이스에 실시간으로 저장한다. Redis 외에도 MySQL 관계형 데이터베이스를 구축하여 트윗들과 트윗이 게재된 시간정보를 추후 분석을 위하여 저장하였다. 키워드를 저장하기 위해 중요한 용어를 추출하는 확률 기반의 한국어 언어 형태 분석기인 KTS(http://kldp.net/projects/kts/)를 사용하였다. KTS가 높은 수준의 정교한 확률기반의 분석을 제공함에도 트윗은 비정형화된 개인적인 용어들이나 전문용어들을 다수 포함하고 있기 때문에 어휘 사전을 이에 맞게 수정하였다.

본 연구를 위한 실험문헌은 위의 시스템을 통해 2012년 10월 1일부터 2012년 10월 31일까지 약 3주간 트위터 한국어 홈페이지에 게재된 트윗 중 본문

에 “박근혜”, “문재인”, “안철수”, “대선”이라는 단어가 출현한 1,737,969건의 트윗을 수집하여 구축되었다.

3.2 텍스트 마이닝 기법

이 장에서는 본 연구에 사용된 텍스트 마이닝 기법에 대해서 다형 토픽 모델링(Multinomial Topic Modeling)과 소셜 네트워크 분석을 위한 커뮤니티 디텍션 기법을 중심으로 설명한다.

3.2.1 다항 토픽 모델링 기법

정보검색과 텍스트 마이닝 등의 영역 등에서 문헌 모델링이란 개별 문헌과 문헌집단을 해당 문헌에 출현한 단어를 통해 표현하는 방법을 의미한다. 토픽 모델링이란 이러한 문헌 모델링의 한 방법으로서 2003년 Blei에 의해 제안된 초창기 토픽 모델링 기법중의 하나인 LDA(Latent Dirichlet Allocation) 알고리즘을 기반으로 하는 절차적 확률분포 모델이다. 즉, LDA에서 문헌은 특정 확률에 의해 선택된 단어들로 구성된 토픽들의 집합으로 표현된다. 본 연구에서 사용된 토픽 모델링 기법은 Mimno와 McCallum가 2008년 제안한 Dirichlet-multinomial regres sion(DMR)이다. DMR은 Blei의 LDA를 확장한 것으로 문헌-주제 분포에 기반한 long-linear prior를 포함함으로써 저자나 발행처, 참고문헌, 날짜와 같은 문헌의 문헌의 특성들을 임의로 조절 가능하게 한다.

3.2.2 이용자 네트워크 분석 기법

소셜 네트워크 분석이란 개인, 집단, 사회의 관계를 네트워크로 파악하는 연구방법론이다. 현재 인문, 경제, 공학, 웹 사이언스 등 다양한 분야에서 소셜 네트워크 분석과 관련한 많은 연구가 진행되어 왔다. 그러나 기존의 소셜미디어를 네트워크 분석의 대상으로 한 연구들의 대부분은 트위터의 팔로우/팔로잉 등으로 대표되는 인맥을 분석단위로 하기 때문에, 단방향 관계 형성이 가능한 트위터 네트워크의 특성을 적절히 반영하지 못하였다. 또한 실세계에서 발생하는 사회적 이슈와 그 변화에 따른 이용자 네트워크의 동적인 변화를 밝혀내지 못한다는 한계점을 가진다. 따라서 본 연구에서는 콘텐츠, 즉 사회적 이슈의 출현과 변화를 반영하는 트위터 이용자 네트워크를 추출하고 그 특성을 밝혀내기 위해 이용자들이 작성한 멘션과 그 멘션의 방향성을 분석의 기본단위로 하였다. 네트워크 분석과 시각화 등을 위하여 Open Source Java 라이브러리인 JUNG(Java Universal Network/Graph Framework)을 사용하였으며, JUNG 패키지에서 제공하는 커뮤니티 디텍션과 페이지랭크 알고리즘을 이용하여 2012년 대선과 관련하여 특정한 이슈를 다루는 비가시적 커뮤니티와 영향력 있는 이용자의 특성을 밝혀내고자 하였다. 커뮤니티 디텍션은 네트워크 구조를 분석하기 위한 효율적인 도구로서 거대한 집단의 특성을 탐색할 수 있을 뿐만 아니라 복잡한 네트워크에 대해 더 나은 설명을 제공한다. 본 연구에서는 규모가 큰 소셜 네트워크 분석에 여타의 커뮤니티 디텍션 기법보다 빠르고 효율적인 SCAN(Structural Clustering Algorithm for Networks) 알고리즘을 사용하였다.

실험 결과 4,668,343개의 에지와 136,754개의 노드로 구성된 1,537개의 컴포넌트가 멘션 기반 트위터 네트워크상 나타났다.

4. 사례연구 : 2012년 한국 대통령 선거를 중심으로

이 장은 트위터 마이닝 시스템을 이용하여 2012

년 한국 대선을 대상으로 진행한 사례 연구이다. 본 사례연구는 트위터 데이터를 기반으로 질의어에 의한 단어 동시출현빈도 비교, 토픽 모델링, 멘션 기반의 이용자 네트워크 분석을 통해 이루어진다.

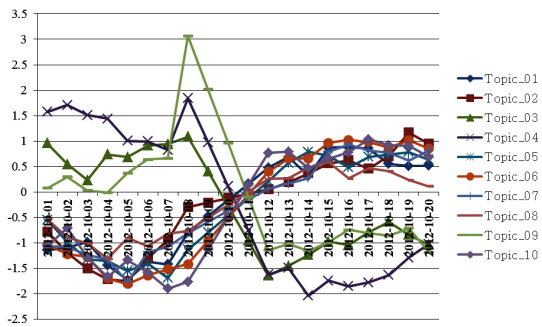
4.1 검색어에 의한 용어의 동시출현빈도 비교

다음은 유력한 대선 후보인 “박근혜”, “문재인”, “안철수” 세 후보의 이름을 질의어로 사용하여 도출한 동시 출현 단어의 일부분이다.

<Table 1> 에서 대선, 지지선언, 여론조사, 각 대선 후보의 이름과 같은 대선과 연관된 일반적인 용어들이 고빈도어로 나타나는 것을 알 수 있다. 반면에 굵은 글씨로 표시되어 있는 질의어 “박근혜”의 박정희와 육영수, 질의어 “문재인”의 투표시간 연장과 후보단일화, 질의어 “안철수”의 다운계약서, 후보단일화는 각 후보를 구분 짓는 특징적인 단어로 볼 수 있다. 그리고 위의 용어들과 연관되어 있는 일련의 사건들은 실제로 각 후보들의 핵심 이슈이기도 하다. 특히, 안철수 후보와 문재인 후보는 단일화, 투표시간연장, 개념인터뷰 와 같이 서로 중복되는 단어가 박근혜 후보에 비해 많이 나타나며, 서로의 이름 또한 2위와 4위에 위치하는 등 두 후보는 트윗의 내용적 측면에서 유사하며 관련성이 높은데 이는 “문재인”, “안철수” 두 후보가 후보 단일화를 의논해 왔기 때문이다.

4.2 다항 토픽 모델링

시계열에 따라 분석된 10개 토픽의 변화 추이는 다음과 같다. 각각의 토픽은 확률 분포의 변화에 따라 Rising issue와 Falling issue의 두 가지 추이(tendency)를 보였으며, 7개의 토픽(topic #1, 2, 5, 6, 7, 8, 10)은 Rising 패턴을, 나머지 3개의 토픽(topic #3, 4, 9)은 Falling 패턴을 보인다.



<Figure 2> Topic Trends Overview

<Table 1> Term Co-occurrence Comparison by Queries

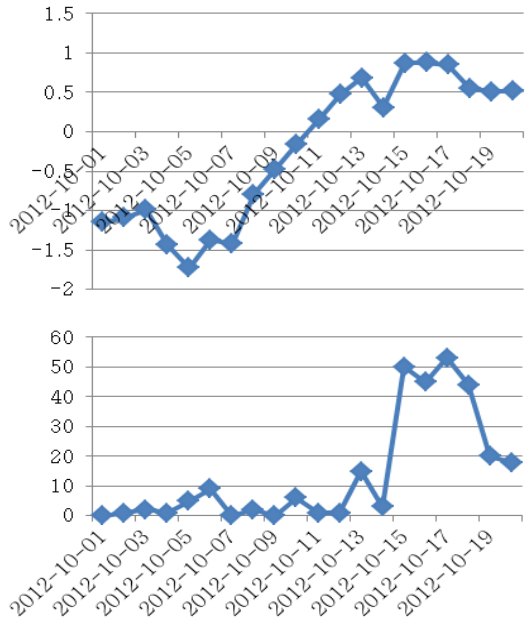
	박근혜		안철수		문재인	
1	대선	634540	문재인	62770	문재인	143663
2	지지선언	589728	안철수	53554	대선	92062
3	박근혜	530307	박근혜	38750	지지선언	62766
4	여론조사	492762	대선	35468	안철수	57420
5	지지	82428	지지선언	11672	여론조사	54494
6	선언	80904	여론조사	10630	박근혜	49785
7	무소속안철수	52990	민주통합당	6858	민주통합당	8552
8	문재인	46614	새누리당	6752	새누리당	7818
9	안철수	40078	문재인캠프	6540	문재인캠프	7796
10	snspage	38748	김정숙	5740	**투표시간연장**	**7192**
11	**육영수**	**20326**	**다운계약서**	**4678**	문재인tv	5916
12	**박정희**	**13059**	**단일화**	**3240**	김정숙	5768
13	박근혜정책	10210	개념인터뷰	2702	**단일화**	**4140**

<Table 2> Topic Analysis

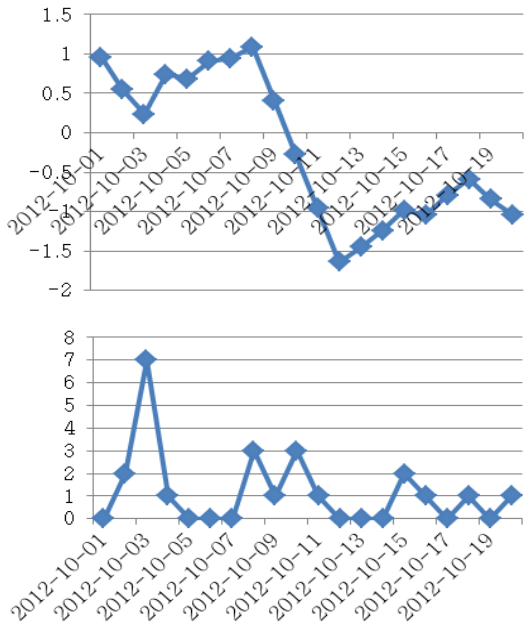
Topic	Description		
	Label	Terms	Type
01	정수 장학회	**박근혜, 정수장학회, 안철수, 대선, 문재인, MBC, 최필립, 새누리당, 부산일보**	R
02	대선 후보	**박근혜, 문재인, 안철수, 후보, 대선, 대통령**	R
03	박근혜 지지율	**박근혜**, 후보, 안철수, 대선, 새누리당, 문재인, **단일화**, 대통령, **지지율**, 선거	F
04	안철수 의혹	**안철수**, 박근혜, **논문**, **표절**, **의혹**, **다운계약서**, **서울대**	F
05	대선 후보	**박근혜, 문재인, 안철수, 후보, 대선, 대통령**	R
06	후보 단일화	**안철수**, 박근혜, **문재인**, 대선, 후보, **무소속**, **단일화**	R
07	박근혜 슬로건	**박근혜**, 문재인, 안철수, **나라**, **내**, **꿈이**, **이루어지는**	R
08	박근혜 캠프 구성	**박근혜**, 문재인, 안철수, 민주당, **캠프**, **김경재**, **이**, **대선**, **장악한**, **종북세력**, **막으러**, **들어왔다**	R
09	대선 후보	**후보, 박근혜, 안철수, 새누리당, 무소속, 민주통합당, 문재인**	F
10	NLL 포기 의혹	**박근혜**, **문재인**, **NLL**, 안철수, **노무현**, **정문헌**, **민주통합당**	R

트위터 상에 나타난 대선 관련 토픽 변화가 실제 사회적 이슈와 어떠한 관계를 갖는지 알아보기 위해 같은 기간 관련된 신문기사 및 실제 사건에 대한 비교 분석을 실시하였다.

첫 번째 토픽인 '정수장학회'에 관련된 트위터 상 토픽 변화와 신문기사의 양은 Rising tendency를 보였다. 특히 10월 15일부터 정수장학회 의혹 및 이사진 사퇴에 관련한 기사들이 집중 보도 되었는데, 트위터에서는 이보다 빠른 10월 10일경부터 관련된 트윗의 증가가 이루어지고 있다. 이를 통해 활발히 논쟁이 되는 이슈가 신문기사보다 트위터에서 더욱 빠르게 전파되고 있음을 확인할 수 있다.



<Figure 3> Comparison between Topic1 (Up) and Related News Articles (Down)



<Figure 4> Comparison between Topic4 (Up) and Related News Articles (Down)

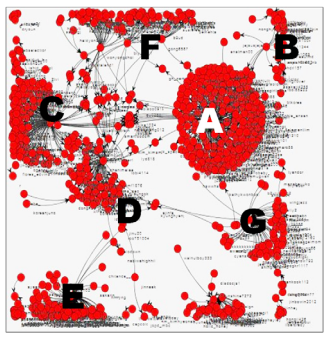
네 번째 토픽인 '안철수 의혹'에 관련된 트위터 상 토픽 변화와 신문기사의 양은 Falling tendency를 보였다. 해당 토픽과 관련하여 이 기간 동안에는 MBC의 안철수 후보와 관련된 의혹에 관한 보도(10월 1일)와 이에 대한 해명(10월 2일)이 있었다. 이에 따라 해당 기간 동안 트위터와 신문기사에서 모두 관련 토픽을 언급하였다. 또한 안철수 후보의 세금 탈루 등의 의혹(10월 8일), 군복무 기간 논문 발표에 대한 의혹(10월 15일) 등이 국정감사에서 거론되어 일시적인 토픽 및 신문기사 양의 변화를 보였다. 하지만 10월 7일 안철수 후보가 의혹을 해명하기 위한 기자회견을 함으로써 해당 토픽은 전반적으로 Falling tendency를 보인다. 위 분석을 통해 신문기사는 해당 토픽에 대한 속보성과 세부 사건 변화의 관찰에 적합하며, 트위터는 전체 맥락에서 해당 토픽의 영향력 변화를 관찰하는데 유용함을 확인할 수 있다.

4.3 네트워크 분석

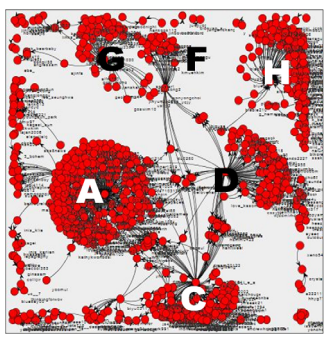
기존 소셜 네트워크 서비스는 이용자들의 실제 사회적 관계를 바탕으로 쌍방향적 관계를 유지하는 것에 초점을 두고 있다. 그러나 트위터에서는 관심이 있는 정치인이나 연예인 또는 기업 등의 계정들을 팔로우하여 상대방의 동의가 없어도 가능한 단방향적 관계를 형성한다. 즉, 트위터의 팔로우 및 팔로잉 관계는 친분을 쌓기 위한 사회적인 관계를 의미하는 것이 아니라, 트윗이라는 하나의 '매체'를 구독하기 위한 소유의 기능을 한다고 볼 수 있다. 따라서 본 연구에서는 기존의 상호연결성에 기반한 연구들이 트위터 상 네트워크를 분석하는데 갖는 한계점을 극복하기 위해 콘텐츠, 즉 멘션에 동시 출현한 자질과 그 빈도를 기반으로 대선과 관련된 트윗을 게재한 총 136,754명의 이용자 네트워크를 분석하였다. 또한 멘션을 기반으로 완전히 상호 연결된 이용자 집단을 찾아내기 위해 커뮤니티 디텍션 기법을 사용하였다.

4.3.1 멘션 기반 이용자 네트워크 분석

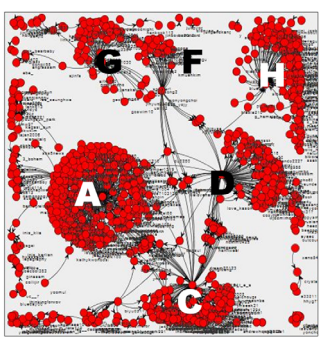
2012년 대선 후보 중 영향력 있는 후보로 평가되던 박근혜, 문재인, 안철수 후보를 질의어로 선택하여 대선과 관련한 멘션 기반 이용자 네트워크 분석을 실시하였다. 추출된 네트워크의 시각화 결과는 다음과 같다.



Query "박근혜" Network



Query "안철수" Network



Query "문재인" Network

<Figure 5> Mention-based User Network Visualization

각 커뮤니티에 속한 이용자 프로파일과 수집된 멘션 내용에 기반하여 분석한 결과 박근혜와 안철수 네트워크는 상당 부분(커뮤니티 A, C, D, F) 일치하며 박근혜와 문재인 네트워크, 안철수와 문재인 네트워크의 경우에는 C커뮤니티가 일치함을 확인할 수 있었다. 또한, 박근혜 네트워크의 A, B 클러스터와 안철수 네트워크의 A 클러스터, 그리고 문재인 네트워크의 A, B 클러스터의 경우 보수적 정치 성향의 이용자들로 확인되었으며, 나머지 클러스터는 진보 성향 이용자의 집합으로 나타났다. 이를 통해 각 후보의 정치적 정체성과는 달리 이용자들은 자신의 성향과 다른 후보 또한 언급하고 있으며, 보수적 성향의 이용자들은 하나의 큰 군집을 이루며 상호작용하는 반면 진보적 성향의 이용자들은 betweeness가 높은 핵심 노드를 매개로 작은 그룹을 형성하여 연결되어 있음이 확인되었다.

다음은 멘션에서 세 후보 각각을 언급한 이용자들과 두 후보를 동시에 언급한 이용자 ID 통계이다. <Table 3>에서 보다시피, 세 명의 대선 후보 중 두 후보가 동시에 등장하는 멘션의 비율이 매우 높다. 즉, 트위터 이용자들은 자신의 정치적 성향과는 별개로 세 후보 모두를 언급하고 있었다.

<Table 3> The Number Twit Mention by Candidates' Name

박근혜	문재인	안철수
202,780	149,338	202,215
박근혜 & 문재인	박근혜 & 안철수	문재인 & 안철수
184,207	195,013	148,750

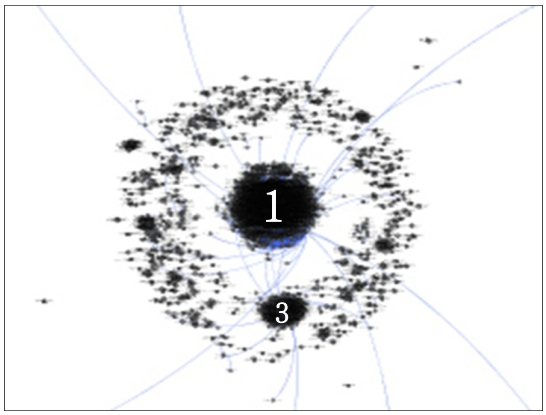
4.3.2 커뮤니티 디텍션

앞 절의 네트워크 분석결과를 바탕으로 완전히 상호연결 된 이용자 집단을 찾아내고 그 특성을 알아보기 위해 커뮤니티 디텍션 알고리즘을 사용하였다. 이용자간 주고 받은 멘션을 기반으로 커뮤니티를 추출한 결과 442개의 커뮤니티가 생성되었으며, 8개의 특징적인 커뮤니티가 관찰되었다. <Table 4>는 추출된 커뮤니티 사이즈(이용자 수), 각 커뮤니티 멘션의 고빈도 출현어, 그리고 토픽모델링과 연계하여 분석한 결과이다.

<Table 4> Major Community Analysis

No.	Size	Frequently Occurred Terms	Related Topic
1	2,995	후보, 박근혜, 안철수, 국민, 대통령, 문재인, 단일화	후보 단일화
3	537	안철수, 후보, 박근혜, 대선, 국민, 단일화, 문재인	후보 단일화
17	81	대선, 흑색선전, 후보자, 깨끗한, 선거문화, 선거	공정선거
11	50	대선, 선거, 비방, 흑색선전, 후보자, 깨끗한	공정선거
33	40	상호존중, 통일정책, 질서, 제정, 실천, 국회	-
74	32	후보, 대선, 대통령, 도발, 평화, 안철수, 박근혜	-
39	24	후보, 박근혜, 대선, 행복, 국민	박근혜 후원
340	22	상처, 기대감, 관광, 여론조사, 전화통화	-

1번과 3번 커뮤니티의 경우 토픽 모델링 결과 출현했던 단어들이 유사하게 출현하였으며, 각 커뮤니티에서 랜덤 표집으로 이용자를 추출하여 프로파일과 멘션을 확인한 결과 1번 커뮤니티에는 보수적 정치 성향의 이용자가, 3번 커뮤니티에는 진보적 정치 성향의 이용자가 주로 존재하는 것으로 확인되었다. 17번과 11번 커뮤니티의 경우 흑색선전, 깨끗한, 선거문화 등 특징적인 단어들이 출현하였으며 해당 이용자의 프로파일과 멘션을 분석한 결과 공정선거홍보 및 지방자치단체 선거관리위원회 계정들로 확인되었다. 39번 커뮤니티는 '박근혜를 사랑하는 모임'이 주최하고 '해핑'이 후원하는 박근혜 대통령 후보 공식 후원 업체로 박근혜 관련 콘텐츠를 생성하는 것으로 확인 되었다. <Figure 6>은 전체 커뮤니티의 시각화 결과이다.



<Figure 6> Community Detection Visualization

5. 결론

본 연구에서는 실시간 트위터 트렌드 마이닝 시스템을 개발하였다. 이 시스템은 1) 실시간으로 트위터에서 발생하는 모든 트윗을 수집, 저장하고 2) 특정 이슈를 중심으로 발생하는 토픽을 시계열로 추적하며 3) 멘션 기반으로 생성된 이용자 네트워크의 특성을 규명하고자 하였다. 본 연구는 최신 기법을 사용하여 트위터에서 생성되는 사회적 트렌드를 마이닝할 수 있게 했다는 점에서 주요한 의의가 있다. 또한 2012 년 한국 대선에 관한 사례 연구를 수행하여 구현한 시스템을 검증해 보았다.

그 결과, 이용자들의 트위터 이용행태는 사회적 이슈의 출현과 변화를 적절히 반영하고 있으며, 특정 이슈의 전체 맥락에서 해당 토픽의 변화를 관찰하는데 유용하다는 점을 볼 수 있었다. 특히, 논쟁적인 이슈의 경우 다른 미디어보다 빠르게 전파됨을 볼 수 있었다. 또한, 멘션 기반 이용자 네트워크를 추출한 결과, 이용자들은 자신의 정치적 성향과 다른 후보에 대해서도 트윗을 작성하고 있었다. 또한 페이지랭크 값을 기준으로 네트워크에서 영향력이 높다고 판단된 이용자의 프로파일과 해당기간에 작성한 트윗을 분석한 결과, 토픽 모델링을 통해 추출된 토픽 전반에 걸쳐 멘션을 주고받고 있음이 확인되었다.

본 연구는 다음과 같은 점에서 한계점을 갖는다. 첫째, 연구를 수행하기 위해 수집한 트위터 데이터 샘플이 한국의 모든 유권자를 대표하지 못한다. 둘째, 본 연구의 데이터는 대선 후보의 이름이나 대선 이라는 질의어를 가지고 있는 트윗으로 한정되어 있다. 따라서 해당 이슈에 관한 논의 임에도 위의 단어들을 포함하고 있지 않은 트윗들은 배제된다. 셋째, 본 연구는 기존의 분석 소프트웨어를 사용하였기 때문에 트위터와 같은 단문 메시지를 분석하는데 특화되어 있지 못하다. 특히 신조어 등 마이크로블로깅에서 사용되는 용어나 문맥이 반영되기가 어렵다.

후속연구에서는 특정 트윗들의 맥락을 좀 더 자세하게 포착하고 놓친 부분들을 포함시킬 수 있는 방법이 연구되어야 할 것이다. 또한 감성 분석(Sentiment Mining) 등을 추가적으로 수행한다면 해당 이슈에 대한 여론의 변화 및 형성 과정을 심도 있게 관찰할 수 있을 것이며, 이를 통해 소셜 이슈 예측 모형(Social Issue Prediction Model)을 구현할 수 있을 것으로 기대된다. 또한, 트위터에 서지 인용 개념을 적용하여 콘텐츠 기반의 영향력자나 오피니언 리더를 발견하거나, 토픽 모델링 결과를 통해 탐색된 커뮤니티 구조의 관련성(co-relation)을 연구할 것이다. 시계열 상 다양한 조건에 따라 변화하는 소셜 네트워크 패턴을 조사해보는 것 또한 흥미로운 연구 과제이다.

참고문헌

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

Cha, M., H. Haddadi, F. Benevenutoz, and K. P. Gummadi, "Measuring user influence in Twitter : the million follower," *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*(2010).

Chang, J., J. Boyd-Graber, C. Wang, S. Gerrish, and D. Blei, "Reading tea leaves : how humans interpret topic models," *Neural Information Processing Systems*(2009), 288~296.

Cho, H. S. and J. Y. Kim, "Political Communication and Civic Participation Through Blogs and Twitter," *Cyber Communication*, Vol.29, No.2 (2012), 95~130.

Choi, D. J., M. H. Min, J. K. Kim, and J. H. Lee, "A Study on topic tracking using microblog," *Proceedings of KIIS Spring Conference*, Vol.21(2011), 80~82.

Drezner, D. W. and F. Henry, "Introduction-blogs, politics and power : a special issue of public choice," *Public Choice*, Vol.134, Issue.1-2(2007), 1~13.

Gam, M. A. 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.

Huberman, B. A., D. M. Romero, and F. Wu, *Social networks that matter : Twitter under the microscope*, SSRN, 2008. Available at http://ssrn.com/abstract=1313405(Accessed 12 May, 2013).

Jansen, B. J., M. Zhang, K. Sobel, and A. Chowdury," Twitter power : tweets as electronic word of mouth," *JASIST*, Vol.60, Issue.11(2009), 1~20.

Weng, J., E. P. Lim, J. Jiang, and Q. He, "Twitterrank : Finding Topic-sensitive Influential Twitterers," *Proceedings of the 3rd ACM international conference on Web search and data mining*(2010), 261~270.

Jansen, H. J. and Koop, R., "Pundits, Ideologues, and Ranters : the british columbia election online," *Canadian Journal of Communication*, Vol.30, Issue.4(2005), 613~632.

Java, A., X. Song, T. Finin, and B. Tseng, "Why we twitter : understanding microblogging usage andcommunities," *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, (2007), 56~65.

Kwak, H. Y., C. H. Lee, H. S. Park, and S. Moon, "What is Twitter, a social network or a news media?," *Proceedings of the 19th International conference on WWW*, (2010), 591~600.

Kim, Y., 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(2012), 143~156.

Kim, Y. H., "Prediction of structure of spread of public opinion at Twitter with special emphasis on by-election for Seoul mayor," *Political Communication*, Vol.23(2011), 103~139.

Lee, K. M., H. Namgoong, E. H. Kim, G. Y. Lee, and H. K. Kim, "Analysis of multi-dimensional interaction among SNS users," *Journal of Korean Society for Internet Information*, Vol.12, No.2(2010), 113~122.

Livne, A., M. Simmons, E. Adar, and L. Adamic, "The party is overhere : Structure and content in the 2010 election," *Proceedings of 5th ICWSM*(2011).

Mimno, D. M. and McCallum, A., "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression," *UAI*(2008), 411~418.

O'Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls : linking text sentiment to public opinion time-series," *Proceedings of 4th ICWSM*(2010),

122~129.

Seol, K. S., J. D. Kim, H. N. Shim and D. G. Baik, "Intimacy measurement between adjacent users in social networks," *Journal of KIISE*, Vol.38, No.2(2012), 335~341.

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.

Song, X., Y. Chi, K. Hino, and B. Tseng, "Identifying Opinion Leaders in the Blogosphere," *Proceedings of the 16th ACM Conference on information and Knowledge Management* (2007), 971~974.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter : what 140 characters reveal about political sentiment," *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*(2010), 178~185.

Williams, C. and G. Gulati, "The political impact of facebook : Evidence from the 2006 midterm elections and 2008 nomination contest," *Politics and Technology Review*, Vol.1(2008), 11~ 21.

Williams, C. and G. Gulati, "What is a Social Network Worth? Facebook and Vote Share in the 2008 Presidential Primaries," *In Annual Meeting of the American Political Science Association*(2008), 1~17.

Xu, X., N. Yuruk, Z. Feng, and T. A. J. Schweiger, "SCAN : a structural clustering algorithm for networks," KDD '2007 *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 824~833.

Yan, E., Y. Ding, S. Milojević, and C. R. Sugimoto, "Topics in dynamic research communities : An exploratory study for the field of information retrieval," *Journal of Informetrics*, Vol. 6, Issue.1(2012), 140~153.

Zhao, W., X. J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li, "Comparing Twitter and Traditional Media Using Topic Models," *Advances in Information Retrieval Lecture Notes in Computer Science*, Vol.6611(2011), 338~349.

Abstract

Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques

Jung-hwan Bae* · Ji-eun Son** · Min Song***

Social media is a representative form of the Web 2.0 that shapes the change of a user's information behavior by allowing users to produce their own contents without any expert skills. In particular, as a new communication medium, it has a profound impact on the social change by enabling users to communicate with the masses and acquaintances their opinions and thoughts. Social media data plays a significant role in an emerging Big Data arena. A variety of research areas such as social network analysis, opinion mining, and so on, therefore, have paid attention to discover meaningful information from vast amounts of data buried in social media. Social media has recently become main foci to the field of Information Retrieval and Text Mining because not only it produces massive unstructured textual data in real-time but also it serves as an influential channel for opinion leading. But most of the previous studies have adopted broad-brush and limited approaches. These approaches have made it difficult to find and analyze new information. To overcome these limitations, we developed a real-time Twitter trend mining system to capture the trend in real-time processing big stream datasets of Twitter. The system offers the functions of term co-occurrence retrieval, visualization of Twitter users by query, similarity calculation between two users, topic modeling to keep track of changes of topical trend, and mention-based user network analysis. In addition, we conducted a case study on the 2012 Korean presidential election. We collected 1,737,969 tweets which contain candidates' name and election on Twitter in Korea (http://www.twitter.com/) for one month in 2012 (October 1 to October 31). The case study shows that the system provides useful information and detects the trend of society effectively. The system also retrieves the list of terms co-occurred by given query terms. We compare the results

* Dept. of Library and Information Science, Yonsei University
** Dept. of Library and Information Science, Yonsei University
*** Corresponding Author: Min Song
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

of term co-occurrence retrieval by giving influential candidates' name, 'Geun Hae Park', 'Jae In Moon', and 'Chul Su Ahn' as query terms. General terms which are related to presidential election such as 'Presidential Election', 'Proclamation in Support', 'Public opinion poll' appear frequently. Also the results show specific terms that differentiate each candidate's feature such as 'Park Jung Hee' and 'Yuk Young Su' from the query 'Guen Hae Park', 'a single candidacy agreement' and 'Time of voting extension' from the query 'Jae In Moon' and 'a single candidacy agreement' and 'down contract' from the query 'Chul Su Ahn'. Our system not only extracts 10 topics along with related terms but also shows topics' dynamic changes over time by employing the multinomial Latent Dirichlet Allocation technique. Each topic can show one of two types of patterns-Rising tendency and Falling tendency-depending on the change of the probability distribution. To determine the relationship between topic trends in Twitter and social issues in the real world, we compare topic trends with related news articles. We are able to identify that Twitter can track the issue faster than the other media, newspapers. The user network in Twitter is different from those of other social media because of distinctive characteristics of making relationships in Twitter. Twitter users can make their relationships by exchanging mentions. We visualize and analyze mention based networks of 136,754 users. We put three candidates' name as query terms-Geun Hae Park', 'Jae In Moon', and 'Chul Su Ahn.' The results show that Twitter users mention all candidates' name regardless of their political tendencies. This case study discloses that Twitter could be an effective tool to detect and predict dynamic changes of social issues, and mention-based user networks could show different aspects of user behavior as a unique network that is uniquely found in Twitter.

Key Words : Social Media Mining, Twitter Trend Miming System, Topic Modeling, Network Analysis, Community Detection, Korean Presidential Election, Big Data

저 자 소 개

배정환

연세대학교 문헌정보학 학사를 취득했으며 동 대학원 석사과정에 재학 중이다. 주요 관심분야는 텍스트 마이닝에 기반한 바이오와 소셜미디어 빅데이터 분석, 디지털 도서관 그리고 HCI이다.

손지은

숙명여자대학교 문헌정보학 및 영어영문학 복수 학사학위를 취득했으며 연세대학교 문헌정보학과 석사과정에 재학 중이다. 주요 관심분야는 소셜미디어 마이닝, 기록관리학, 도서관 서비스 및 이용자 연구이다.

송 민

Prof. Song has a background in Text Mining, Bioinfomatics, Information Retrieval and Information Visualization. Prior to Yonsei, he was an AssociateProfessor 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.