트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구*

Topic-Network based Topic Shift Detection on Twitter

진설아 (Seol A Jin)**
허고은 (Go Eun Heo)***
정유경 (Yoo Kyung Jeong)****
송 민 (Min Song)*****

초 록

본 연구는 높은 접근성과 간결성으로 인해 방대한 양의 텍스트를 생산하는 트위터 데이터를 분석하여 토픽의 변화 시점 및 패턴을 파악하였다. 먼저 특정 상품명에 관한 키워드를 추출한 후, 동시출현단어분석(Co-word Analysis)을 이용하여 노드와 에지를 통해 토픽과 관련 키워드를 직관적으로 파악 가능한 네트워크로 표현하였다. 이후 네트워크 분석 결과를 검증하기 위해 출현빈도 기반의 시계열 분석과 LDA 토픽 모델링을 실시하였다. 또한 트위터 상의 토픽 변화와 언론 기사 검색결과를 비교한 결과, 트위터는 언론 뉴스에 즉각적으로 반응하며 부정적 이슈를 빠르게 확산시키는 것을 확인하였다. 이를 통해 기업은 대중의 부정적 의견을 신속하게 파악하고 이에 대한 즉각적인 의사결정 및 대응을 위한 도구로 본 연구방법을 활용할 수 있을 것으로 기대된다.

ABSTRACT

This study identified topic shifts and patterns over time by analyzing an enormous amount of Twitter data whose characteristics are high accessibility and briefness. First, we extracted keywords for a certain product and used them for representing the topic network allows for intuitive understanding of keywords associated with topics by nodes and edges by co-word analysis. We conducted temporal analysis of term co-occurrence as well as topic modeling to examine the results of network analysis. In addition, the results of comparing topic shifts on Twitter with the corresponding retrieval results from newspapers confirm that Twitter makes immediate responses to news media and spreads the negative issues out quickly. Our findings may suggest that companies utilize the proposed technique to identify public's negative opinions as quickly as possible and to apply for the timely decision making and effective responses to their customers.

키워드: 트위터, 토픽 추적, 동시출현단어분석, 네트워크 기반 분석, 시계열 그래프
LDA, latent Dirichlet allocation, twitter, topic detection, co-word analysis, network-based analysis, time-series graph

---

* 본 연구는 2012년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2012-2012S1A3A2033291).
** 연세대학교 문헌정보학과 대학원(sula.jin@gmail.com)
*** 연세대학교 문헌정보학과 대학원(goeun.heo@yonsei.ac.kr)
**** 연세대학교 문헌정보학과 대학원(yk.jeong@yonsei.ac.kr)
***** 연세대학교 문헌정보학과 부교수(min.song@yonsei.ac.kr) (교신저자)
■ 논문접수일자 : 2013년 2월 15일 ■ 최초심사일자 : 2013년 2월 26일 ■ 게재확정일자 : 2013년 3월 25일
■ 정보관리학회지, 30(1), 285-302, 2013. 〔http://dx.doi.org/10.3743/KOSIM.2013.30.1.285〕

1. 서 론

최근 급격한 이용자 증가를 보이고 있는 트위터는 이용자의 자체적 메시지 작성뿐만 아니라 정보의 링크가 가능한 미디어로써 다양한 출처로부터의 풍부한 정보를 포함하는 중요한 정보원의 역할을 하고 있다(최돈정, 이성우, 김재광, 이지형, 2011). 하지만 전 세계 이용자들에 의해 매우 빠른 속도로 생산되는 트윗의 방대한 데이터를 모두 검토하여 의미 있는 정보를 이끌어 내는 것은 사실상 불가능하다. 따라서 최근에는 트위터 상의 정보를 직관적으로 이해할 수 있도록 하는 연구들이 진행되고 있다(최돈정 외, 2011; 하용호, 임성원, 김용혁, 2012).

모바일로 이용가능한 높은 접근성과 짧은 글자 수로 제한된 간결성의 특성을 가지는 트위터는 기존 매체들에 비해 이용자들의 관심사 변화를 보다 빠르게 반영하기 때문에 기업의 자사 제품 평가, 선거와 관련된 대중 의견 파악 등을 위해서 활발히 사용되고 있다(O' Connor, Balasubramanyan, Routledge, & Smith, 2010; Tumasjan, Sprenger, Sandner, & Welpe, 2010; Bermingham & Smeaton, 2011). 특히 이윤을 목적으로 하는 기업은 보다 신속하고 정확한 의견 파악을 필요로 한다. 전선규(1996)는 제품에 대해 부정적인 경험을 겪은 소비자에게 제품의 상세정보를 제공하여 이성적인 판단을 유도하도록 하며, 이를 통해 효과적인 의사소통을 이끌어 낼 수 있음을 강조하면서 기존 고객 및 잠재고객의 의견 전파와 이에 대한 신속한 대처의 중요성을 역설하였다. 또한 길이가 긴 문헌을 대상으로 한 기존의 분석 방법들과 달리 트위터는 짧은 문장으로 구성되어 있기 때문에 의미 있는 정보를 추출하는 데에 보다 간결한 방법을 활용하여 분석의 효율성을 높일 수 있다. 위와 같이 간결성과 즉시성을 가지는 트위터의 특성을 고려하여 본 논문에서는 특정 제품에 대한 토픽 키워드를 추출하여 계량정보학 분야에서 내용 분석을 위해 많이 쓰이는 방법인 동시출현단어분석을 이용하였다. 일반적으로 동시출현단어분석(Co-word analysis)은 특정 문헌의 단어가 동시에 출현한 패턴을 분석하여 해당 문헌이 가지고 있는 지적구조를 파악하는 방법이다. 이를 바탕으로 하나의 트윗에 동시에 출현한 단어의 쌍을 이용하여 트위터 상의 토픽 키워드와 토픽이 변화하는 시점 및 패턴을 파악할 수 있으며, 키워드 간의 조직적 관계가 표현된 네트워크를 이용하여 시간의 흐름에 따라 변하는 토픽 추적 방법을 제안하였다. 확률에 의존한 일반화의 한계를 가지는 양적 분석 기법들 중에서 관계의 복잡성을 반영하는 기법으로서 많이 쓰이는 네트워크 분석은 노드(node)와 에지(edge)를 통해 토픽에 해당되는 의견의 구조적 관계를 파악하고 관련 키워드를 살펴볼 수 있다. 또한 기존의 오피니언 마이닝, 감정 분석과 같은 의견추출 기법의 극성(Polarity) 판단과 같은 절차를 거치지 않아도 되는 장점을 가진다. 이러한 결과를 검증하기 위하여 키워드 출현 빈도의 시계열 분석과 토픽 모델링에서도 일관된 결과를 보이는 지를 확인하였다.

본 연구에서는 특정 상품명을 포함하는 트윗을 추출하여 각 트윗당 단어의 동시출현빈도를 기반으로 네트워크를 그린 후, 토픽 변화를 추적하였다. 또한 포털 검색엔진의 언론 기사와의 비교 분석을 통해 이슈 검증 및 상품

을 둘러싼 트위터 상의 토픽 변화가 어떠한 특성을 나타내는 지도 파악하였다. 본 연구는 이를 통해 최신 소셜미디어 데이터에 기존의 계량적 문헌분석방법을 접목시켜 효율적인 분석이 가능한지에 관한 탐색적 연구의 성격을 가진다.

2. 선행 연구

트위터에 관한 초기 연구들은 트윗 행태와 이용자에 관한 기초통계 분석 위주의 연구들이었으나(Java, Song, Finnin, & Tseng, 2007; 정혜란, 지숙영, 이중식, 2010; 황유선, 심홍진, 2010), 최근에는 트윗의 내용을 분석하여 유용한 정보를 발견하는 연구가 증가하고 있다.

우선, 방대한 양의 트윗 데이터를 이용하여 이용자의 의견 및 감정을 분석하는 연구들이 있다. 오피니언 마이닝과 감정분석은 의견의 객관성·주관성 및 감정의 극성을 분류하는 어휘 사전 혹은 학습 데이터 구축을 기반으로 행해져 왔다(Esuli & Sebastiani, 2006; Strapparava, Gliozzo, & Giuliano, 2004). 이러한 기계학습과정(machine learning)을 최소화하기 위한 연구들 중 리뷰의 이용자 별점을 긍정·부정 의견의 기준으로 활용하여 트위터의 특성인 자유로운 형식의 짧은 문장을 고려한 연구도 이루어졌다(송종석, 이수원, 2011). 그 외에 이모티콘을 학습 데이터의 분류 기준으로써 감정분석에 활용한 연구(Go, Bhayani, & Hunag, 2009), 해시태그와 스마일 레이블을 함께 활용한 연구(Davidiv, Oren Tsur, & Rappoport, 2010), 해시태그를 키워드로 활용하여 준 지도학습(semi-supervised learning)을 통해 감정 극성을 분석한 연구(Wang, Wei, Liu, Zhou, & Zhang, 2011)가 있다. 하지만 위의 연구들은 부분적 요소들만 이용함으로써 전체 메시지가 나타내고 있는 내용 및 맥락을 충분히 반영하지 못하는 한계가 있다.

또한 이용자가 원하는 주제를 나타내는 타겟과 쿼리(query)를 기반으로 감정을 분류한 연구로는 Jiang, Yu, Zhou, Liu, Zhao(2011)의 연구가 있다. 이 연구에서는 그래프 기반의 최적화(optimization) 과정을 통해 정확도를 증가시켰으며, 이를 통해 토픽을 대상으로 한 그래프 기반 분석이 가지는 유효성을 확인하였다. Asur와 Huberman(2010)은 영화 박스오피스 성적을 예측하기 위해 트윗을 대상으로 오피니언 마이닝을 실행한 결과, 시간의 흐름에 따른 트윗 비율을 고려한 경우 약 90%를 넘는 결정계수(R^2)값을 보였다.

트윗의 내용을 분석하는 연구의 또 다른 흐름으로는 기하급수적으로 쏟아지는 트윗의 토픽을 파악하고 분류하는 연구가 있다. 이용자의 프로필, 작성된 트윗 메시지 등을 기반으로 기계학습을 통해 이용자들의 정치적 성향, 관심사를 자동 분류하는 연구(Pennacchiotti & Popescu, 2011)부터 비계층적 군집화 기법인 K-means, 계층적 군집화 기법인 HAC, 그래프 군집화 기법인 MCL을 적용하여 관심사가 유사한 트위터 사용자들을 군집화한 연구(김성훈, 최돈정, 김재광, 정혜욱, 이지형, 2011), 트위터와 위키피디아의 토픽 검색 결과를 결합하여 학습 세트를 구축한 후 준 지도 클러스터링 기법을 적용한 연구(Chen, Shipper, & Khan, 2010), 다양한 시간 스케일에서 주제를 추출할

수 있는 NMF(Non-negative Matrix Factorization) 클러스터링 기법을 적용하여 트위터의 트렌드를 분석한 연구(하용호, 임성원, 김용혁, 2012)까지 다양한 클러스터링 기법을 적용하여 토픽과 트렌드를 파악하고자 하였다. 또한 Sakaki, Toriumi, Matsuo(2011)는 시간의 흐름에 따른 트위터의 토픽 변화 양상을 살펴보기 위해 단어의 출현빈도를 사용하였다. 트윗에 출현한 동일본 대지진 사건에 관한 단어들의 빈도수 변화와 이용자들의 네트워크를 통해 긴급 상황에서 일어나는 트위터의 토픽 변화를 살펴보았다. 위의 연구들을 통해 트위터 상의 토픽이 가지는 예측성과 연구 가치를 확인할 수 있었다.

이용자의 의견 및 감정뿐만 아니라 트위터 상의 특정 상품 관련 토픽이 오프라인과 연계되어 어떠한 특성을 나타내는지 동시출현단어들을 기반으로 생성된 네트워크 분석을 통해 살펴보았다.

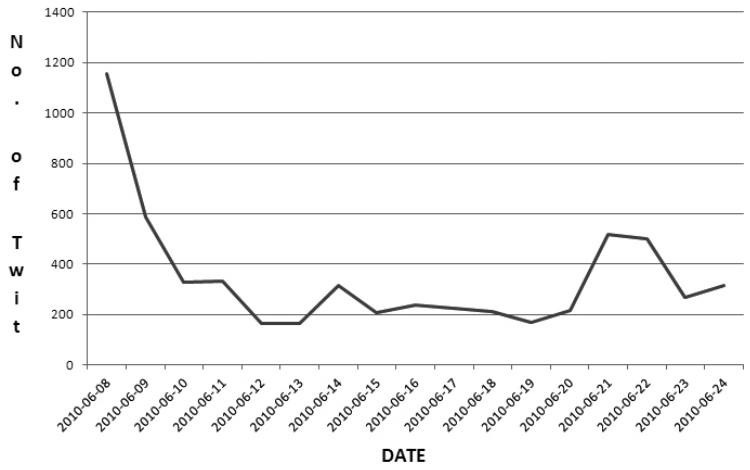
3. 실험 설계

3.1 데이터 수집

국내 트위터 이용자 11,379명을 대상으로 2010년 5월 16일부터 2010년 8월 16일까지 작성된 총 2,647,727개의 트윗(이원태, 차미영, 양해륜, 2011)을 데이터베이스화 하였으며 이 중 특정 제품에 대한 시간적 변화를 파악하기 위해 질의어(query term)를 선정하였다. 따라서 프레젠테이션을 통해 '아이폰4'가 발표되었던 2010년 6월 8일부터 미국의 출시일인 2010년 6월 24일 사이의 트윗 중 상품명 '아이폰'을 포함하는 트윗 총 10,866개를 분석대상으로 삼았다. 질의어 '아이폰'을 포함하는 트윗의 날짜별 추이는 〈그림 1〉과 같다.

3.2 전처리(preprocessing)

수집된 데이터를 데이터베이스화하여 토큰화

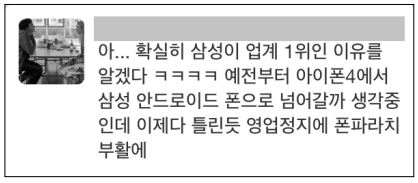


〈그림 1〉 질의어 '아이폰'을 포함하는 트윗의 날짜별 개수

(Tokenization) 과정을 거친 후 Lucene Korean Analyzer 3.5 형태소 분석기를 사용하여 토픽 키워드로서 쓰일 180,626개의 명사를 추출하였다. 그 후, 수집된 키워드에서 멘션(mention)으로 인해 발생되는 영문 아이디, 링크 URL, '어제', '모두' 등과 같이 토픽으로서 의미를 가지지 않는 불용어와 '어흑', 'ㅋㅋㅋ' 등과 같이 인터넷 용어를 합쳐 총 584개의 키워드를 제거하였다. 또한 예를 들어 'iphone4', '아이폰포' 등의 용어는 모두 '아이폰4'로, '흰둥이', '화이트'는 '흰색'으로 바꾸는 키워드 통합을 실행하였다. 그리하여 한 트윗당 평균 10개 미만의 키워드를 포함하는 총 91,427개의 키워드로 정리되었다. 마지막으로 이용자들에게 많이 회자되는 이슈가 토픽으로서 의미를 가진다는 것을 전제로 단순출현빈도가 20회 이하인 키워드는 분석에서 제외하였다.

3.3 동시출현(co-occurrence) 키워드 생성

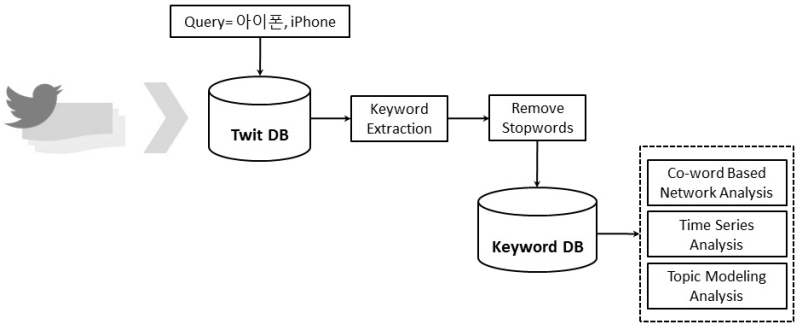
전처리 과정 이후 한 트윗에서 동시 출현한 키워드를 대상으로 동시출현 빈도를 산출하였다. 〈그림 2〉와 같이 '아이폰'을 포함하는 트윗에서 키워드는 '삼성', '안드로이드폰', '업계', '영업정지' 등이 있다. 이 경우 '아이폰4'와 '삼성'은 동시출현빈도가 1이 된다. 이 과정에서도 역시 20회 이하의 동시출현빈도를 가진 페어는 제거하였다.



〈그림 2〉 트위터 예시

3.4 트위터 토픽 키워드 분석

본 연구의 전체적인 연구 설계는 〈그림 3〉과 같다. 전처리 과정을 통하여 특정 기간 동안의 트윗에서 동시출현한 키워드 페어를 생성하였으며, '아이폰' 상품명에 해당하는 트윗을 다차원적으로 분석하기 위하여 총 세 가지의 분석 기법을 적용하였다. 우선 동시출현 키워드 기반의 네트워크 분석과 연결정도 중심성을 고려하여 선정한 키워드의 출현빈도 기반 시계열 분석



〈그림 3〉 연구 설계

을 시도하였으며, 마지막으로 토픽 모델링 기법 중 확률 기반 문헌 분석 모델인 LDA(Latent Dirichlet Allocation) 모델을 적용하였다. 각 분석 기법에 대한 추가적인 설명은 다음과 같다.

20회 이상의 동시출현빈도를 가진 키워드 페어를 대상으로 네트워크 분석을 시도하였으며, 데이터 시각화 도구인 'Gephi'를 이용하여 시각화하였다. 키워드 페어의 동시출현빈도가 높을수록 내용적 연관성이 높다는 가정을 바탕으로 노드(node)는 토픽 키워드를, 에지(edge)는 동시출현빈도를 나타내며 방향성이 없는 날짜별 네트워크를 그린 후 특징을 보이는 네트워크를 중심으로 분석을 실행하였다.

또한 출현빈도 기반의 시계열 그래프 분석을 위해 가장 큰 규모의 네트워크인 2010년 6월 8일의 토픽 네트워크에서 동시출현빈도에 영향을 받는 연결정도 중심성(degree centrality)을 고려하여 상위 20개의 키워드를 선정한 후, 해당 노드에 해당하는 페이지랭크(PageRank) 값을 산출하였다(연결정도 중심성과 페이지랭크 값 산출은 시각화 도구인 'Gephi'를 활용하였다). 그 후, 이에 대한 시계열 분석을 시도하여 키워드의 변화 양상을 살펴보고자 했다. 또한 오프라인에서도 비교대상인 경쟁제품 키워드의 하루 단위 출현빈도를 시계열 그래프로 표현하여 각각의 특징을 비교 분석하였다.

마지막으로 토픽 모델링 기법을 적용하였다. Blei, Ng, Jordan, Lafferty(2003)에 의해 제안된 토픽 모델링 기법 중 하나인 LDA 모델은 각 문헌집단은 토픽의 확률적인 혼합체이며, 각 토픽은 단어의 분포로 이루어져 있다는 전제 하에 확률을 기반으로 토픽을 추출하는 기법이다. 본 연구에서는 자바 기반의 오픈소스 툴킷인 MALLET을 이용하여 LDA 모델을 실행하였으며 10개로 나뉘어 산출된 토픽과 각 토픽을 구성하는 단어들을 확인하였다.

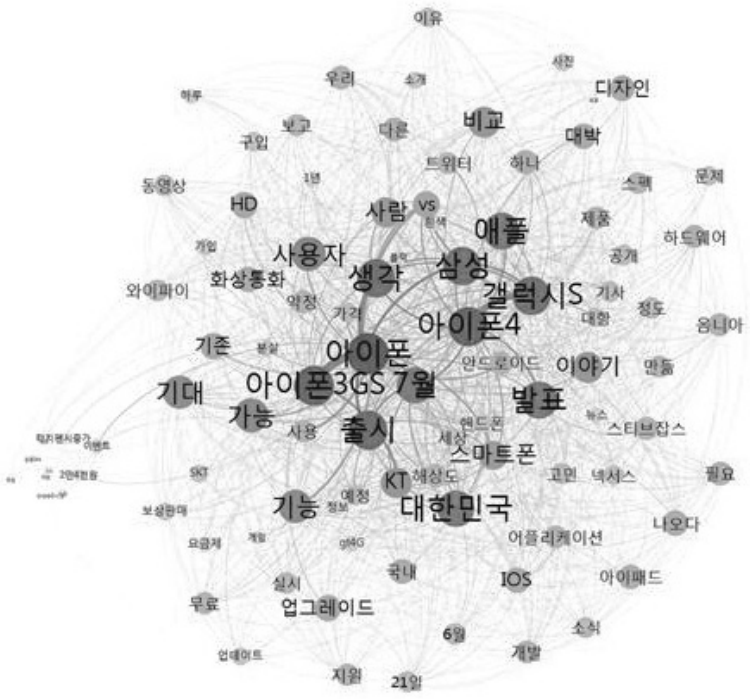
이러한 동시출현단어들을 이용한 네트워크 분석 결과의 타당성을 검증하기 위해 시계열 그래프 분석, 토픽 모델링 기법을 통해 나타난 결과들과 비교분석하였다.

4. 실험 결과 분석

4.1 토픽 기반 동시출현 단어 네트워크 분석

2010년 6월 8일부터 2010년 6월 24일까지 총 17일의 기간 동안 '아이폰'을 포함하는 트윗 키워드들의 관계를 나타내는 네트워크를 하루 단위로 만들어 분석하였다. 포털 검색엔진 '네이버'에서 질의어 '아이폰'을 통해 검색된 각 날짜 당일의 언론 기사와 트윗 데이터를 비교하여 네트워크를 분석하였으며, 트위터 내에서 일어나는 토픽 전파의 특성을 살펴보았다.

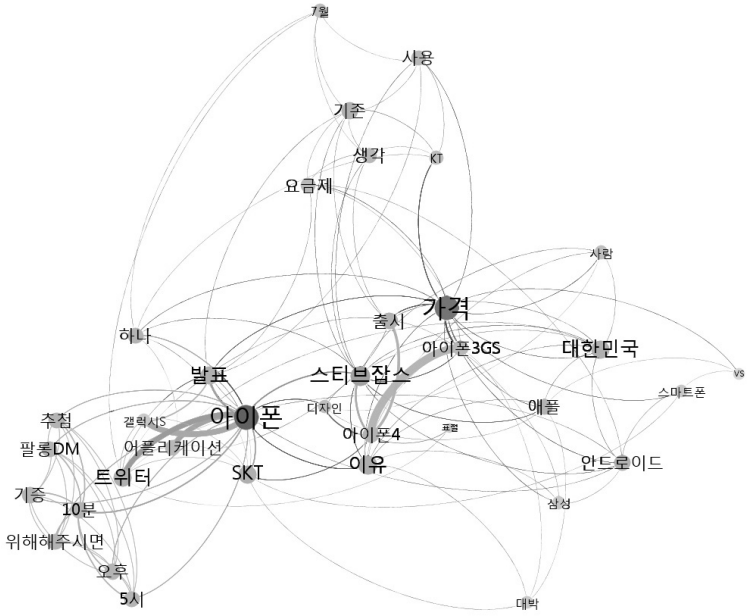
〈그림 4〉는 스티브잡스의 '아이폰4' 출시 발표 프레젠테이션이 열린 2010년 6월 8일의 전체 네트워크 구조이다. 네트워크의 중심부를 차지한 크기가 큰 노드를 살펴보면 '애플', '발표'와 함께 '아이폰', '아이폰4', '아이폰3GS'의 빈도수가 높게 나타난 것을 확인할 수 있다. 또한 '7월'은 당시 '대한민국'에서 'KT' 통신사의 '아이폰4' 출시가 7월로 예상되는 기사가 발표되어 출현빈도가 높게 나타났다. 한편 미국에서 '아이폰4' 출시 발표를 한지 8시간 후인 국내 오후시각에는 '아이폰4'와 경쟁 제품으로 꼽히는 '삼성'의



〈그림 4〉 2010년 6월 8일의 토픽 네트워크

'갤럭시S'의 출시가 발표되어 대한민국 언론들이 두 제품을 비교하는 기사들을 쏟아내었다. 이는 소셜미디어상에서도 이슈화 되어 애플과 삼성, 아이폰과 갤럭시를 비교하는 트윗들이 자주 출현하고 기사도 링크되어 크기가 큰 노드로서 자리하게 되었다.

'아이폰'을 포함하는 트윗의 수가 〈그림 1〉에서처럼 6월 8일에 집중적으로 높게 나타났기 때문에 6월 9일 이후의 네트워크들은 8일의 네트워크보다 작고 분할된 모습을 보이게 된다. 프레젠테이션의 다음날인 6월 9일의 네트워크인 〈그림 5〉에서는 '가격' 노드의 크기가 커진 것을 볼 수 있는데, 9일의 언론 기사에서는 KT가 기존 아이폰3GS를 반값에 가격 인하 판매 소식을 발표한 것과 매치된다. 또한 '아이폰4'와 '아이폰3GS'간의 링크가 가장 굵게 나타난 것을 확인할 수 있는데, 이는 새롭게 출시된 '아이폰4'와 기존의 '아이폰3GS'를 직접적으로 비교하는 트윗들이 많이 출현하면서 이 두 단어의 동시출현 빈도수가 가장 높게 나타났다. 〈그림 6〉은 6월 10일의 네트워크 구조로 미국에서 기존제품을 '아이폰4'로 교환해주는 '보상판매' 계획을 발표한 날짜로 확인되었다. 한편 국내 'KT'에서는 "아이폰 보상판매 계획이 없다."라는 점을 밝히면서 '아이폰4', '애플', '보상판매' 단어가 트윗상에서 이슈화되었다. 11, 12, 13일은 10일의 네트워크와 구별되는 특징적 변화가 일어나지 않았다. 'KT'에서 기존 '아이폰'의 '약정'승계 검토에 대한 기사와 '아이폰'전용의 평생 '요금제' 출시를 발표한 6월 14일의 토픽 네트워크를 나타

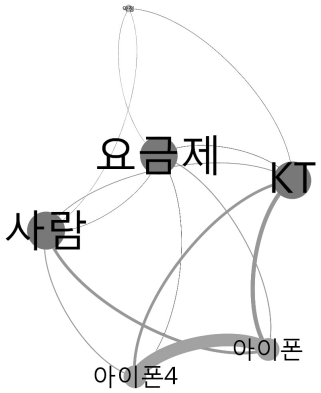


〈그림 5〉 6월 9일의 토픽 네트워크

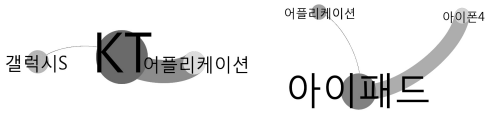
〈그림 6〉 6월 10일의 토픽 네트워크

낸 〈그림 7〉에서도 '요금제' 및 'KT'의 단어 출현빈도가 높게 나타났으며, 기존에 존재하지 않았던 '약정'노드 또한 다른 주요 키워드들과 동시에 출현하였다. 〈그림 8〉의 좌측에 위치한 6월 16일은 KT에서 '아이패드'를 독자적으로 들여오면서 '어플리케이션' 개발 계획을 발표한 날로 2년 전의 아이패드 출시와 더불어 태블릿 PC의 성공여부에 대한 관심이 소셜미디어 상에서 많이 나타남으로 인해 많이 출현하였다. 〈그림 8〉의 우측 네트워크인 6월 17일에는 전날인 16일에 이어서 KT에서 독자적으로 개발한 클라우드 컴퓨터 형식의 웹하드인 'Ucloud' 어플리케이션의 오픈 발표가 있었다. 이 당시에 국내 앱스토어에 등록된 어플리케이션의 숫자는 약 6천개

에 불과하였기 때문에, 트위터 상에서 'KT'와 '어플리케이션'에 대한 관심이 급증한 것으로 확인할 수 있다.

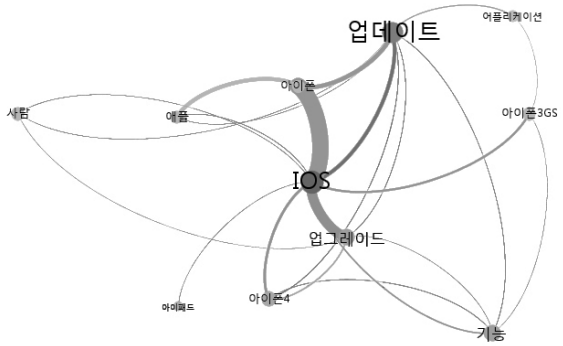


〈그림 7〉 6월 14일의 토픽 네트워크



〈그림 8〉 6월 16일(좌), 17일(우)의 토픽 네트워크

〈그림 9〉는 애플에서 자사 아이폰 사용자들을 대상으로 한 운영체제인 'IOS 4'의 업그레이드를 실시한다고 밝힌 6월 21일의 토픽 네트워크이다. 따라서 'IOS' 단어를 중심으로 '업그레이드', '업데이트' 등의 단어가 높은 빈도로 출현하였으며, 업그레이드 후 'IOS4'의 신규 및 개선 '기능'에 대한 정보를 트위터 상에서 함께 공유했던 것으로 나타났다. 〈그림 10〉은 6월 23일의 네트워크 구조로 21, 22일의 'IOS4' 업데이트에 맞춰 트위터 버전이 3.0.1로 '업그레이드'되었으며, 〈그림 11〉은 본 연구의 분석대상 기간 중 마지막 날인 6월 24일의 네트워크 구조로서 미국에서 '아이폰4'가 출시되어 판매가 시작된 날이다. 특징적 화제에만 집중되었던 9일~23일과 다르게 아이폰과 관련한 다양한 노드가 다시 네트워크를 구성하고 있음을 확인할 수 있다. 이 날은 특히 '문제' 단어의 동시출현빈도가 매우 높게 나타났는데, 이에 대해 '네이버' 포털 사이트를 통해 검색한 6월 24일의 언론 기사와 당일의 트윗을 비교해 보았다. 먼저 언론 기사 검색 결과, 총 196건의 '아이폰'을 포함한 언론 기사가 나타났고, 기사의 내용을 확인하여 아이폰과 관련된 문제점에 대해 언급한 언론 기사 10건을 확인하였다. 기사에서는 '수급문제', '액정결함', '통화품질', 'IOS4.0 업그레이드 시 발생된 주소록 데이터가 사라졌다거나 G메일 등과 싱크 문제가 발생했다는 등의 사고 사례', '아이폰4의 사전 주문 폭주로 자사와 AT&T 웹사이트의 서버가 다운되는 문제' 등을 언급하였다. 한편 6월 24일의 데이터 셋인 총 317건의 트윗 중에서 아이폰의 기능 및 품질 문제 등을 다룬 기사를 링크한 트윗 17건과, 아이폰 이용시 발생한 문제점에 대해 언급한 트윗 53건을 포함하여 총 70건의 트윗을 확인하였다. 이와 같이 아이폰의 문제를 언급한 언론 기사는 5.1%인 반면 트윗 데이터는 22%로 4배 이상 늘어난 점을 확인할 수 있다. 이를 통해 언론사의 언론 기사 자체에서는 부정적인 문제에 큰 비중을 두지 않았음에도 불구하고 트위터에서는 이러한 크고 작은 '문제'들을 놓고 부정적인 견해를 드러내는 비율이 높으며, 링크 및 RT를 통해 정보를 확산하는 경향을 보이는 점을 발견할 수 있었다. 이러한 트위터의 선별적 정보 채택 및 전파가 전통적 언론이 제시하는 주요 뉴스뿐만 아니라 언론에서 강조하지 않은 예외적인 흥미성 이슈들에 대해



〈그림 9〉 6월 21일의 토픽 네트워크

서도 이루어진다는 것을 확인할 수 있으며 이는 선행연구(김은미, 이주현, 2011)에서도 나타났다. 부정적 의견은 긍정의견에 비해 쉽게 상쇄되지 못한다는 연구(전선규, 1996)를 바탕으로 부정적 정보가 보다 많이 회자되고 전파되는 트위터 상의 토픽에 대해 신속한 대응이 필요함을 인식할 수 있다.

〈그림 10〉 6월 23일의 토픽 네트워크

〈그림 11〉 6월 24일의 토픽 네트워크

17일간의 토픽 네트워크와 동일 질의어를 입력하여 검색된 언론 기사를 비교한 결과, 트위터는 사안이 공개되는 당일의 이슈에 상당히 예민하게 반응하고 있음을 확인할 수 있었다.

4.2 키워드 빈도 분석 및 네트워크 중심성 분석

앞 절에서 제시한 네트워크 분석을 바탕으로 네트워크의 중심성 척도를 계산하였다. 〈표 1〉은 〈그림 4〉의 2010년 6월 8일 토픽 네트워크의 연결정도 중심성 값을 기준으로 선정한 상위 20개의 키워드이며, Brin과 Page(1998)의 페이지랭크 알고리즘을 사용하여 페이지랭크 값 또한 나타냈다. 각 키워드에 대한 연결정도 중심성과 페이지랭크 값은 일반적으로 비례하는 결과를 보였지만 예외적으로 연결정도 중심성 값은 낮은 반면 페이지랭크 값에서 높은 값을 가지는 3개의 키워드를 발견했다(〈표 2〉 참조). '이벤트', '2만4천원', '터치펜시중가' 키워드를 포함하는 트윗을 살펴본 결과, 이벤트를 홍보하는 트윗으로서 약 20회 RT된 것으로 확인되었다. 이러한 사례를 통해 일반적으로 웹 환경에서 중

요한 웹 페이지를 인용 또는 링크한 웹 페이지에 가중치를 부여하는 페이지랭크 값을 통해 트위터 상에서 영향력 있는 트윗이 리트윗(retweet: RT)되는 현상을 살펴볼 수 있었다. 이와 같이 RT가 많이 이루어진 토픽의 식별을 통해 특정 상품에 대한 폭발적인 화제성, 무분별한 정보전파 등과 관련된 커뮤니케이션 파악 및 대처가 가능하다.

〈표 1〉 노드값 상위 20개 키워드

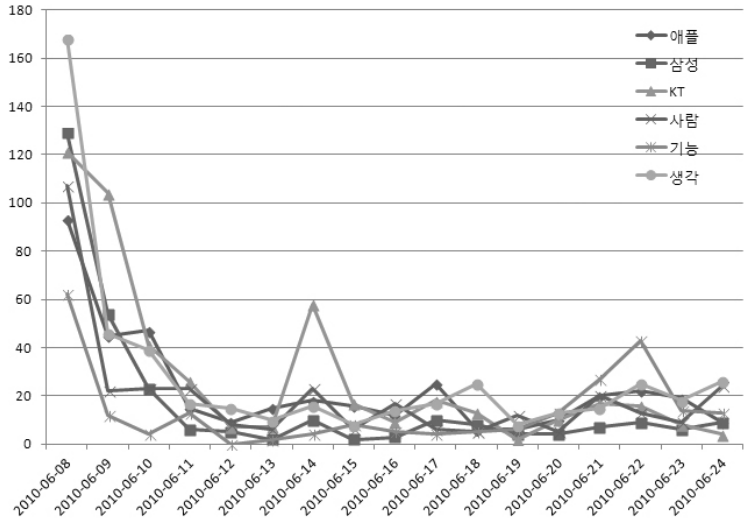
Label	Degree	PageRank
아이폰4	67	0.0218
아이폰3GS	67	0.0194
갤럭시S	67	0.0190
삼성	64	0.0186
출시	64	0.0181
생각	63	0.0183
애플	63	0.0179
발표	60	0.0172
7월	60	0.0172
가능	59	0.0159
기능	53	0.0150
기대	52	0.0148
비교	51	0.0144
이야기	48	0.0137
스마트폰	47	0.0134
사람	46	0.0136
대한민국	45	0.0131
KT	43	0.0126
사용자	41	0.0121
기존	41	0.0119

〈표 2〉 페이지랭크 값이 높게 나타난 3개 키워드

노드명	연결정도 중심성	순위	페이지랭크	순위
이벤트	15	87	0.011674	23
2만4천원	10	90	0.01015	39
터치펜시중가	10	91	0.01015	40

또한 핵심 키워드의 빈도 변화를 시간의 흐름에 따라 분석함으로써 해당 토픽의 변화 양상을 살펴보았다. 연결정도 중심성 값의 상위 20개 키워드에 대한 시계열 그래프를 모두 그려본 결과, 오르내리는 변동을 보여주며 시간의 흐름에 따른 토픽 분석이 유의미하다고 판단되는 6개의 키워드를 선정하였다. 선정된 키워드는 '애플', '삼성', 'KT', '사람', '기능', '생각'이며 이에 대한 시계열 그래프는 〈그림 12〉와 같다. 〈그림 12〉는 앞서 기술한 6개의 키워드에 대한 시계열 그래프로써 X축은 2010년 6월 8일부터 2010년 6월 24일까지의 시간을 나타내며, Y축은 키워드의 출현빈도를 나타낸다. 본 연구에서 설정한 기간 중 6월 8일은 미국 아이폰 출시 발표일로 모든 키워드가 6월 8일에 가장 높은 빈도수를 보임을 확인할 수 있다. 특히 '생각' 키워드는 무려 160번이 넘는 출현빈도를 나타냈는데 이는 아이폰에 대한 구체적인 생각들을 공유하는 트윗이 많이 나타남으로 인해 발생한 것으로써 트위터 이용자의 아이폰에 대한 관심을 확인할 수 있는 키워드로 판단되었다. 또한 '생각', '사람' 키워드의 높은 출현빈도를 통해 트위터에는 개개인의 의견과 대상을 지칭할 때 쓰이는 일반적인 단어들이 많이 쓰임을 확인할 수 있었다. 예를 들어 "사람들이 그러던데~", "제 생각엔~", "~하다는 생각"과 같은 트윗의 패턴들을 확인할 수 있었다.

시계열 그래프 상에서 특징을 보이고 있는 키워드에 대해 중점적으로 살펴보면, 우선 6월 8일 이후 토픽의 출현빈도가 일반적으로 하향세를 보이고 있으나 당시 아이폰의 유일한 통신사였던 'KT'가 6월 14일에 두드러지게 나타난 이유는 아이폰 3G 약정 승계 검토와 아이폰 전

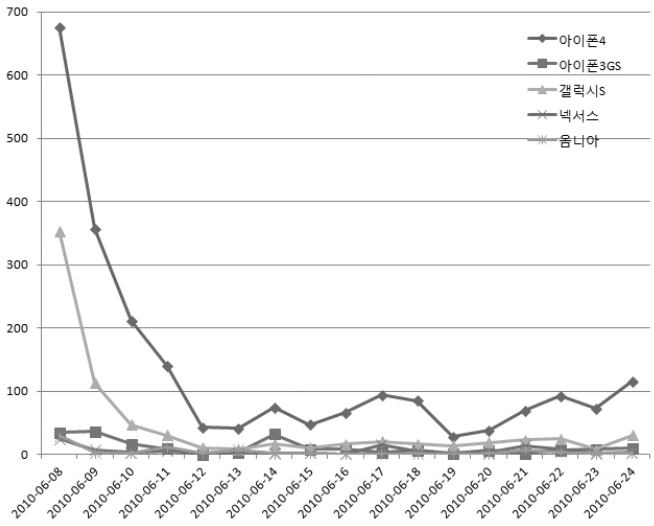


〈그림 12〉 6개 키워드 시계열 그래프

용 '평생 요금제'를 출시하기로 공식 발표하여 트위터 상에서도 화두가 되었던 것으로 확인되었다. 또한 '기능'은 6월 22일 한국에서 아이폰 'IOS4'로 업데이트를 발표하여 6월 21일부터 많은 아이폰 이용자들이 IOS 업그레이드와 관련하여 기능에 대한 언급횟수가 높았다. 이와 같이 14일, 21일의 토픽 네트워크 구조와 일관적인 맥락을 보이며, '기능' 키워드의 경우에는 단순빈도가 아닌 질의어와의 동시출현빈도를 보여주는 네트워크 분석에서 하루 더 빠르게 토픽 키워드로서 언급된 점을 볼 수 있었다.

〈그림 13〉은 스마트폰의 경쟁사 제품 키워드를 대상으로 한 시계열 그래프이다. 6월 8일 발표되었던 애플사의 '아이폰4'와 '아이폰3GS', 삼성사의 '갤럭시S', 그리고 추가적으로 구글의 '넥서스', 삼성사의 '옴니아' 스마트폰을 비교하였다. 세 키워드 중 '아이폰4'와 '갤럭시S'는 6월 8일에 가장 높은 출현빈도를 보인 후 급격하게 출현빈도가 적어지는 유사한 패턴을 보이며, '아이폰3GS'는 상대적으로 출현빈도 자체뿐만 아니라 뚜렷한 변화 양상 또한 보이지 않았다. 트윗 데이터를 분석해보았을 때 트윗의 패턴은 1) '아이폰4'와 '아이폰3GS'를 비교하거나 2) 아이폰4'와 '아이폰3GS' 및 '갤럭시S'를 모두 비교하거나 3) '아이폰4'와 '갤럭시S'를 비교한 총 세 가지의 패턴을 확인할 수 있었는데 이 중 '아이폰3GS'보다는 새롭게 출시된 '아이폰4'와 '갤럭시S'를 직접적으로 비교하여 언급하는 3번 패턴의 트윗이 상당수를 차지하였다. 한편, '넥서스', '옴니아'의 경우 첫째 날인 6월 8일에 24회, 29회 빈도로 가장 많이 출현하긴 했지만 대부분 10회 이하의 빈도를 나타내어 '아이폰'과 '갤럭시S'의 경쟁구도에 큰 역할을 하지 못하고 있다는 점을 확인할 수 있었다.

지금까지 시계열 그래프를 통해 2010년 6월 8일부터 2010년 6월 24일까지의 총 17일 동안 출현빈도가 높게 나타난 키워드들을 대상으로 날짜별 변화 양상을 파악해보았다. 이를 통해



〈그림 13〉 스마트폰 제품 키워드 그래프

동시출현빈도 기반의 토픽 네트워크 분석과 단순빈도 기반의 시계열 그래프가 일관성 있는 흐름을 보여주며, 특히 네트워크 분석에서는 키워드들 간의 관계를 보여줌으로써 키워드들이 나타내는 이슈의 주제를 보다 손쉽게 파악할 수 있음을 알 수 있었다.

4.3 토픽 모델링 결과

네트워크 분석과 키워드 빈도분석에서 나타난 토픽 변화 양상을 바탕으로, 트윗에서 나타난 키워드들이 어떠한 주제그룹을 형성하는 지를 파악하기 위해 토픽 모델링(Topic modeling)을 실행하였다. 하나의 트윗을 LDA 모델에서의 문헌으로 간주하고 실험 대상 트윗을 입력하여 10개의 토픽을 추출하였다.

〈표 3〉에서 보듯이 2번 토픽은 가장 큰 규모를 가지는 6월 8일의 토픽 네트워크에서 아이폰과 함께 크기가 큰 노드에 속하는 아이폰의 경쟁 제품, '갤럭시S', '삼성'과 '안드로이드', '옴니아' 등 각각의 스마트폰 제품을 비교하는 트윗이 하나의 토픽을 형성하고 있다. 또한 7번 토픽은 'KT', '요금제', 'SKT' 등의 단어로서 구성되어 6월 14일의 토픽 네트워크와 매치되며 8번 토픽은 'IOS', '업데이트', '기능', '어플리케이션'의 단어들로 구성되어 IOS4의 업그레이드 실시를 발표한 직후인 6월 21일의 토픽 네트워크와 부합되는 것을 볼 수 있다.

이를 통해 앞 절의 네트워크 분석이 나타내는 당일의 이슈를 토픽 모델링에서도 하나의 토픽으로 제시함을 볼 수 있었다. 이후, 해당 날짜별 토픽 모델링을 통해 분석을 실행한다면 특정 이슈에 대한 토픽의 세부 주제 변화를 파악하는 데에도 유용할 것으로 보인다. 이와 같이 토픽 분석에 관해 상호보완적 관계를 가지는 네트워크 분석과 토픽 모델링 분석을 결합한 연구가 추가적으로 이루어진다면 보다 상세한 토픽의 다이내믹스 파악이 가능할 것이다.

〈표 3〉 토픽 모델링 결과

토픽	토픽 구성 단어
1	트위터 유저 사람 부탁 시작 보고 친구
2	안드로이드 애플 스마트폰 갤럭시S 삼성 생각 옴니아 기사 대한민국
3	케이스 생각 트위터 흰색 하나 사람 핸드폰
4	생각 아이패드 사용 기능 터치 GS 사람 스마트폰 아이팟
5	트위터 사진 어플리케이션 동영상 스티브잡스 이벤트 게임 회사
6	어플리케이션 무료 게임 트위터 사용 추천 개발 다운
7	KT GS 요금제 사용 SKT 사용자 가입
8	어플리케이션 업데이트 방법 가능 IOS 사용 기능 아이튠즈 접속
9	월 출시 일 KT 분 이벤트 시 GS 트위터
10	배터리 충전 넌 아이팟 터치 하나 핸드폰 고민 생각

5. 결 론

최신 소셜미디어 데이터에 기존의 계량적 문헌분석방법인 동시출현단어분석을 접목한 본 연구에서는 사람들의 관심사 변화를 빠르게 반영하는 트위터 데이터 중 특정 상품명을 포함하는 트윗을 대상으로 토픽 변화를 추적하였다. 트위터의 메시지(트윗)와 작성날짜 정보만을 사용한 간결한 정보로 신속하게 관련 토픽을 추출하고 토픽 간의 관련성을 측정하기 위해 동시출현단어분석을 이용하였다. 트윗에서 추출한 키워드들 간의 동시출현빈도 기반 네트워크를 만들어 하루 단위로 변하는 네트워크 구조 분석을 실행하였다. 그 후, 연결정도 중심성을 고려한 상위 20개의 키워드 중 토픽의 변화 양상을 파악할 수 있는 키워드를 선정하여 출현빈도 기반의 시계열 그래프를 그려 특정 기간 동안에 나타난 각 토픽의 흐름을 비교 분석하였다. 더불어 문헌 당 단어가 출현하는 확률을 기반으로 토픽을 추출하는 LDA기반의 토픽 모델링 기법을 적용하여 트윗에서 나타난 키워드들이 주제적으로 어떠한 토픽 그룹을 형성하는지 확인하였다. 추가적인 시계열 그래프와 토픽 모델링 결과 분석이 동시출현빈도 기반의 토픽 네트워크 분석과 일관적인 맥락을 나타낸다는 점에서 본 연구의 네트워크 분석 결과의 효용성을 입증할 수 있었다. 이뿐만 아니라 단순빈도 측정을 통한 키워드 추출 수준의 시계열 그래프보다 네트워크 분석이 각 키워드들 간의 관계를 보여줌으로써 관련 키워드 파악, 토픽의 전체적 양상을 손쉽게 파악할 수 있다는 장점 또한 살펴볼 수 있었다. 또한 토픽 모델링 분석과의 상호보완적 관계를 통해 보다 깊이 있는 토픽 분석에 대한 가능성을 확인하였다. 본 연구에서 사용된 분석 대상은 특정 상품으로 범위를 한정하여 토픽 변화를 살펴보았으나, 토픽 모델링을 통해 범위를 한정하지 않고 트위터 상의 전체 토픽의 변화 양상을 살펴보는 후속 연구가 가능할 것이다. 기업, 이용자 그룹 등과 같은 추가적인 파라미터를 적용하여 토픽을 추출하는 DMR (Dirichlet-multinomial regression, Mimno & MacCallum, 2008) 기법을 활용한다면 추가적 파라미터를 조건으로 생성된 토픽의 변화 양상을 비교 분석할 수 있을 것이다. 또한 동시출현

단어분석을 기반으로 구축한 네트워크에 그래프 마이닝(Graph mining)기법들을 적용하여 토픽들의 연관성을 추적하는 연구도 수행할 예정이다.

본 연구를 통해 관찰된 트위터 상의 커뮤니케이션 특징은 다음과 같다. 트위터 상에서 나타나는 토픽 변화와 언론 기사 검색 결과를 비교 분석한 결과, 트위터는 언론 뉴스에 즉각적으로 반응하였으며 언론이 다루는 비중과 무관하게 상품에 관한 부정적 이슈를 빠르게 확산시키는 특성을 보였다.

본 연구에서 활용한 신속하고 직관적인 네트워크 분석방법뿐만 아니라 이와 같은 특성을 기반으로 기업은 트위터를 이용해 대중이 만들고 전파하는 부정적 의견을 파악하고 이에 대한 즉각적인 의사결정 및 대응을 위한 마케팅 도구로 활용할 수 있다.

참 고 문 헌

김성훈, 최돈정, 김재광, 정혜욱, 이지형 (2011). 트위터 게시물을 이용한 공통 관심사를 지닌 사용자 그룹 발견. 한국지능시스템학회 학술발표 논문집, 21(2), 129-131.

김은미, 이주현 (2011). 뉴스미디어로서의 트위터. 한국언론학보, 55(6), 152-180.

송종석, 이수원 (2011). 상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축. 정보과학회 논문지: 소프트웨어 및 응용, 38(3), 157-168.

이원태, 차미영, 양해륜 (2011). 소셜미디어 유력자의 네트워크 특성: 한국의 트위터를 중심으로. 언론정보연구, 48(2), 44-79.

전선규 (1996). 불만족한 소비자의 구매 후 행동. 마케팅, 30(10), 22-26.

정혜란, 지숙영, 이중식 (2010). 국내 트위터 유저분석을 위한 예비연구: “익스트림 헤비 유저”의 트위터 로그를 중심으로. 한국 HCI학회 논문지, 5(1), 37-43.

최돈정, 이성우, 김재광, 이지형 (2011). 마이크로 블로그를 통한 그래프 기반의 토픽 추출에 관한 연구. 한국지능시스템학회 논문지, 21(5), 564-568.

하용호, 임성원, 김용혁 (2012). 내용기반 트윗 클러스터링을 통한 트렌드 분석. 한국정보과학회 학술발표논문집, 39(2B), 210-212.

황유선, 심홍진 (2010). 트위터에서의 의견 지도력과 트위터 이용패턴: 이용동기, 트윗 이용패턴, 그리고 유형별 사례분석. 한국방송학보, 24(6), 365-404.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. Retrieved from http://arxiv.org/abs/1003.5699

Bermingham, A., & Smeaton, A. F. (2011). On using twitter to monitor political sentiment and

predict election results. In: Sentiment Analysis Where AI Meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP). Retrieved from http://doras.dcu.ie/16670/1/saaip2011.pdf

Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022. Retrieved from http://jmlr.csail.mit.edu/papers/v3/blei03a.html

Chen, Q., Shipper, T., & Khan, L. (2010). Tweets mining using WIKIPEDIA and impurity cluster measurement. IEEE ISI 2010, 141-143. http://dx.doi.org/10.1109/ISI.2010.5484758

Davidiv, D., Oren Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 241-249.

Esuli, A., & Sebastiani, F. (2006). Determining term subjectivity and term orientation for opinion mining. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), 193-200. Retrieved from http://acl.ldc.upenn.edu/eacl2006/main/papers/13_1_esulisebastiani_192.pdf

Go, A., Bhayani, R., & Hunag, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford University.

Java, A., Song, X., Finnin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07), 56-65.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent twitter sentiment classification. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 151-160.

Mimno, D., & MacCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI '08). Retrieved from http://people.cs.umass.edu/~mccallum/papers/dmr-uai.pdf

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. Proceedings of International AAAI Conference on Weblogs and Social Media, 122-129.

Pennacchiotti, M., & Popescu, A. M. (2011). A machine learning approach to twitter user classification. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 281-288. Retrieved from

https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2886

Sakaki, T., Toriumi, F., & Matsuo, Y. (2011). Tweet trend analysis in an emergency situation. Proceedings of the Special Workshop on Internet and Disasters (SWID '11), Article No. 3. http://dx.doi.org/10.1145/2079360.2079363

Strapparava, C., Gliozzo, A., & Giuliano, C. (2004). Pattern abstraction and term similarity for word sense disambiguation: IRST at Senseval-3. Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3), 229-234.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 178-185.

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011). Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11), 1031-1040.

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Choi, Don-Jung, Lee, Sung-Woo, Kim, Jae-Kwang, & Lee, Jee-Hyong (2011). A study on graph-based topic extraction from microblogs. Journal of Korean Institute of Intelligent Systems, 21(5), 564-568.

Ha, Yong-Ho, Lim, Seong Won, & Kim, Yong-Hyuk (2012). Trend analysis through content-based tweet clustering. Proceedings of the Korean Information Science Society Conference, 39(2B), 210-212.

Hwang, Yoo Sun, & Shim, Hong-Jin (2010). Opinion leadership on twitter and twitter use : Motivations and patterns of twitter use and case study of opinion leaders on twitter. Korean Journal of Broadcasting, 24(6), 365-404.

Jun, Sun Kyu (1996). Postpurchase behavior of discontented consumers. Marketing, 30(10), 22-26.

Jung, Hye Lan, Ji, Soo Kyoung, & Lee, Joong Seek (2010). Preliminary research for Korean twitter user analysis focusing on extreme heavy users twitter log. Journal of the HCI Society of Korea, 5(1), 37-43.

Kim, Eun Mee, & Lee, Ju Hyun (2011). The diffusion of news through twitter and the emerging media ecosystem. Korean Journal of Journalism & Communication Studies, 55(6), 152-180.

Kim, Sung-Hun, Choi, Don Jung, Kim, Jae Kwang, Jung, Hye-Wuk, & Lee, Jee-Hyong (2011).

Discovering twitter user group with common interests by tweets. Proceedings of the Korea Fuzzy Logic and Intelligent Systems Society Conference, 21(2), 129-131.

Lee, Won-Tae, Cha, Mee Young, & Yang, Hae Ryun (2011). Network properties of social media influentials: Focusing on the Korean twitter community. Journal of Communication Research, 48(2), 44-79.

Song, Jong Seok, & Lee, Soo Won (2011). Automatic construction of positive/negative feature-predicate dictionary for polarity classification of product reviews. Journal of KIISS: Software and Applications, 38(3), 157-168.