ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org

J Intell Inform Syst 2015 March: 21(1): 143~159 http://dx.doi.org/10.13088/jiis.2015.21.1.143

트윗 데이터를 활용한 IT 트렌드 분석

이진백
SAS Korea
(jin-baek.yi@sas.com)

이충권
계명대학교 경영정보학과
(cklee@kmu.ac.kr)

차경진
강원대학교 글로벌비즈니스학과
(kjcha7@kangwon.ac.kr)

불확실한 환경변화에 대처하고 장기적 전략수립을 위해 기업에게 있어서 IT 트렌드에 대한 예측은 오랫동안 중요한 주제였다. IT 트렌드에 대한 예측을 기반으로 새로운 시대에 대한 인식을 하고 예산을 배정하여 빠르게 변화하는 기술의 추세에 대비할 수 있기 때문이다. 해마다 유수의 컨설팅업체들과 조사기관에서 차년도 IT 트렌드에 대해서 발표되고는 있지만, 이러한 예측이 실제로 차년도 비즈니스 현실세계에서 나타났는지에 대한 연구는 거의 없었다. 본 연구는 현존하는 빅데이터 기술을 활용하여 서울지역을 중심으로 지난 8개월동안(2013년 5월1일부터 2013년12월31까지) 정보통신산업진흥원과 한국정보화진흥원에서 2012년 말에 발표한 IT 트렌드 토픽이 언급된 21,589개의 트윗 데이터를 수집하여 분석하였다. 또한 2013년에 나라장터에 올라온 프로젝트들이 IT트렌드 토픽과 관련이 있는지 상관관계분석을 실시하였다. 연구결과, 빅데이터, 클라우드, HTML5, 스마트홈, 테블릿PC, UI/UX와 같은 IT토픽은 시간이 지날수록 매우 빈번하게 언급되어 졌으며, 이 같은 토픽들은 2013년 나라장터 공고 프로젝트 데이터와도 매우 유의한 상관관계를 가지고 있는 것을 확인할 수 있었다. 이는 전년도(2012년)에 예측한 트렌드들이 차년도(2013년)에 실제로 트위터와 한국정부의 공공조달사업에 반영되어 나타나고 있는 것을 의미한다. 본 연구는 최신 빅데이터툴을 사용하여, 유수기관의 IT트렌드 예측이 실제로 트위터와 같은 소셜미디에서 생성되는 트윗데이터에서 얼마나 언급되어 나타나는지 추적했다는 점에서 중요한 의의가 있고, 이를 통해 트위터가 사회적 트랜드의 변화를 효율적으로 추적하기에 유용한 도구임을 확인하고자 할 수 있었다.

주제어 : IT 트렌드, 트위터, SNS, IT토픽

논문접수일 : 2015년 3월 1일 논문수정일 : 2015년 3월 18일 게재확정일 : 2015년 3월 19일
투고유형 : 국문급행 교신저자 : 이충권

1. 개요

빅데이터(Big Data)는 기업의 내외부에서 발생하는 다양한 유형의 데이터를 저장, 관리, 분석함으로써 의미 있는 정보를 얻는 기술을 말한다. 다양한 종류의 대규모 데이터에 대한 생성, 수집, 분석, 표현을 그 특징으로 하는 빅데이터 기술의 발전은 다변화된 현대 사회를 더욱 정확하게 예측하여 효율적으로 작동케 하고 개인화된 현대 사회 구성원 마다 맞춤형 정보를 제공, 관리, 분석 가능케 하며 과거에는 불가능했던 기술을 실현시키기도 한다. 특히, 정부나 공공기관들이 자발적으로 공개하는 데이터는 새로운 비즈니스의 기회를 제공하고 있다. 예를 들어, 전기나 수도, 지하철 등과 같은 공공자원에 관한 여러 가지 데이터는 상권의 분석에 활용되고 있다.

특히, 최근에는 하드웨어와 소프트웨어의 발전으로 기하급수적으로 증가하는 데이터에 대한 분석 및 처리가 가능해지면서 개인이 발생시키는 각종 모바일 데이터에 대한 분석이 가능하게 되었다. 즉, 트위터와 같은 모바일 데이터를 분석하여 비즈니스에 활용하려는 시도가 이루어지

게 된 것이다. 예를 들어, 미국 국세청은 탈세 및 사기범죄 예방을 위하여 페이스북이나 트위터 등의 사회관계망 데이터를 분석해 범죄자 집단을 찾아내고 감시할 수 있는 기능을 갖추고 사기범죄 및 탈세 관련사건을 예측에 활용하고 있다. 삼성전자는 외부 사회관계망 분석업체와 협력하여 트위터와 블로그에 올라오는 비정형 텍스트 데이터를 수집하고 분석하여 여론을 파악하여 신제품의 판매가능성과 전략을 수립하거나 다른 사람들에게 파급력이 높은 빅 마우스들을 파악하여 관리하고 있다. 이처럼 정보기술의 발달은 사회관계망에서 발생하는 데이터를 수집하고 분석하여 의미 있는 정보를 얻을 수 있는 가능성을 열어 놓았다.

정보시스템의 연구에 있어서 IT 트렌드에 대한 예측은 오랫동안 중요한 주제였다. IT 트렌드에 대한 예측을 기반으로 새로운 시대에 대한 준비를 하고 예산을 배정하여 빠르게 변화하는 기술추세에 대비할 수 있기 때문이다. 학문적으로는 1990년대 초부터 정보시스템 관리에 있어서 중요한 이슈들을 식별하는 연구를 기업측면(Niederman et al., 1991)과 공공측면(Caudle et al., 1991)에서 이루어졌었고, 최근에는 IT 관리자들에게 의미가 있는 이슈들을 조사하여 분석하였다(Luftman and Derksen, 2012). 그리고, 실무적으로는 Gartner Group이나 한국정보화진흥원과 같은 전문기관들이 해마다 차년도의 IT 트렌드를 예측하여 발표하였다. 그러나, 이러한 트렌드 예측이 얼마나 정확했는지를 확인한 연구는 거의 없었다.

본 연구는 해마다 공공기관에서 발표하는 차년도의 IT 트렌드 예측이 트위터와 같은 사회관계망에서 실제로 얼마나 언급되는지를 살펴보고자 한다. 또한 한국정보산업진흥원이나 한국정보화진흥원에서 발표하는 차년도 IT 트렌드의 용어들이 실제로 트위터에서 언급되는 빈도를 조사하여 그 상관관계를 밝히고, IT 트렌드가 실제 비즈니스에 나타나는지를 살펴보기 위하여 한국의 공공기관 발주 사이트인 나라장터를 검색하여 해당 용어들이 포함된 프로젝트들의 빈도를 살펴보고자 한다.

2. 이론적 배경

2.1. 빅데이터

Wikipedia에 따르면, 빅데이터란 기존 데이터베이스 관리도구로 데이터를 수집, 저장, 관리, 분석할 수 있는 역량을 넘어서는 대량의 정형 또는 비정형 데이터 집합(Manyika et al., 2011) 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술을 의미한다(Gantz and Reinsel, 2011). 이러한 정의는 기술과 그 활용이라는 측면에서 과학의 한 분야로서의 빅데이터를 개념적으로 정의하고 있다. 실무적인 관점에서 Beyer and Laney(2012)는 향상된 시사점(Insight)과 더 나은 의사 결정을 위해 사용되는 비용 효율이 높고, 혁신적이며, 대용량, 고속 및 다양성의 특성을 가진 정보 자산이라고 하였고, Manyika et al. (2011)는 일반적 데이터베이스SW가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터, 그리고, IDC는 다양한 종류의 대규모 데이터로부터 낮은 비용으로 가치를 추출하고 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처라고 정의하였다.

오랫동안 데이터의 중요성은 주로 그 품질에 근거하여 평가되었다. 예를 들어, Garbage In

Garbage Out은 데이터의 품질에 따라서 그 출력물인 정보의 품질이 결정된다는 것이다. Klein (2001)은 어떤 정보의 원천(Source)인 데이터의 품질에 따라서 정보의 품질이 결정된다고 하였다. 이러한 관점에서 기존의 많은 연구들은 데이터 품질을 구성하는 다양한 속성들을 밝혀내었다. Wang and Strong (1996)의 연구는 데이터의 정확성뿐만 아니라 관련성(Relevancy), 표현(Representation), 그리고 접근성(Accessibility)도 중요하다는 것을 밝혔다. Wang에 의해 시작된 데이터 품질에 관한 연구는 다른 연구자들에 의해 더욱 정교하게 발전되었고, 더 많은 데이터 속성들을 밝혀내었고, 데이터의 용도와 사용자에 따라 데이터의 품질이 다르게 나타난다고 하였다(Madnick et al., 2009). 그러나, 데이터의 품질을 결정하는 가장 근본적인 요인이라고 할 수 있는 양(Amount, Volume)과 원천(Source)에 대한 관심은 상대적으로 부족하였다. 이것은 현실적으로 너무 많은 원천으로부터 생산되는 많은 데이터가 존재하지만, 그것을 저장하고 처리하여 정보를 만드는데 있어서 한계가 있다는 가정이 깔려 있었기 때문이다. 빅데이터 기술은 대용량 데이터의 수집과 처리를 가능하게 하였고, 공공기관이나 통신분야에서 공개된 다양한 유형의 데이터는 외부 데이터의 수집을 가능하게 하였다.

빅데이터에 대한 관심이 집중되고 이를 저장, 활용할 수 있는 기술이 발달하면서 다양한 영역에서 이를 수집하고 분석하여 활용하는 사례가 나타나고 있다. 기업은 이윤 창출의 목적을 위해서, 공공기관은 정책의 수립과정에서 그리고 정책의 시행도중 발생하는 시행착오를 줄이기 위해서 빅데이터를 활용하고 있다. 이렇듯 다양한 분야에서 목적을 위해 빅데이터를 활용하고 있으며, 이를 위한 빅데이터는 대부분 트위터, 페이스북, 구글을 비롯한 검색엔진에서 생성되는 소셜 네트워크 서비스의 활용 데이터가 많은 부분을 차지하고 있다. 트위터는 최근에 개발되어진 short-message system에 불과했지만, 지금은 전세계 1.4억 명이 넘는 이용자가 하루 평균 3.4억개 이상의 트윗을 발생시킨다(Yoon et al., 2013). 이런 소셜 네트워크에서 발생되는 대량의 데이터는 그 자체로도 의미와 정보를 가질 수 있다. 소셜 네크워크의 활용 초기에는 소셜 네트워크 자체의 수집과 분석을 통해 특정 제품에 대한 반응, 선거시 특정 정당 및 후보자에 대한 반응들을 조사하여 활용하였다. 이제 기업과 기관에서는 외부의 데이터에 내부의 데이터를 결합 시킴으로써 그동안 파악하지 못했던 다양한 결과를 얻고자 시도하고 있다. 기업 내부에서 관리되는 고객의 ID와 소셜 네트워크에서 고객의 ID간의 유사성을 파악하여 해당 기업의 고객이 소셜 네트워크에서 쏟아내는 정보만을 찾아서 고객이 원하는 바가 무엇인지를 빠르고 정확하게 파악하여 마케팅 및 제품의 기능 연구 개발에 사용할 수가 있다.

2.2. 트위터 데이터 분석

기업과 기관에서는 내부의 데이터뿐만 아니라 페이스북, 트위터, 카카오톡과 같은 외부자료를 수집하고 분석하여 그동안 파악하지 못했던 다양한 결과를 얻고자 시도하고 있다. 이런 사회관계망에서 발생되는 대량의 데이터를 분석하여 특정 제품에 대한 소비자들의 반응(Yoon and Kwon, 2012), 선거결과의 예측(Cha, 2012)등과 같은 정보를 생성하여 활용하고 있다. 트위터는 전세계 1억명이 넘는 이용자가 하루 평균 2억개

이상의 트윗을 발생시킨다. Song(2014)과 Bae et al.(2013;2014)는 2012년 대통령 선거 당시의 174만개 트윗을 분석하여 민심의 흐름을 추적하였고, Ha et al.(2014)는 특정 IT박람회와 관련된 트윗을 분석하여 정보교환의 사회연결망을 분석하고자 하였다. Lee(2013)는 신문기사로부터 사회적인 이슈를 찾아내고, 그 이슈에 대한 트위터상에서 대중의 긍정과 부정의 감성을 파악할 수 있는지를 연구하였다. 또한, 코레일를 대상으로 트위터의 사용으로 기업의 이미지가 좋아질 수 있는지도 연구되었다(Ju et al., 2012).

2.3. IT 트렌드 예측

최근 세계의 경제는 경영환경의 복잡성과 불확실성이 심화되어 기업이 글로벌 경쟁우위 및 미래 성장 동력을 확보하기 위해서는 기존 제품과 신기술 간의 융합 등을 통해 새로운 가치를 창출이 절실하다. 이러한 유망기술 예측을 통해 경영의 불확실성을 미리 준비하는 것이 중요한 요소로 부각되고 있다. 미래에 대한 복잡성과 불확실성이 증대함에 따라 기업이 지속적으로 경쟁하기 위해서는 기업의 한정된 자원과 시간을 고려하여 미래 환경을 예측하고 미래 기술을 확보하기 위한 장기적 전략 수립이 중요하다. 이를 위해서는 예측하고자 하는 사안에 대해 적합한 방법론을 이용하여 신뢰성 있게 미래를 예측하는 것이 중요할 것이다(Ko et al., 2013).

미래예측에 관련된 연구는 장기적인 전략을 수립하기 위해 외부체제의 복잡성과 불안정성이 증대하고, 급격한 변화로 인한 충격의 증대를 반영하여 예측자체의 정확도보다는 최악의 경우에 대비하면서 최선의 목표를 추구하기 위한 전략적인 접근에 중점을 두는 추세이다(Jung, 2006). 이러한 미래예측을 위한 조사방법은 접근방법에 따라 규범적 방법, 탐구적 방법, 직관적 방법 등으로 구분 되며, 탐구적 방법은 현재에 대한 이해를 바탕으로 미래에 무엇이 발생할지 예측하는 기술능력 지향적인(Capability-oriented) 예측방법이다. 대표적인 탐구적 방법으로는 시나리오(Scenario), 델파이(Delphi), 분석적 계층절차(AHP; Analytic Hierarchy Process)등이 있다. 규범적 기법은 희망하는 미래 또는 미래의 니즈를 충족하기 위한 기술능력 등을 예측하는 것으로, 목표지향적인(Goal-Oriented) 예측방법이다. 대표적인 규범적 기법으로는 시뮬레이션(Simulation), 연관수목법(Relevance tree)등이 있다. 또한, 미래예측 방법론은 데이터 원천에 따라 문헌기반 기법과 전문가기반 기법으로 구분할 수 있다(Kostoff, 1999; Kostoff and Geisler, 1999). 문헌기반 기법은 문헌, 특허 등 대용량의 데이터로부터 예측 분석을 하기 위하여 데이터마이닝, 네트워크 분석 등에 초점을 맞추고 있는 방법이며, 전문가 기반 기법은 기술 발전 방향예측 및 향후 발생 가능한 시나리오를 개발하기 위하여 관련 전문가의 지식을 활용하는 방법이다. 특히 문헌기반 기법은 기존 문헌에 대해 계량적인 방법으로 분석함으로써, 연구 동향을 파악하는데 유용하다(Ko et al., 2013).

IT(Information Technology) 분야에서는 매년 연말이 되면 다음 년도에 유행할 IT트렌드에 대해서 유수의 컨설팅 업체들과 조사기관에서 발표를 한다. Gartner의 경우 향후 3년간 IT업계에 상당한 영향을 미칠 잠재력을 가진 10대 IT기술을 발표하며, 국내에서는 정보통신산업진흥원에서 2013년 IT트렌드 예측을 위해 IT업계종사자 723명을 대상으로 하여 2012년 9월 5일부터 14일까지 실시한 설문 조사 결과를 바탕으로 2012

년 10월에 IT산업 10대 이슈를 발표하였다. 이러한 각 기관의 IT트렌드 발표를 통해서 기업은 향후 유행할 기술에 대한 전망과 이를 통한 비즈니스의 가치 창출과 수익향상을 기대하고 준비할 수 있으며, IT종사자는 미래에 유행할 기술에 대한 가치 투자를 통해 지속적인 IT 기술 경쟁력을 갖출 수 있다. 그러나 이러한 기관들의 예측이 얼마나 정확한지에 대한 사후분석은 아직까지 없었다. 따라서 본 연구는 트위터와 같은 사회관계망에서 주고받는 대화들을 수집하고 분석하면, IT 트렌드의 변화를 확인하고 예측하는데 도움이 될 것이라고 가정하였다.

본 연구는 매년 연말이 되면 국내외의 전문기관이나 기업에서 발표되는 차년도에 유행할 10대 IT트렌드 또는, 유망 IT기술 등의 발표자료에서 거론된 핵심기술과 관련된 용어들이 트위터상에서 트윗되는 빈도를 파악하고자 한다. 이후, 트윗에서 관련용어들이 거론된 빈도와 국가조달시스템인 나라장터(http://www.g2b.go.kr)에서 해당 IT 트렌드와 관련된 조달공고의 빈도를 비교하여 트위터에서 트윗 및 리트윗 되는 IT관련 내용을 수집, 분석하면 전문기관에서 발표된 IT관련 유망 기술 중 실제 현장에서 유용하게 사용되거나, 검토 중인 기술이 무엇인지를 설명할 수 있을 것으로 가정하였다.

3. 데이터

3.1. IT 트렌드 대상 키워드 선정

본 연구에서는 매년 연말이 되면 전문기관들에서 발표되는 차년도에 유행할 10대 IT트렌드 또는, 유망 IT기술등의 발표자료에서 거론된 핵심기술들이 트위터상에서 트윗되는 횟수를 파악하고자 한다. 트위터에서 트윗 및 리트윗 되는 IT관련 내용을 수집, 분석하면 전문기관에서 발표된 IT관련 유망기술 중 실제 현장에서 유용하게 사용되거나, 검토중인 기술이 무엇인지를 설명할 수 있을 것으로 가정하였다. 본 연구를 위해 <Table 1> 에서 보는 바와 같이 정보통신산업진흥원과 한국정보화진흥원에서 2012년에 발표한 차년도에 유망할 것으로 예상된다고 발표한 20가지의 IT 트렌드들 중에서 중복되는 것들을 제거하고 11개의 트렌드를 선정하였다.

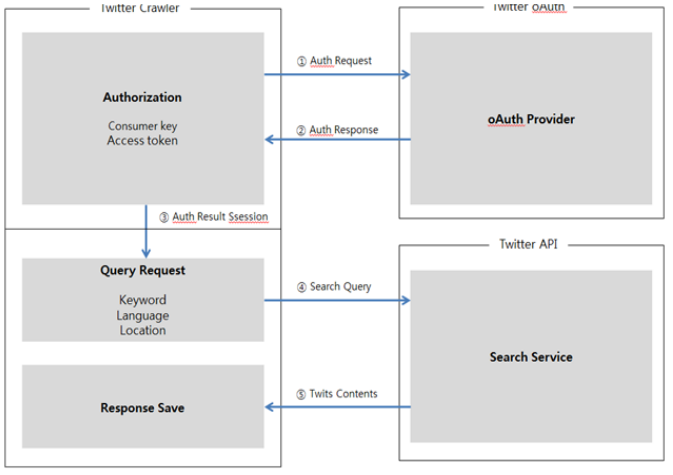
3.2. 데이터 수집

<Table 1>에서 채택된 용어가 포함된 트윗을 수집하기 위해서 트위터에서 제공하는 Rest 방식의 API를 활용하였다. 해당 프로그램은 API를 통해 키워드 쿼리를 전송하고 이에 대한 응답으로 전송되는 트윗을 저장하는 크롤러 역할을 수행한다. 해당 API를 통해 사용자가 설정한 쿼리에 대해 질의를 하고 응답하는 결과를 파일로 저장하는 형태는 Text와 XML을 지원하며, 저장하는 문서의 템플릿을 제공하는 형태로 구성하였다. 트위터의 API를 이용하기 위해서는 아이디와 비밀번호가 아닌 Consumer key와 Access Token을 이용해야 하며, 해당 Key와 Token은 트위터의 API URL(https://dev.twitter.com/apps)을 통해 Application을 등록하여 발급받는다. 트위터 API는 1.1 버전을 이용하였으며, 트위터 API 1.1의 rate limit으로 인해서 한 번의 질의로 최대 18,000개의 트윗 데이터를 수집할 수 있다.

트윗 데이터를 수집할 시에 관련 지역에 대한 위치 지정은 도시의 경계를 가급적 벗어나지 않으면서 가능한 많은 지역을 포함할 수 있는 중심좌표와 반경을 지정하는 방식을 사용하였다

〈Table1〉 2013 IT Trends announced by NIPA and NIA

National IT Industry Promotion Agency (NIPA)	National Information Society Agency (NIA)	Selected IT Topics
Introduction and Utilization of Big Data technology	Big Data	Big Data
Introduction and Diffusion of Cloud Computing	Cloud Computing	Cloud Computing
New Types of Security Threat	Security	Security Intelligence
Smart Home and Home Appliances Service	Smart Home/Home Appliances	Smart Home
HTML5	HTML5	HTML5
Social Media and Enterprise	Social Network Service	SNS
Differentiated Contents and Service Competition	Smart Car	Smart Car
IT Policy and New Government	Green IT	Green IT
Next Generation Semiconductor and Display	Tablet PC	Next Generation Semiconductor
		Tablet PC
The Importance of Patent and Intellectual Property Protection	UI/UX	UI/UX



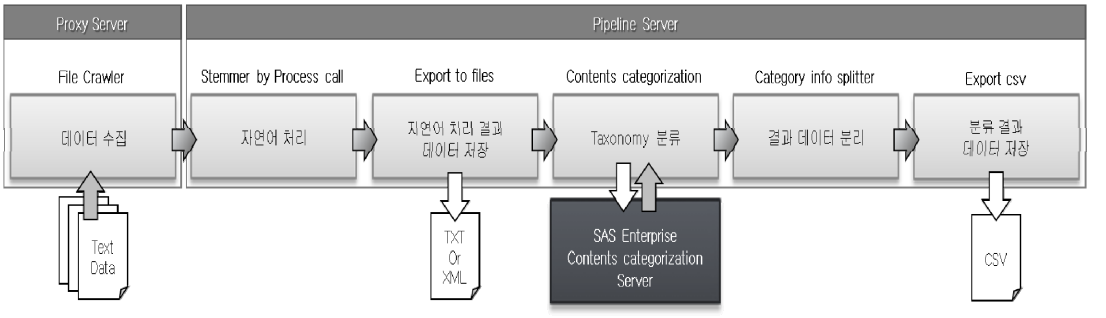
〈Figure 1〉 Program Execution Structure Using Twitter API

(Yoon et al., 2013). 트위터 수집을 위한 중심좌표는 위도 37.568889와 경도 126.976667이었고, 반경은 51.2 km이었다.

3.3. 데이터 분석 방법

비정형 텍스트 데이터의 수집과 분석을 위하여 SAS IRS(Information Retrieval Studio)를 사용하였다. SAS IRS는 Web, Feed(Twitter, Facebook, SNS등), File기반의 비정형 데이터를 수집하고, 자연어 처리(Natural Language Processing), 비정형 데이터 분류를 위한 SAS ECC 서버 가동 및 결과 데이터 수집 및 처리, 분류 결과 File 생성 등에 모든 과정을 UI(User Interface)를 통하여 쉽게 사용자가 이용할 수 있으며 해당 SAS IRS에 프로세스는 <Figure2>와 같이 6단계를 거쳐서 진행된다. <Figure2>의 주요 프로세스 정의는 다음과 같다.

• **Stemmer by Process call** 은 비정형 텍스트 데이터를 컴퓨터가 처리할 수 있도록 변환하는 자연어 처리(Natural Language Processing) 프로세스이다. 이 프로세스를 통하여 File Crawler를 통해 저장된 비정형 텍스트 데이터(Body)는 형태소(뜻(의미)을 가진 가장 작은 말의 단위) 분석을



〈Figure 2〉 SAS IRS (Information Retrieval Studio) Process

통하여 자연어 처리(Natural Language Processing) 되어 SAS IRS 서버에 저장되는 과정이다.

• **Export to File**은 사전처리가 완료되어 SAS IRS 서버에 저장되어있는 데이터를 파일 형태로 저장하는 프로세스이다. 이 프로세스를 통하여 Stemmer by Process call 프로세스에 결과 데이터는 사용자가 설정한 특정 위치 폴더에 ID(Filename)와 Body(Data)가 포함된TXT 또는 XML 파일형태로 저장됨을 뜻한다.

• **Contents Categorization**은 수집된 비정형 텍스트 데이터를 분류하기 위하여 ECC(Enterprise Contents Categorization) 서버에 수집된 데이터를 전달하고 분류결과 데이터를 다시 IRS 서버에 저장하는 프로세스이다. 이 프로세스를 통하여 Stemmer by Process call 프로세스에 결과 데이터는 ECC(Enterprise Contents Categorization) 서버로 전달되며, 전달된 데이터는 사용자가 정의한 Taxonomy(분류기법, 분류체계)를 기반으로 자동분류 되어진다.

수집된 트윗은 6개의 과정을 통하여 키워드의 포함여부를 판단하게 되는데, 2013년 12월 11일 수집 된 데이터 중 “https://twitter.com/Borichaa/status/410742412030787584” 트윗을 예를 들어 설명하면, 첫번째 단계에서는 File Crawler를 통해 수집된 트위터 데이터를 ‘Table 2-1) Result 1’과 같이 저장한 후, Stemmer by Process call 프로세스를 통하여 트위터의 내용을 ‘Table 2-2) Result 2’과 같은 형태의 자연어 변환 처리시켜 ‘Table 2-3) Result 2’과 같은 텍스트 파일 형태로 저장된다. 그 다음 단계인, Contents categorization 과정에서는 자연어 처리된 텍스트 데이터를 SAS ECC서버를 통해 IT트렌드 대상 선정용어 중 어느 하나에 포함되는지 여부를 판단하게 된다. 또한 Category info splitter에서는 IT트렌드대상 선정용어 중 어느 분류에 포함되는지를 판단하여 저장하게 된다. 아래의 <Table 2>에서는 대상이 되는 트위터 데이터가 ‘빅데이터’와 ‘클라우드’ 두 개의 대상 용어를 포함하고 있으므로 ‘빅데이터’는 ‘Table2의 Result 5-1’ 형태로, '클라우드'는 ‘Table2의 Result 5-2’ 형태로 나뉘어 ‘Table2의 Result 6’과 같이 csv 파일로 저장이 되어진다. File Crawler는 사용자가 설정한 특정 위치 폴더와 하위폴더에 존재하는 다양한 형태의 파일(TXT, CSV, HTML, Excel등)을 자동으로 수집하는 프로세스이다. 이 프로세스를 통하여 사용자는 데이터 갱신 주기, 파일 인코딩 형식 및

〈Table 2〉 Tweet data processing results - example

1) Result File Crawler		410742412030787584, “[지디넷코리아] 클라우드와 빅데이터 기술이 확산되면서 변화에 유연하게 대응할 수 있는 소프트웨어 개발 방법론인 '애자일'에 주목해야 한다는 목소리도 커졌다. 애자일은 최선을 다 하더라도 IT 프로젝트는... http://t.co/F6339nboLy”
2)Result Stemmer by Process Call		410742412030787584, "[지 디 넷 코리아] 클라우드 와 빅데이터 기술 이 확산되다 면서 변화 에 유연하다 게 대응하다 ㄹ 수 있다 는 소프트웨어 개발 방법론 이다 ㄴ ' 애자 이다 ㄹ ' 에 주목하다 어야 하다 ㄴ다 는 목소리 도 커지다 었 다 . 애 자일 은 최선 을 다 하다 더라도 IT 프로젝트 는 ... http://t.co/F6339nboLy”
3)Result Export to File		410742412030787584.txt
4)Result Contents Categorization		410742412030787584,Top/IT_Trend/BigData:1.0:빅데이터;Top/IT_Trend/Cloud:1.0:클라우드, "[지 디 넷 코리아] 클라우드 와 빅데이터 기술 이 확산되다 면서 변화 에 유연하다 게 대응하다 ㄹ 수 있다 는 소프트웨어 개발 방법론 이다 ㄴ ' 애자 이다 ㄹ ' 에 주목하다 어야 하다 ㄴ다 는 목소리 도 커지다 었 다 . 애 자일 은 최선 을 다 하다 더라도 IT 프로젝트 는 ... http://t.co/F6339nboLy"
5)Result Category info Splitter	Result 5-1	410638704303038464,Top/IT_Trend/BigData,1.0,빅데이터, "[지 디 넷 코리아] 클라우드 와 빅데이터 기술 이 확산되다 면서 변화 에 유연하다 게 대응하다 ㄹ 수 있다 는 소프트웨어 개발 방법론 이다 ㄴ ' 애자 이다 ㄹ ' 에 주목하다 어야 하다 ㄴ다 는 목소리 도 커지다 었 다 . 애 자일 은 최선 을 다 하다 더라도 IT 프로젝트 는 ... http://t.co/F6339nboLy"
	Result 5-2	410638704303038464,Top/IT_Trend/Cloud,1.0,클라우드, "[지 디 넷 코리아] 클라우드 와 빅데이터 기술 이 확산되다 면서 변화 에 유연하다 게 대응하다 ㄹ 수 있다 는 소프트웨어 개발 방법론 이다 ㄴ ' 애자 이다 ㄹ ' 에 주목하다 어야 하다 ㄴ다 는 목소리 도 커지다 었 다 . 애 자일 은 최선 을 다 하다 더라도 IT 프로젝트 는 ... http://t.co/F6339nboLy"
6) Result Export CSV		IT_TREND_KOR_YYYYMMDD.csv

〈Table 3〉 Contents categorization Output

Column Name	Column Description	Example
ID	Filename Information	389199910278422528
Body	Natural language processing text data	410637593684897792,Top/IT_Trend/BigData,1.0,빅데이터;Top/IT_Trend/Cloud,1.0,클라우드,"HDS , 2014 년 아태 지역 IT 시장 전망 발표 … 빅데이터 클라우드 확산 시기 http://t.co/CuY95BJdmq @ eyeball 에서"
Match words	User defined Category classification keywords	빅데이터, 클라우드
Relevance	Category classification keyword relevance	1.0,1.0
Category	User defined Category ID	Top/IT_Trend/BigData ,Top/IT_Trend/Cloud

수집 파일에 대한 파일크기, 파일생성일 기준 수집범위 등에 설정을 통하여 다양한 방법에 파일 데이터를 수집을 할 수 있으며, 이렇게 수집된 데이터는 File Crawler 프로세스를 통하여 ID(Filename), Body(Text Data) 형태의 정형화된 데이터로 <Table3>과 같이 SAS IRS 서버에 저장

〈Table 4〉 Category info splitter Output

INPUT	OUTPUT
410637593684897792,Top/IT_Trend/BigData,1.0,빅데이터;Top/IT_Trend/Cloud,1.0,클라우드,"HDS , 2014 년 아태 지역 IT 시장 전망 발표 … 빅데이터 클라우드 확산 시기 http://t.co/CuY95BJdmq @ eyeball 에서"	410637593684897792,Top/IT_Trend/BigData,1.0,빅데이터,"HDS , 2014 년 아태 지역 IT 시장 전망 발표 … 빅데이터 클라우드 확산 시기 http://t.co/CuY95BJdmq @ eyeball 에서"
	410637593684897792,Top/IT_Trend/Cloud,1.0,클라우드,"HDS , 2014 년 아태 지역 IT 시장 전망 발표 … 빅데이터 클라우드 확산 시기 http://t.co/CuY95BJdmq @ eyeball 에서"

된다. 스는 CSV 파일 생성에 있어 데이터 추가방법

	A	B	C	D	E	F
1	id	categories	relevance	matchwords	body	
2	413799296824123000	Top/IT_Trend/Smart_Home	1	스마트 홈	[올레 핫 클릭] 스마트 홈 폰 HD mini 추천하다 고 추천 받다 으면 모두 에게 백화점 상품권 을 드리다 0	
3	413805642923524000	Top/IT_Trend/BigData	1	빅 데이터	BIG to GREAT Story 업계 최대 빅 데이터 로 행복 , 진화하다 다 ! http://t.co/04a9fGDaIE http://t.co/FCcI3	
4	413806465149714000	Top/IT_Trend/BigData	1	빅 데이터	빅 데이터 의 속성 http://t.co/yR3JR5XzSo	
5	413806473861275000	Top/IT_Trend/Tablet_PC	1	태블릿 PC	태블릿 PC 처럼 얇다 은 슬림 노트북 소니 바이오 탭 11 http://t.co/xq9VnToQxF	
6	413810387545387000	Top/IT_Trend/BigData	1	빅 데이터	" " 빅 데이터 시장 , 외세 맞서다 어 국산 SW 뭉치다 어야 " " good read http://t.co/8dJX4k1Jmi	
7	413824566675255000	Top/IT_Trend/BigData	1	빅 데이터	[성남 뉴스] 망막 에 구멍 이 … ' 망막박리 ' / 우세 준 최남 경 교수 건강보험심사평가원 빅 데이터 이용 ,	
8	413826461892878000	Top/IT_Trend/Cloud	1	클라우드	[네이버] 뉴스 : 세계 최대 클라우드 기업 ' 아마존 ' 중국 진출 http://t.co/fX0Q7DDRnN	
9	413829722796072000	Top/IT_Trend/Cloud	2	클라우드,클라우드	하둡도 클라우드 에서 쓰다 는 흐름 이 확산되다 는 모습 . 클라우데 라 , 마침내 퍼블릭 클라우드 로 입성	
10	413830184614121000	Top/IT_Trend/Cloud	2	클라우드,클라우드	클라우드 만나다 ㄴ 퀴즈 왕 ' 왔슨 ' , 이제 ' 벤처기업 어머니 ' 로 # 클라우드 @ wikitree http://t.co/kJ98k	
11	413830998090993000	Top/IT_Trend/BigData	1	빅 데이터	사이버 사 ' 정치 글 ' 대량 삭제 의혹 … 군 " 빅 데이터 통하다 어 복제 추진 " http://t.co/LITjOIIiU9	
12	413833150293221000	Top/IT_Trend/Cloud	1	클라우드	" @ ha0504 : 노인 들 을 개엿 으로 여기 는 4 인방 정동영 , 유시민 . 천정배 , 김용 민 http://t.co/9il2ldkbl	
13	413834151066759000	Top/IT_Trend/Cloud	1	클라우드	[ICON 동향 정보] 우리나라 , 2 년 연속 클라우드 컴퓨팅 국가경쟁력 상위권 유지 … BSA 조사 발표 " htt	
14	413841108775759000	Top/IT_Trend/BigData	1	빅 데이터	고객 맞춤 마케팅 유통업체 의 빅 데이터 활용 사례 (매일경제신문) 살롱 에서 도 관심 을 갖다 고 유용하	
15	413844179265921000	Top/IT_Trend/Cloud	2	클라우드,클라우드	RT @ delight412 : 하둡도 클라우드 에서 쓰다 는 흐름 이 확산되다 는 모습 . 클라우데 라 , 마침내 퍼블릭	
16	413847079874990000	Top/IT_Trend/Tablet_PC	1	태블릿 PC	http://t.co/eCsnVvk8xs 삼성전자 , 10.5 인치 AM OLED 태블릿 PC 다음 달 출시 윈도우 OS 달다 고 출시	
17	413851638282338000	Top/IT_Trend/Cloud	2	Cloud,클라우드	Cloud security by Tresorit : http://t.co/UUXj2uFGip 보안 을 강조하다 는 클라우드 저장 소 . 5GB 까지 무	
18	413852724204027000	Top/IT_Trend/BigData	1	빅 데이터	RT @ DikesEye : SK , 네이버 가 개인 정보 를 모으다 는 이유 . 개인 정보 가 돈 이 되다 는 ' 빅 데이터 세	
19	413855008606543000	Top/IT_Trend/BigData	1	빅 데이터	RT @ Naburangii : 빅 데이터 산업 활성화 를 위하다 어 ? 공개되다 ㄴ 개인 정보 의 활용 수집 에 새롭다	
20	413855092316438000	Top/IT_Trend/BigData	1	빅 데이터	RT @ CBar _ GaKa : 방통 위 ' 빅 데이터 개인 정보 보호 가이드라인 안 ' 마련 " 공개되다 ㄴ 개인 정보 수	
21	413855409921732000	Top/IT_Trend/BigData	1	빅 데이터	RT @ coreacom : RT @ slownewskr : [빅 데이터 로 창조 경제 ? 주민등록 제도 먼저 폐지하다 라 !	슬로
22	413858822864662000	Top/IT_Trend/Cloud	1	클라우드	편리하다 ㄴ 웹서비스 를 다루 는 클라우드 기술 - 업무 를 효율 화하 는 시간 단축 기술 http://t.co/X6aJW	
23	413879005318635000	Top/IT_Trend/Tablet_PC	1	태블릿 PC	MS , ' 서 피스 프로 2 ' 최신 업데이트 배포 중단 ! 배터리 충전 문제 발견 : 마이크로소프트 의 태블릿 PC	

〈Figure 3〉 Export csv Results

Category info splitter 는 ECC(Enterprise Contents Categorization)서버의 분류결과를 이용하여 Category(Category ID) 기준으로 데이터를 자동으로 분리하는 프로세스이다. 이 프로세스를 통하여 Contents categorization 프로세스 분류된 결과를 Category(Category ID) 기준으로 <Table 4>와 같이 Row 단위로 분리되어 SAS IRS 서버에 저장된다.

Export csv는SAS IRS 서버에 저장되어있는 데이터를 사용자가 설정한 특정 위치 폴더에 CSV 파일 형태로 저장하는 프로세스이다. 이 프로세스는 CSV 파일 생성에 있어 데이터 추가방법(데이터 추가, 재생성), 데이터 구분자, 파일 인코딩 유형, 이용 컬럼 등에 설정을 통하여 사용자에 분석적 데이터 생성 요건이 반영된 CSV 파일을 생성한다. 해당 파일에는 Contents categorization 과정에서 설정한 컬럼이 <Figure 3>과 같이 생성된 것을 확인할 수 있다.

4. 분석결과

트위터 API를 활용하여 2013년 5월 1일부터

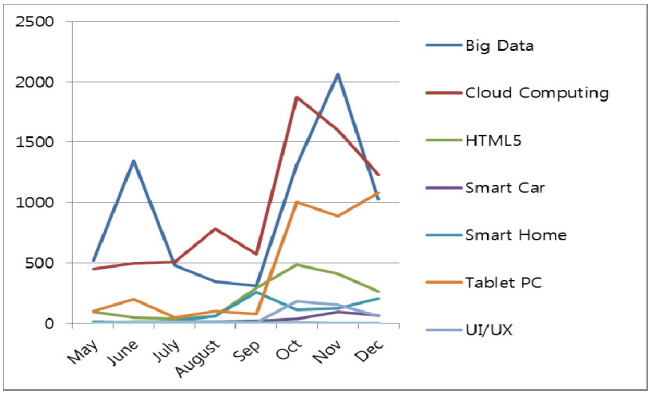
〈Table 5〉Collected tweet by each IT Trend Topic

	May	June	July	August	Sep	Oct	Nov	Dec	Total
Big Data	520	1,345	476	345	307	1,312	2,064	1,028	7,397
Cloud Computing	448	496	510	784	571	1,871	1,598	1,233	7,511
HTML5	95	50	37	60	290	487	406	262	1,687
Smart Car	1	0	0	12	19	37	95	66	230
Smart Home	14	7	18	58	260	114	126	202	799
Tablet PC	99	197	46	101	75	1,005	889	1,080	3,492
UI/UX	0	11	14	13	6	184	154	57	439
New Security Threat	0	0	0	0	0	0	0	0	0
SNS	0	7	7	0	0	8	3	0	25
GreenIT	0	0	0	0	0	5	1	3	9
next generation semiconductor’	0	0	0	0	0	0	0	0	0
Total	1,182	2,119	1,115	1,381	1,537	5,033	5,347	3,943	21,589

2013년 12월 31일까지 서울에서 트윗을 수집하였다. 한국에서는 <Table 1>에서 채택된 IT 트렌드 11개에 대한 결과는 <Table 5>와 같다. 최초에 수집된 트윗데이터에서 테블릿 PC가 실제로 언급된 빈도가 14,201건이였다. 그러나 2013년 5월 20일부터 2013년 6월 16일까지 하이투자증권에서 SNS를 통한 이벤트로 인하여 태블릿 PC를 언급하여 해당 내용은 수집 건수의 합계에서 제외하였다. 예를 들어, 트윗의 내용이 "RT @ hi_hiclass : [하이 투자 증권 SNS 이벤트] 신규 고객 님께 태블릿 PC 매매 수수료 1년간 무료 혜택 드리다"와 같은 트윗은 수집된 데이터에서 제거된 것이다. 결과적으로 한국에서 테블릿 PC의 트윗 빈도는 <Table 5>에서 보듯이 3,492개가 되었다.

<Table5>에서 트윗 수집결과에서 관찰되는 바와 같이 빅데이터와 클라우드 관련 토픽이 트위터에서 가장 많이 언급되었으며, HTML5, 스마트카, 스마트홈, 테블릿PC, UI/UX등도 빈번하게 트위터상에 언급된 것으로 나타났다. 하지만 그밖에 신종보안, 그린IT, 차세대반도체 등은 전혀 언급되지 않았거나, 매우 드물게 언급되어진 것으로 나타났다. 이는 이 토픽들이 아직 보편화되지 않은 합성어이기 때문에 트위터상에 다른 단어들로 언급은 되었을 수 있으나, 정확히 검색어와 같은 용어로 언급되진 않았기 때문으로 해석될 수 있다.

<Figure4>는 지난 8개월간의 IT토픽변화를 보



〈Figure 4〉 IT Topic Trend Changes

〈Table 6〉 Listed Projects on Nara Market in 2012 and 2013

IT Trend Topic	Listed Projects In 2012	Listed Projects In 2013	Collected Tweets
Big Data	9	64	7,397
Cloud Computing	48	90	7,511
Tablet PC	26	63	3,492
HTML5	6	15	1,687
Smart Home	0	0	799
UI/UX	3	2	439
Smart Car	0	0	230
SNS	3	1	25
Green IT	8	0	9
next generation semiconductor	2	1	0
New Security Threat (Security Intelligence)	2	0	0

〈Table 7〉 Listed Projects on Nara Mart - Example(Big Data)

Task	Project No.	Category	Project Name	Agency	Demand Agency	Contract Method	Input Date / Due Date
Service	20130628831-00	Urgent	A Development of the Shared Application System on Big Data	Seoul Public Procurement Service	National Information Society Agency	General - Negotiation	2013-06-26 18:12 / 2013/07/09 13:30
Service	20130617-00	General	Expressway Transport data analysis using big data technology	The Korea Expressway Corporation	The Korea Expressway Corporation	Limit	2013-06-24 0:00 / 2013/08/05 10:00

여주고 있다. 빅데이터와 클라우드와 관련된 트윗은 5월에서 9월까지 어느 정도 언급되다가 10월부터 트윗의 수가 급격하게 늘어나는 현상을 발견할 수 있는데, 이는 10월부터 시작되는 잇따른 차년도 IT트렌드 발표와 빅데이터와 클라우드 같은 토픽들이 또다시 등장하여, 사람들의 재관심을 받게 되는 현상으로 추측해 볼 수 있다. 이는 다시 말해, IT토픽 트렌드 예측발표가 서울지역에서 발생하는 트윗 데이터에 직접적인 영향을 주는 것을 의미하기도 한다.

또한, 본 연구에서는 트위터에서 수집된 데이터를 실제 비즈니스 상황에서 활용되는 정도를 파악하고 비교하고자 하였다. 이를 위해, 한국정부의 공공기관 조달을 담당하는 나라장터 인터넷 홈페이지에서 2012년과 2013년의 국내 IT트렌드 관련 용어(2013년)를 검색하여 조달공고의 건수 데이터를 확보하여 <Table 6>와 같이 트윗 빈도와 비교하였다. <Table 6>와 <Table 7>은 나라장터에서 빅데이터와 관련하여 조달공고를 발표한 사례이다.

〈Table 8〉 Correlation Analysis between Listed Projects on NaraMarket and Collected Tweets

	NaraMarket (2012)	NaraMarket (2013)	Collected Tweets
NaraMarket (2012)	1	0.87988	0.74939
		0.0004	0.0079
NaraMarket (2013)	0.87988	1	0.95179
	0.0004		<.0001
Collected Tweets	0.74939	0.95179	1
	0.0079	<.0001	

또한, 나라장터의 발주 건수와 수집결과가 어느 정도 일치한다는 것을 <Table 6>에서 보여주었다. 통계적인 유의성을 살펴보기 위하여 상관관계 분석을 실하여 <Table 8>와 같은 결과를 얻을 수 있었다.

트윗의 수집결과는 2012년 나라장터와 0.74939에 양의 상관관계를 보이고, 2013년도 나라장터의 IT트렌드와는 0.95179에 강한 양의 상관관계를 가지는 것을 확인할 수 있다. 이러한 분석결과를 통하여 IT 트렌드에 대한 트윗 수집결과는 2012년도 보다 2013년도 나라장터에 더욱 강한 상관관계와 유의수준 보여준다. 이것은 트윗 수집결과는 2013년 나라장터에 IT트렌드 분석을 위한 예측 변수로써 이용이 가능하다는 것을 확인할 수 있다. 다시 말해서, 2012년도에 예측한 트렌드들이 2013년도에 실제로 한국정부의 공공조달사업에 반영되어 나타났다는 것이다.

5. 결론

본 연구에서는 사람들의 관심사 변화를 빠르게 반영하는 트위터 데이터 중 전문기관에서 예측한 특정 IT 토픽들을 포함하는 트윗을 대상으로 예측된 IT트렌드와 실제 트위터상의 IT 트렌드를 비교 분석하였다. 이를 위해 SAS의 빅데이터 툴을 이용하여 한국정보통신산업진흥원과 한국정보화진흥원에서 발표한 11개의 IT트렌드 용어가 실제로 트위터에서 얼마나 언급이 되어왔는지를 살펴보았는데, 트위터 API를 활용하여 2013년 5월부터 12월까지 서울을 중심으로 반경 51.2km까지 언급된 트윗데이터를 수집하였다. 트위터 데이터 분석결과, 전문기관에서 예측한 IT 트렌드 토픽이 실제로 트위터에 자주 언급되고 있으며, 특히 이 트윗빈도는 공공기관의 조달공고데이터와도 통계적으로 유의한 상관관계를 가지고 있는 것으로 나타났다.

본 연구는 설문조사나 전문가 인터뷰를 통하여 IT 트렌드를 분석하는 기존연구들과는 차별화되며, IT중심도시인 서울에서 실제로 지난 8개월동안 트윗되어지는 데이터를 분석하였다는 면에서 그 의미가 크다고 할 수 있다. 하지만, 본 연구에서 사용된 분석대상은 예측된 특정 IT 토픽 트윗으로 그 범위를 한정하여 토픽 변화를 살펴보았는데, 후속연구에서는 토픽 모델링을 통해 범위를 한정하지 않고 트위터 상의 전체 IT관련 토픽의 변화 양상을 살펴보는 것도 방법이 될 수 있을 것이다. 또한 동시출현 단어분석을 기반

으로 구축한 그래프 마이닝 기법들을 적용하여 토픽들의 연관성을 추적하여 융합될 수 있는 IT 토픽들에 대한 연구도 가능할 것이다.

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, J. -h., N. -g. Han, and M. Song, "Twitter Issue Tracking System by Topic Modeling Techniques," *Journal of Intelligence and Information Systems,* Vol.20, No.2(2014), 109~122.

Beyer, M. and D. Laney., "The Importance of Big Data: A Definition," Gartner group, 2012. Available at https://www.gartner.com/doc/2057415 /importance-big-data-definition (Downloaded 10 February, 2015).

Cha, J. P., "Big Data Mining For United State Presidential Election," *IT & Future Strategy*, National Information Society Agency, Vol.12, 1~28, 2012.

Caudle, S. L., W. L. Gorr, and K. E. Newcomer, "Key Information Systems Management Issues for the Public Sector," *MIS Quarterly*, Vol.15, No.2(1991) 171~188.

Gantz, J. and D. Reinsel, "Extracting Value from Chaos," *IDC IVIEW*, 2011. Available at http://www.emc.com/digital_universe (Downloaded 10 February, 2014)

Ha, K. M., H. S. Moon., I. Y. Choi, and J. Kim, "A Network Analysis of Information Exchange using Social Media in ICT Exhibition," *Journal of Intelligence and Information Systems,* Vol.20, No.2(2014), 1~17.

Jung, J. H., "Methodology For Future Prediction," *National Economy*, Vol.17, No.10(2006), 118~125.

Ju, H. -J., J. -Y. Cho, T. -H. Kim, and J. -W. Jeong, "Effects of motives for social media use on corporate image," *The Korean Journal of Local Government Studies*, Vol. 16, No.3(2012), 51~67.

Kho, J., K. Cho, and Y. Cho, "A Study on Recent Research Trend in Management of Technology Using Keywords Network Analysis," *Journal of Intelligence and Information Systems*, Vol.19, No.2(2013), 101~123.

Klein, B. D., "User Perceptions of Data Quality: Internet and Traditional Text Sources," *The Journal of Computer Information Systems*, Vol.41, No. 4(2001), 9~14.

Kostoff, R. N., "Science and Technology Innovation," *Technovation*, Vol.19, No.10(1999), 593~604.

Kostoff, R. N., and E. Geisler., "Strategic Management and Implementation of Textual Data Mining in Goverment Organization," *Technology Analysis and Strategic Management*, Vol.11, No. 4(1999), 493~525.

Lee, K. H., and K. J. Lee, "Twitter Sentiment Analysis for the Recent Trend Extracted from the Newspaper Article," *KIPS Transactions on Software and Data Engineering*, Vol.2, No.10(2013), 731~738.

Luftman, J., and B. Derksen., "Key Issues for IT Executives 2012: Doing More with Less," *MIS Quarterly Executive*, Vol.11, No.4(2012), 207~218.

Madnick, S. E., R. Y. Wang, Y. W. Lee., and H. Zhu., "Overview and Framework for Data and Information Quality Research," *ACM Journal of Data and Information Quality*, Vol.1, No.1(2009), 1~22.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers., "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institue, 2011. Available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (Downloaded 14 November, 2014).

Niederman, F., J. C. Brancheau., and J. C. Wetherbe, "Information Systems Management Issues for the 1990s," *MIS Quarterly*, Vol.15, No. 4(1991), 475~500.

Song. M., "Reading Others' Mind Through Text Mining," *Future Horizon*, Vol.20, No.2(2014), 8~9.

Wang, R. Y., and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, Vol.12, No.4(1996), 5~33.

Yoon, M. Y., and J. E. Kwon, "Global Case Studies on Big Data," *ICT Issue Weekly*, National Information Society Agency, 2012. Available at http://www.nia.or.kr/bbs/board_view.asp?boardid=20111128132107445 8&Order=020201&id=10764 (Downloaded 14 November, 2014).

Yoon, J., S. Kim., B. Lee., and B. -Y. Hwang, "A Correlation Analysis between the Social Signals of Cold Symptoms Extracted from Twitter and the Influence Factors," *Journal of the Korean Multimedia Society*, Vol.16, No.6(2013), 667~677.

Abstract

An Analysis of IT Trends Using Tweet Data

Jin Baek Yi* · Choong Kwon Lee** · Kyung Jin CHA***

Predicting IT trends has been a long and important subject for information systems research. IT trend prediction makes it possible to acknowledge emerging eras of innovation and allocate budgets to prepare against rapidly changing technological trends. Towards the end of each year, various domestic and global organizations predict and announce IT trends for the following year. For example, Gartner Predicts 10 top IT trend during the next year, and these predictions affect IT and industry leaders and organization's basic assumptions about technology and the future of IT, but the accuracy of these reports are difficult to verify. Social media data can be useful tool to verify the accuracy. As social media services have gained in popularity, it is used in a variety of ways, from posting about personal daily life to keeping up to date with news and trends. In the recent years, rates of social media activity in Korea have reached unprecedented levels. Hundreds of millions of users now participate in online social networks and communicate with colleague and friends their opinions and thoughts. In particular, Twitter is currently the major micro blog service, it has an important function named 'tweets' which is to report their current thoughts and actions, comments on news and engage in discussions. For an analysis on IT trends, we chose Tweet data because not only it produces massive unstructured textual data in real time but also it serves as an influential channel for opinion leading on technology. Previous studies found that the tweet data provides useful information and detects the trend of society effectively, these studies also identifies that Twitter can track the issue faster than the other media, newspapers. Therefore, this study investigates how frequently the predicted IT trends for the following year announced by public organizations are mentioned on social network services like Twitter. IT trend predictions for 2013, announced near the end of 2012 from two domestic organizations, the National IT Industry Promotion Agency (NIPA) and the National Information Society Agency (NIA), were used as a basis for this research. The present study analyzes the Twitter data generated from Seoul (Korea) compared with the predictions of the two organizations to analyze the differences. Thus, Twitter data analysis requires various natural language processing techniques,

* SAS Korea
** Corresponding author: Choong Kwon Lee
1095 Dalgubeol-daero, Daegu 704-701, Republic of Korea
Tel: 053-580-6416, Fax: 053-580-6364, E-mail: cklee@kmu.ac.kr
*** Department of Global Business, Kangwon National University

including the removal of stop words, and noun extraction for processing various unrefined forms of unstructured data. To overcome these challenges, we used SAS IRS (Information Retrieval Studio) developed by SAS to capture the trend in real-time processing big stream datasets of Twitter. The system offers a framework for crawling, normalizing, analyzing, indexing and searching tweet data. As a result, we have crawled the entire Twitter sphere in Seoul area and obtained 21,589 tweets in 2013 to review how frequently the IT trend topics announced by the two organizations were mentioned by the people in Seoul. The results shows that most IT trend predicted by NIPA and NIA were all frequently mentioned in Twitter except some topics such as ‘new types of security threat’, ‘green IT’, ‘next generation semiconductor’ since these topics non generalized compound words so they can be mentioned in Twitter with other words. To answer whether the IT trend tweets from Korea is related to the following year's IT trends in real world, we compared Twitter’s trending topics with those in Nara Market, Korea's online e-Procurement system which is a nationwide web-based procurement system, dealing with whole procurement process of all public organizations in Korea. The correlation analysis show that Tweet frequencies on IT trending topics predicted by NIPA and NIA are significantly correlated with frequencies on IT topics mentioned in project announcements by Nara market in 2012 and 2013. The main contribution of our research can be found in the following aspects: i) the IT topic predictions announced by NIPA and NIA can provide an effective guideline to IT professionals and researchers in Korea who are looking for verified IT topic trends in the following topic, ii) researchers can use Twitter to get some useful ideas to detect and predict dynamic trends of technological and social issues.

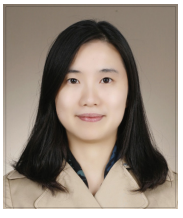
Key Words : IT Trends, Tweeter, Tweet Data, Social Network Service

Received : March 1, 2015 Revised : March 18, 2015 Accepted : March 19, 2015
Type of Submission : Fast Track Corresponding Author : Choong Kwon Lee

저 자 소 개

이 진 백
현재 한국쌔스소프트웨어에서 컨설팅 팀장으로 재직중이다. 대구계명대학교 경영대학에서 경영정보를 전공하였으며, 동 대학에서 경영정보 석사 학위를 취득하였다. 주요 관심분야는 빅데이터와 Text 분석이다.

이 충 권
현재 계명대학교 경영정보학과 MIS전공 교수로 재직중이다. 미국 University of Nebraska-Lincoln에서 박사학위를 취득하고 Georgia Southern University에서 교수로 재직하였다. 주 연구분야는 정보기술인력, 빅데이터, 정보전략이다.



차 경 진
현재 강원대학교 글로벌비즈니스학과 교수로 재직중이다. 호주UTAS(University of Tasmania)에서 BIS(Business Information System)을 취득하고, 동 대학에서 명예학사(Honors Degree)를 취득하였으며, ANU(Australian National University)에서 박사학위를 취득하였다. 주 연구분야는 IT Value Measure, 빅데이터 마이닝, 추천시스템, SmartWork 이다.