ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

소셜 미디어에 기반한 이벤트 정보의 시공간적 시각화를 통한 추이 분석☆

Trend Analysis using Spatial-Temporal Visualization of Event Information based on Social Media

오 효 정[1] 윤 보 현[2] 유 철 중[3] 김 용[4*]
Hyo-Jung Oh Bo-Hyun Yun Cheol-Jung Yoo Yong Kim

요 약

본 논문의 주안점은 다양한 매스 미디어에 나타난 이벤트(event) 정보를 자동으로 인식하고, 이를 시간 및 장소 축으로 시각화함으로써 특정 이벤트의 시간의 흐름에 따른 장소 이동의 추이를 분석하는 데에 있다. 특히 사용자가 직접 작성한 소셜 미디어에 기반하여 이벤트를 추출하고 그들 간의 연속성 분석을 통해 해당 이벤트의 변화 방향성과 사회적 영향을 가늠할 수 있다. 연속성 이벤트의 특성을 규명하기 위해 2년간의 뉴스 기사 및 트윗(tweet)을 수집하여 관련 도메인 선정을 위한 전수조사를 수행하였다. 수행 결과, '질병'과 '여가'도메인을 선정, 본 논문에서 제안한 시각화 방법을 적용한 사례 연구를 통해 시간 및 장소 관점에서의 시각화를 통한 추이 분석의 효용성과 제안된 방법의 유용성을 검증하였다. 특히 단순 사실기반의 연속성 시각과 결과와 사용자의 관심도가 반영된 소셜 미디어에 기반한 연속성 시각화 결과를 비교한 결과, 같은 시기의 이벤트들이라 하더라도 사회적으로 미치는 파장이 큰 장소 이동의 흐름을 파악할 수 있음을 보였다.

☞ 주제어 : 소셜 미디어, 정보 시각화, 자동정보 추출, 추이 분석

ABSTRACT

The main focus of this paper is to analyze trend of event informations in a variety of mass media by graphical visualization in axis of the time and location. Especially, continuity analysis based on user-generated social media can reflect the social impact of a certain event according to change time and location and their directional changes. To reveal the characteristics of continuous events, we survey the data set collected from news articles and tweets during two years. Based on case studies on 'disease' and 'leisure', we verify the effectiveness and usefulness of our proposed method. Even though some events occurred during same period, we showed directional changes which have high-impact in social media referred user interest's, compared with fact-based continuous visualization results,

☞ keyword : Social Media, Information Visualization, Automatic Information Extraction, Trend Analysis

1. 서 론

최근 정보통신 기술이 발전하고 무선 단말기 보급이 폭발적으로 증가함에 따라 소셜 웹 (social web) 서비스의 사용자들은 매년 증가하는 추세이며, 소셜 웹 서비스로부터 생성된 데이터를 분석하여 사용자들에게 다양한 정보 서비스를 제공하는 기업들도 증가하고 있다[1]. 뿐만 아니라 웹기반 정보환경이 웹 2.0 패러다임으로 전환되면서 학술, 연구 분야뿐 아니라 일상 생활 속에서의 정보이용 행태도 기존의 단순 정보 습득이 아닌 양방향 정보 생산의 형태로 변화하고 있다[2,3].

이러한 변화에 따라 뉴스 전달의 매체가 전통적인 오

[1] Electronics and Telecommunications Research Institute (ETRI), Daejeon, 305-700, Korea (first author)
[2] Dept. of Computer Education, Mokwon University, Daejeon, 203-729, Korea
[3] Dept. of Software Engineering, Chonbuk National University, Jeonju-si, 570-756, Korea.
[4] Dept. of Library & Information Science, Chonbuk National University, Jeonju-si, 570-756, Korea.
* Corresponding author (yk9118@jbnu.ac.kr)
[Received 21 July 2014, Reviewed 23 July 2014, Accepted 22 September 2014]
☆ 본 연구는 2014년 미래창조과학부 및 정보통신기술진흥센터의 SW컴퓨팅산업원천기술개발사업의 일환으로 수행하였음. [과제번호: 10044577, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]"

☆ 본 연구는 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업지원을 받아 수행된 것임(No. 한국연구재단에서 부여한 과제번호 : 2010-0022201)

http://dx.doi.org/10.7472/jksii.2014.15.6.65

프라인 인쇄물에서 웹과 모바일로 진화하고 있으며[4], 최근에는 오직 사회망 서비스(Social Network Service, 이하 SNS)를 통하여 뉴스를 제공하는 인터넷 신문 사이트도 많이 등장하고 있다. 특히 SNS의 폭발적인 인기에 힘입어, 개인이 직접 목격하고 경험한 사건/사고 내용의 실시간 전파가 가능해짐에 따라 소셜 미디어가 기존 전통적인 언론사보다 더 빠르고 파급력 있는 뉴스보도의 역할도 겸하고 있는 실정이다[3]. 일반적으로 '소셜 미디어(social media)'는 웹 뉴스, 블로그, 및 SNS에 기반한 페이스북(facebook), 트위터(twitter), 팟캐스트(Podcat) 등을 포함하는 광의의 개념으로 쓰이며[5], 본 논문에서는 이 중에 웹 뉴스와 트위터를 분석 대상으로 한다.

한편, 쏟아지는 정보의 홍수 시대에 대응해 방대한 양의 웹 뉴스 혹은 비정형 텍스트에서 자동적으로 정보를 추출하여 그 주제를 요약하거나 핵심적인 사건 혹은 이벤트를 추출하는 연구가 진행되어 왔다. 특히 그날 처음 발생한 사건, 또는 같은 사건을 다루는 기사들을 탐색하고, 예전에 발생한 사건과 관련된 사건인지를 추적해 나가는 기술을 사건 탐색 및 추적 (Topic Detection and Tracking, 이하 TDT) 기술 이라고 한다[6].

일반적인 의미의 '이벤트(event)'는 문제가 되거나 관심을 끌만한 일을 말하는 반면, 텍스트(text) 정보처리를 위한 정보추출(Information Extraction) 관점에서의 '이벤트'는 사용자 혹은 주어진 주제에 해당하는 '사건'에 해당하는 것으로, 정보추출 대상(target)을 의미한다. TDT 이벤트는 이전 이벤트와 관련하여 어떤 사건이 특정 시간 혹은 특정 장소에서 발생하는 이벤트를 말하며 예를 들어 선거 운동, 재난 복구 과정 등이 있다. 초창기의 TDT 기술은 이벤트의 주제에 따라 문서를 군집/분류하는 수준에 머물렀지만[7], 이후에는 텍스트 문장내의 다양한 어휘 단위자질(lexical feature)과 공기 빈도를 활용하거나[8, 9], 각 카테고리 별로 분류된 용어 자질을 사용하여 이벤트를 추적하는 방향으로 발전하였다[10, 11]. 그러나 이러한 기술은 스트림(stream)으로 입력되는 데이터에서 자동으로 새로운 이벤트를 탐지하거나 입력된 질의(query)나 키워드(keyword)에 해당하는 주제를 탐지하는 것이 대부분이라, 이벤트의 연속적인 변화 추이 분석을 통해 이벤트 정보의 확산 방향 및 미디어간의 확산 양상을 분석하는 본 연구와는 차이가 있다. 특히 본 논문에서와 같이 관련 이벤트들을 시간과 공간의 두 축을 동시에 고려하여 연속성을 분석하는 연구는 없었다.

한편, 소셜 미디어가 폭발적 증가하는 추세에 힘입어 뉴스 이외의 미디어를 대상으로 한 연구가 활발히 진행되고 있다. 특히 사용자의 위치 정보가 함께 제공되는 트위터 데이터를 기반으로 한 연구가 다수 발표되고 있는데, Muslove 외[12]는 지난 3년간의 트윗 문서셋에서 감성 단어를 추출, 지역 간의 분위기(mood) 변화를 분석한 연구 결과를 보였다. 특히 이 연구에서는 본 논문에서와 같이 지도를 통해 트윗내의 분위기 변화를 시각화 하였는데 이는 단순 감성의 차이를 표현한 것으로 본 논문에서 다루고 있는 연속성을 갖는 이벤트의 흐름 분석과는 차이가 있다.

또한 지진과 같은 자연 재해 현상과 관련된 정보의 경우, 트위터의 '실시간' 업데이트 특성을 활용함으로써 뉴스 미디어에서 제공하는 정보보다 더 빠르게 활용될 수 있다는 연구들도 다수 발표되었다[13, 14]. 그러나 위 연구들 역시 지역 혹은 시간의 흐름 등 단편적인 기준에 의해 관련 정보들의 연관성을 파악했다는 한계가 있다. 특히 미국의 토네이도 전문 사이트인 ©United State Tornades [15]에서는 토네이도 이동 경로를 지도와 함께 시각화 한다. 그러나 이 연구 역시 수작업으로 습득한 자료에 기반한 결과라는 한계가 있다. 또한 대부분의 기존 연구들이 특정 미디어 내에서 분석된 이벤트 정보만을 대상으로 하는 반면, 본 연구에서는 같은 이벤트라 하더라도 서로 다른 미디어로부터 수집된 이벤트들의 시공간적 추이를 분석, 그 차이를 비교함으로써 미디어별 특성을 파악하고자 한다.

정리해보면 본 논문에서는 소셜 미디어로부터 수집된 사건/사고들 중 연관된 이벤트를 탐지하고 시간과 장소의 관점에서 시각화함으로써, 이벤트들 간의 연속성을 분석하고자 한다. 특히 기존 연구가 특정 미디어, 혹은 시간, 장소 등 한정된 기준으로만 이벤트를 분석하는 한계를 탈피하고자 한다. 또한 뉴스 미디어를 대상으로 한 단순 사실(fact) 보도 기반의 추이 분석 결과와 사용자의 관심도가 반영된 개인 미디어에 기반한 추이 분석 결과를 비교함으로써, 같은 시기에 발생한 이벤트들이라 하더라도 사회적으로 관심이 높은 이동 경로를 파악하고자 한다.

2. 소셜 미디어 기반 연속성 이벤트 정의

2.1 이벤트의 시공간적 추이 관점에서의 연속성

본 논문의 주안점은 소셜 미디어에 나타난 이벤트 정보를 자동으로 인식, 이를 시간 및 장소 축으로 시각화함으로써 특정 사건의 시공간적 추이, 즉 '연속성'을 분석하는 데에 있다. 이를 위해 먼저 시공간적 관점에서의 연

속성 이벤트 정보의 특성을 규명하고자 한다.

본 논문에서는 임의의 지역 혹은 임의의 시간에 독립적으로 발생한 사건/사고를 '단발성' 이벤트라고 정의한다. 도난 사건이나 교통사고 등이 이에 해당한다. 반면, 시간 및 장소 관점에서의 '연속성'을 갖는 이벤트 정보란 특정 사건이 처음 발생한 후 시간의 흐름 혹은 장소 이동에 따라 관련 있는 해당 이벤트가 반복 혹은 파생되어 발생하는 경우를 의미한다. 주로 재해나 질병 분야의 사건/사고가 이에 해당한다.

단발성 이벤트에 비하여 연속성 이벤트가 사회적으로 미치는 영향이 크기 때문에 온라인 콘텐츠 상에서 발생한 연속성 이벤트를 자동으로 탐지, 이를 추적할 수 있다면 이벤트의 최초 발생 이후의 이벤트 발생 경로와 확산 범위의 분석이 가능해 지며, 이를 통해 신속하고 효과적인 대응방안을 수립하는데 도움이 된다. 본 논문에서는 다음과 같은 기준에 의해 연속성 이벤트의 선정하기로 한다[16].

1) 사회적으로 파급력이 있다.
2) 전염성 혹은 유행성을 띤다.
3) 초기 발생 이후 계속되는 지속성을 보인다.

세부 기준을 보충 설명하면, 사회적으로 파급력이 있다는 의미는 해당 이벤트가 미디어에 통해 보도, 탐지된 이후 사용자들에게 전파가 얼마나 되었는지에 따른 척도이다. 즉 파급력이 큰 이벤트인 경우 뉴스보도 건수, 트윗의 경우, 리트윗 횟수 등이 이를 반영한다고 할 수 있다. 또한 어떤 이벤트가 특정 지역 한 곳에서만 발생한 경우와 여러 장소에서 발발, 장소를 이동하는 경우, 그에 대한 사회적 중요도는 후자가 더 크다. 더불어 해당 이벤트가 일회성이 아닌 특정 기간 동안 계속되는 지속성을 갖는다면 사회적으로 미치는 영향력은 더욱 배가된다. 이 같은 기준은 본 논문에서 제안된 방법을 통해 시각화한 결과의 활용이 높은 이벤트를 선정하기 위함이다.

2.2 소셜 미디어에 나타난 연속성 이벤트 도메인 선정

먼저 소셜 미디어에 출현한 다양한 이벤트들 중 시간 및 장소 관점에서의 연속성을 갖는 이벤트 주제(topic) 즉 도메인(domain)을 선정하기 위해 지난 2년간 (2012년 1월부터 2013년 12월 사이)의 뉴스 기사를 수집하였다. 표 1은 일간지 중 온라인으로 수집이 용이한 대상 뉴스 매체 목록과 관련 URL이고, 표 2는 해당 기간에 수집된 뉴스를 취합한 결과를 월별로 정리한 것으로, 2012년에는 총 1,437,118건, 2013년에는 1,468,380건의 기사를 수집, 총 2,905,498건의 뉴스기사를 수집하였다.

본 논문에서는 소셜 미디어 수집을 위해 웹 크롤러(crawler)를 자체적으로 개발하였으며, 수집된 문서들은 빅 데이터 분석 및 저장을 위한 오픈소스 프로젝트(Open Source Project)인 하둡(Hadoop) 기반의 HBase (http://hadoop.apach.org)를 통해 색인, 저장되었다. 특히 트윗(tweet) 문서 수집을 위해서는 트위터에서 제공하는 Open API(http://api.twitter.com)를 활용하였다.

(표 1) 뉴스미디어 수집 대상 매체
(Table 1) Target Channels of News Collection

뉴스 매체	수집 URL
국민일보	http://news.kukinews.com
한국일보	http://news.hankooki.com
동아일보	http://news.donga.com
Korea Herald	http://view.koreaherald.com
오마이뉴스	http://www.ohmynews.com
한겨레신문	http://kr.hani.feedsportal.com
LAtimes	http://feeds.latimes.com
NYtimes	http://rss.nytimes.com
국방일보	http://kookbang.dema.mil.kr
조선일보	http://chosun.feedsportal.com
연합뉴스	http://yonhapnews.feedsportal.com/
중앙일보	http://article.joins.com

(표 2) 뉴스미디어 수집 결과
(Table 2) News media Collection

년/월	수집 건수	년/월	수집 건수
201201	123,414	201301	119,945
201202	121,605	201302	102,543
201203	131,927	201303	109,577
201204	121,650	201304	113,616
201205	134,272	201305	113,929
201206	106,979	201306	104,975
201207	126,658	201307	127,250
201208	113,747	201308	135,102
201209	101,001	201309	125,821
201210	120,349	201310	142,205
201211	116,602	201311	131,551
201212	118,914	201312	141,866
총 1,437,118		총 1,468,380	
총합: 2,905,498			

수집된 뉴스미디어로부터 시간 및 장소의 연속성이 유지되는 이슈 이벤트 선정을 위한 기초 조사를 실시하였다. 그 결과, 주로 '보건/의료' 관련 뉴스기사에서 '질병' 도메인의 이벤트가 시간과 장소의 연속성을 보이며 보도됨을 알 수 있었다. 우리나라에서 발생하는 '재해/재난' 분야의 사건/사고의 경우에는 시간의 변화에 따른 지리적 이동반경이 매우 짧아 연속성이 발견되는 이벤트가 드물었다.

상기한 기초조사에 기반하여 2012년 6월부터 2012년 12월 사이에 보도된 '보건/의료' 분야 592,734건의 뉴스를 대상으로 전수 조사를 실시하였다. 조사의 편의를 위해 문서 내에 나타난 질병명, 시간 정보, 장소 정보는 한국전자통신연구원(ETRI)에서 개발한 언어분석 및 이벤트 정보추출기[16]를 활용해 자동으로 추출한 결과를 보정하지 않고 사용하였다. 표 3은 전수 조사 결과이다.

(표 3) 보건/의료 뉴스 조사 결과
(Table 3) Survey Result of 'Health' domain

구 분	뉴스 건수
보건/의료 뉴스 전체	592,734
시간 정보가 표현된 뉴스	314,421 (53%)
장소 정보가 표현된 뉴스	32,142 (5%)
질병명이 나타난 뉴스	115,215 (19.4%)
44개 연속성 질병명이 나타난 뉴스	2,142 (3.6%)
시간+장소+연속성 질병명이 함께 나타난 뉴스	175

표 3에 나타나듯이, 뉴스미디어의 특성상 특정 사건 즉 이벤트의 발생 시간 정보를 포함하는 경우가 많았고 (전체 592,734건 중 314,421건으로 53%), 사건 발생 장소 정보가 포함되는 경우는 상대적으로 적게 나타났다 (전체 중 5%). 또한 '보건/의료' 분야의 뉴스임에도 불구하고 실제 명확한 질병명이 나타난 뉴스는 전체 19.4%에 그쳤다. 위 결과를 바탕으로 본 논문에서 다룰 연속성 이벤트의 대상 도메인으로 '보건/의료' 분야에 출현한 다양한 이벤트 중 '질병' 도메인을 우선 선정하기로 한다.

2.3 소셜 미디어에 나타난 연속성 이벤트 선정

실제 시간 및 장소 축 기반 이벤트 추이 분석을 위해서는 대상 도메인뿐 아니라 해당 도메인의 세부 대상 이슈 이벤트가 선정되어야 한다. 이를 위해 앞서 선정된 '질병' 도메인에 해당하는 뉴스를 질병별로 세부 분석하였으며, 상기 2.1에서 기술한 연속성 이벤트의 특성을 만족시키기 위하여 아래와 같은 기준을 적용하였다.

1) 사회적 파급력이 있는지: 발생빈도(건수)
2) 전염성(유행성)이 있는지: 이동반경
3) 초기 발생 이후 지속성이 있는지: 발생기간

세부 조사 결과, 이러한 기준에 해당하는 질병명은 '구제역', '식중독', '신종플루' 등 44개가 초기 선정되었다[16]. 표 3에 나타나듯이, 해당 질병명이 나타난 뉴스 건수는 2,142건으로 전체 3.6%에 해당한 것으로 분석되었다. 연속성 분석을 위해서는 시간 및 장소 정보가 모두 표현된 뉴스만을 대상으로 하는데, 이런 조건을 만족하는 뉴스는 이 중에서 175건이 수집되었다.

44개 질병 대상 이벤트 후보의 발생 빈도 분포를 분석한 결과, '식중독', '신종플루', '구제역', '비브리오', '인플루엔자/AI', '대장균' 순으로 나타났다[16]. 본 논문에서 제안하는 이슈 이벤트의 연속성 분석을 위해서는 소셜 미디어에 보도하고 있는 이벤트가 앞서 제시한 바와 같이 1) 사회적 파급력과 2) 전염성, 3) 지속성이란 특성을 만족하며, 동시에 미디어 내의 문장에 시간과 장소정보가 명시적으로 표현되어 있는 문서를 대상으로 선정해야 한다. 상기한 6개의 상위 분포 질병 이벤트들의 연속성 적합 조건을 분석한 결과, 표 4와 같은 특성을 발견하였다.

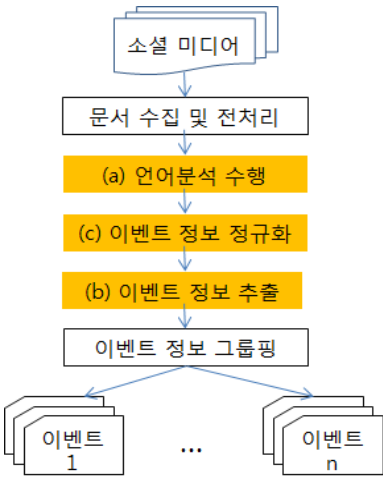
(표 4) 이슈 이벤트 후보별 특성
(Table 4) Characteristics of Issue Event Candidates

질병명	연속성 적합 조건			적합 여부	트윗 빈도
	전염/유행성	지속성	시간/장소		
식중독	O	O	O	O	낮음
신종플루	O	O	해외지명	-	중간
구제역	O	O	O	O	높음
비브리오	-	O	O	-	낮음
인플루엔자	O	O	해외지명	-	중간
대장균	O	-	O	-	낮음

상위 질병 이벤트 중 '비브리오'는 단순 질병군으로 전염성이 없고, '대장균' 역시 지역이나 시간의 흐름에 따른 지속성이 발견되지 않아 대상 후보에서 제외되었다. 연속성 적합 조건을 모두 만족하는 후보 이벤트는 '식중독/신종플루/구제역/인플루엔자'로 선정되었으나, 이 중에서 '신종플루'와 '인플루엔자'와 관련된 뉴스가 대부분 해외에서 동시 다발적으로 발병한 사례를 보도한

경우로, 장소의 연속성을 파악하기 어려운 점을 고려, 제외하였다.

표 4와 같이 시공간 연속성 분석이 적합한 대상으로는 '구제역'과 '식중독'이 선정되었으나, 같은 기간의 트위터 수집 결과를 분석한 결과, '식중독' 질병의 경우 뉴스 미디어에는 다수 발견된 반면 트윗에서의 빈도는 '구제역'보다 상대적으로 적게 분석되었다. 이 같은 특성을 종합적으로 고려한 결과, 본 논문에서는 최종적으로 '구제역' 질병을 이슈 이벤트 후보로 선정, 이들 간에 시간과 장소의 흐름에 따라 어떠한 연속성이 분석되는지를 파악하기로 한다. 참고로 위 선정 결과는 앞서 설명한 바와 같이 2012년 6월부터 2012년 12월 사이의 '질병'관련 뉴스기사의 전수 조사 결과로, 대상 기간을 확대하거나 뉴스 매체 추가를 통해 보다 많은 모수의 뉴스 기사를 분석함으로써 대상 이벤트 후보군의 확장이 가능하다.

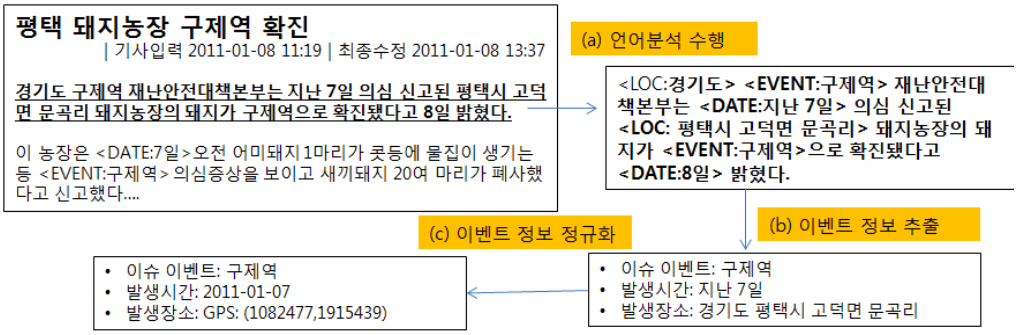


(그림 1) 이벤트 정보 추출 시스템 흐름도
(Figure 1) System Flow of Event Information Extraction

3. 소셜 미디어 기반 이벤트 정보 추출 방법

이벤트 추이 분석을 위해서는 소셜 미디어에 나타난 텍스트 문장에서 이벤트 정보를 추출해야한다. 이를 위해 본 논문에서는 한국전자통신연구원에서 개발한 이벤트 정보추출기[16]를 활용하였다. 그림 1은 이슈 이벤트 정보 추출 과정을 도식화 한 것으로, 먼저 수집된 소셜 미디어로부터 전처리과정을 통해 텍스트 정보를 획득, 언어분석을 수행한다. 언어분석 과정에는 형태소 분석, 개체명 인식 과정이 포함되며, 이때 트윗 문서의 경우 오타나 띄어쓰기 오류 등이 교정된다. 각 문장의 언어분석 결과를 바탕으로 이벤트가 출현할 가능성이 있는 '이벤트 문장'을 선별한다[18]. 이벤트 문장은 이벤트 정보의 핵심 요소로 사건의 구체적인 내용을 포함하고 있는 문장으로, 본 논문에서는 '구제역' 질병이나 '벚꽃' 축제 관련 이벤트가 나타난 문장을 대상으로 삼았다. 이벤트 '발생 시간'과 이벤트 '발생 장소'는 연속성을 분석하기 위한 기준으로 사용된다. 특히 해당 어휘 정보의 빈도 밀도(tf-density)에 기반하여 가중치를 추출하였다[19]. 이벤트 발생 시간과 장소 정보는 언어분석을 통해 인식된 개체명 중 시간 표현어구, 지명, 건물, 주소 등의 정소를 후보로 한다.

그림 2는 2011년 1월 8월의 보도된 '구제역' 관련 기사로부터 이벤트 정보를 자동 추출한 결과를 예시한 것



(그림 2) 소셜 미디어에 나타난 '구제역' 이벤트 정보 추출 결과 예시
(Figure 2) An Example of Event Information Extraction Result for 'foot-and-moth disease'

으로, 기사의 원문(original text)부터 (a) 언어분석을 수행, 이벤트 정보의 후보가 될 수 있는 어휘 자질을 선별한다. 문장 내의 <EVENT>, <LOC>, <DATE> 태그는 각각 이슈 후보 이벤트, 이벤트 발생 시간 정보, 이벤트 발생 장소 정보의 후보를 의미한다[17]. 추출된 후보로부터 이벤트 정보 템플릿(template)을 구축, 최종 추출된 이벤트 정보를 요약하면 그림 2의 (b)와 같다. 특히 실제 문장에 다양한 방식으로 표현된 시간정보와 장소정보는 (c)와 같이 각각 8자리 숫자(YYYY-MM-DD)와 GPS 좌표 체계로 정규화(normalization) 된다[16].

4. 이벤트 정보의 시공간적 시각화를 통한 추이 분석 결과

이번 장에서는 본 논문에서 제안한 소셜 미디어 기반 이벤트 정보의 시공간적 시각화 결과를 통한 추이 분석의 유용성과 제안 방법의 타당성을 검증하기 위해 2가지 사례연구를 실시하였다.

1) [사례연구1]: '질명' 도메인의 '구제역' 이벤트
2) [사례연구2]: '여가/레저' 도메인의 '벚꽃' 이벤트

먼저 소셜 미디어에 기반한 이벤트 정보의 시공간적 연속성 분석의 시각화 결과의 유용성을 검증하기 위해, 앞서 3장에서 설명한 연속성 이벤트의 기준에 의해 선정된 '구제역' 이벤트를 대상으로 연속성 분석 시각화 결과를 제시하였다. 또한, 웹 뉴스와 트위터 간의 상관관계 및 두 소셜 미디어 간의 정보 확산 방향을 비교 분석한다.

[사례연구2]는 '구제역' 이벤트의 선정 과정과 동일한 절차와 기준을 '여가/레저' 도메인으로 확장한 결과로, 본 논문에서 제안된 방법론이 특정 도메인에 국한된 방법이 아님을 보이고자 한다. 특히 제안된 방법을 적용한 결과를 비교 분석하여 연속성 이벤트 시각화 방법의 확장성과 타당성을 검증하고자 한다.

4.1 '질병' 도메인 '구제역' 이벤트 추이 분석

제안된 이벤트 연속성 분석 방법의 효과를 증명하기 위해 '구제역' 사건이 집중 보도된 2010년 11월부터 2011년 2월 사이의 신문기사를 수집, 250건의 문서를 수집하였으며, 같은 기간 내의 트위터 데이터 중 '구제역' 관련 트윗을 수집한 결과, 총 982건이 수집되었다. 해당 문서 집합내의 1,849개의 뉴스 문장과 1,042개의 트윗 문장을 대상으로 상기 4장에서 기술한 이벤트 정보 추출 시스템을 통해 자동으로 이벤트 정보를 추출, 시간 및 장소 축으로 연속성을 분석해보았다.

이슈 이벤트	발생 시간	발생 장소	정규화 좌표	출처/문서
구제역	2010-11-30	경북/안동시/서후면	1104074 1847197	뉴스
Aphtae epizooticae	2010-12-02	경북/영양군	1144091 1853304	트위터
구제역	2010-12-13	경기도/양주시	975049 1959679	보건복지부DB
口蹄疫	2011-01-02	강원도 문막읍	1028115 1923785	뉴스

(그림 3) '구제역' 이벤트 그룹핑 결과
(Figure 3) Grouping Results of 'foot-and-mouth disease' Events

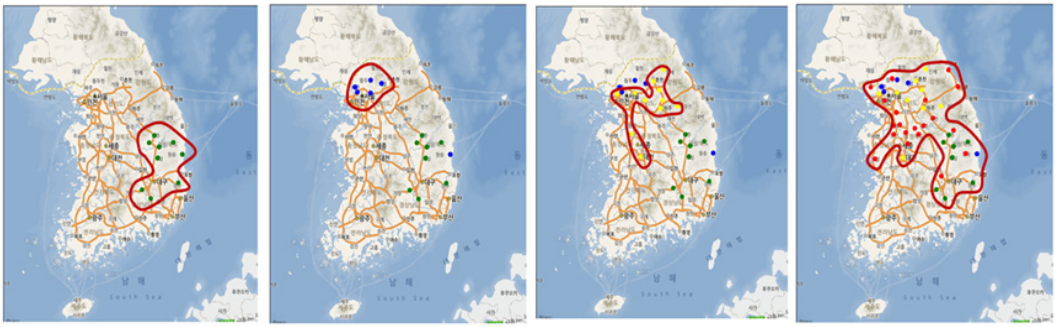
그림 3은 이벤트 정보 추출 시스템을 통해 인식된 '구제역' 이벤트 정보들을 취합, 시간 순으로 정렬된 결과를 예시한 것으로, 소셜 미디어 내에 다양한 형태로 표현된 질병명의 이형태를 인식, 대표 질병명으로 정규화하여 관련 이벤트를 그룹핑(grouping)하였다.

그림 4와 5는 각각 뉴스 미디어와 트위터로부터 분석된 이벤트 그룹핑 결과를 지도를 통해 시각화한 것이다. 지도 시각화를 위해서는 네이버(Naver)에서 제공하는 Open API (http://dev.naver.com/openapi/) 를 활용하였다.

그림 4와 5를 비교해보면, 같은 기간(2010년 11월말부터 2011년 1월 중순까지)에 발생한 '구제역'이벤트의 최종 확산 결과는 유사한 분포를 보이지만, 초기 발견 시점부터 점차 확산되는 전파 양상은 서로 차이가 있는 것을 볼 수 있다. 미디어별로 보다 세부적인 추이 분석을 위해 전파 단계를 구제역이 발생한 시간과 지역을 기준으로 4단계로 분류를 하였다.

그림 4의 뉴스 미디어에서의 '구제역' 이벤트 확산 과정은 살펴보면, 그림 4의 a)는 구제역의 1차 확산 단계로 2010년 11월29일에서 2010년 12월 9일 사이에 발생한 이벤트들을 시각화한 결과이다. 경상북도 안동에서 최초로 발생하여 경상북도 예천군, 의성군, 영양군, 고령군 등 경상북도 내에서 지리적으로 가까운 곳에 위치해 있는 지역들만 이벤트 연속성을 보이고 있다.

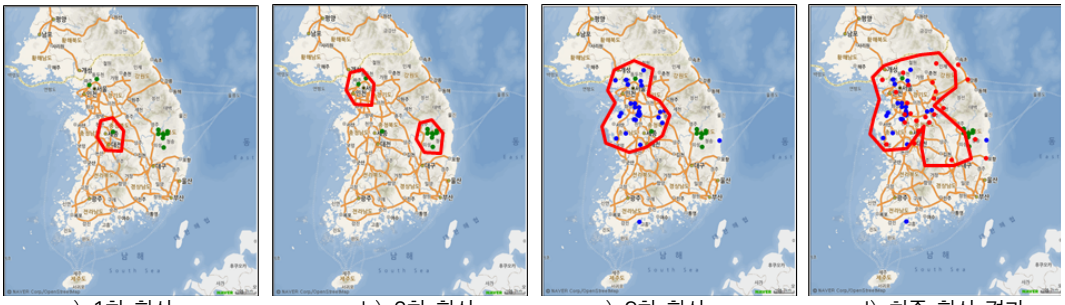
그림 4의 b)은 구제역 2차 확산 단계로 2010년 12월 11일에서 2010년 12월 20일 사이에 발생한 이벤트들의 집합이다. 경기도 양주시, 파주시, 고양시 등 구제역 1차 확산 단계와 비슷한 현상으로 경기도 지역 내에서 지리적으로 가까운 곳에 위치해 있는 지역들만 이벤트 연속성을 보이고 있다.



a) 1차 확산 b) 2차 확산 c) 3차 확산 d) 최종 확산 결과

(그림 4) 뉴스 미디어에서의 '구제역' 이벤트 시각화 결과

(Figure 4) Visualization Result of 'foot-and-mouth disease' Events in News media



a) 1차 확산 b) 2차 확산 c) 3차 확산 d) 최종 확산 결과

(그림 5) 트위터에서의 '구제역' 이벤트 시각화 결과

(Figure 5) Visualization Result of 'foot-and-mouth disease' Events in Twitter

그림 4의 c)는 구제역 3차 확산 단계로 2010년 12월 21일에서 2010년 12월 31일 사이에 발생한 이벤트들의 집합이다. 구제역 3차 확산 단계는 전국적 확산 초기 단계로도 볼 수 있으며 경기도 지역 일대에서 강원도 지역 또는 충청북도 지역으로 이벤트의 연속성이 조금씩 확산되는 현상을 확인할 수 있다.

그림 4의 d)은 구제역 4차 확산 단계로 2011년 1월 1일에서 2011년 1월 15일 기간 사이의 발생한 이벤트 집합이다. 구제역 4차 확산 단계는 구제역이 전국적으로 확산되었음을 확인할 수 있다. 정리해보면, 초기 구제역 1차 확산과 2차 확산 단계에서는 경상북도 지역과 경기도 지역 내에서 지리적으로 가까운 곳만 이벤트 연속성을 보였으나 3차 확산 단계 이후에는 전국적으로 이벤트가 발생하는 현상을 보였다.

그림 5는 같은 기간의 트위터 문서집합에서의 '구제역' 이벤트 확산 과정을 분석한 결과로, 뉴스 미디어 분석결과인 그림 4와 비교해보면 a) 1차 확산 시기에는 충청, 대전 지역에서만 낮은 빈도로 '구제역' 관련 트윗이 인식되었다. b)의 2차 확산 시기에는 뉴스 미디어의 경우, 경상북도 지역에서 발발 후 수도권으로 확산되는 현상을 보인 반면, 트위터에서는 경북과 수도권이 비슷한 시점에 동시 확산된 것으로 분석되었다. 이후 c) 3차 확산 양산은 뉴스 미디어 분석 결과에 비해 더 넓은 수도권 전역으로 확산된 것으로 분석되었으나, 최종 확산 결과는 뉴스 미디어 결과와 유사한 모습을 보였다.

위와 같은 미디어간 결과 비교를 통해, 트위터 기반 연속성 분석 결과는 실제 사용자들이 사는 지역 혹은 관심 지역에 따른 분포를 반영함을 알 수 있다. 특히 충청 지역의 경우에는 뉴스보도보다 먼저 구제역 발발 소식이 언급되었는데, 이는 사용자가 직접 작성한 미디어를 통해 공적 보도 매체인 뉴스보다 더 빠른 사고소식을 전달할 수 있음을 함의한다.

4.2 '여가/레저' 도메인 '벚꽃' 이벤트 추이 분석

본 절에서는 제안한 방법의 효용성을 검증하기 위해 기존 '질병'도메인이 아닌 타 도메인으로의 확장 적용을 시도하였다. 앞서 설명한 바와 같이 연속성을 갖는 이벤트 선정 기준은 사회적으로 관심이 높은 어떤 이벤트가 발생한 후 일정 기간 동안 지리적 변화가 일어나는 이벤트이어야 한다. 다음은 '구제역' 이벤트의 선정 과정과 동일한 절차와 기준에 따라 실시한 기초 조사를 통해 '여가/레저'도메인의 '축제(festival)' 이벤트를 선정하는 과정을 설명한다.

트윗 내용	메타정보
벚꽃 구경하고 오랜만에 던킨!!!! @ 던킨 도너츠 회룡역점	"coordinates" : [127.04770088, 37.72419139], "created_at" : "Sun Apr 15 08:45:39 +0000 2013", "name" : "Ye Heun"
신탄에 벚꽃이 마니 피웠어요 화려한 봄날이었어여	"coordinates" : "null", "created_at" : "Sun Apr 15 12:33:03 +0000 2013", "name" : "문기버ㅁ"
"까악~~ 너무 예쁘다~+^.~* RT "@thanks0408: 밤벚꽃	"coordinates" : [127.42245535, 36.43858324] , "created_at" : "Sun Apr 15 12:34:56 +0000 2013", "name" : "커피&초코브라우니"

(그림 6) '벚꽃' 관련 트윗 문서 예
(Figure 6) Examples of Tweet documents for 'Cheery blossoms'

(표 5) '벚꽃' 관련 추이 분석 대상 트윗 집합
(Table 5) Test Tweet Collection for 'Cheery blossoms

기간	날 짜	트윗 건수
1	2013.02.01.~2013.02.29	28
2	2013.03.01.~2013.03.15	56
3	2013.03.16.~2013.03.31	243
4	2013.04.01.~2013.04.10	915
5	2013.04.11.~2013.04.20	329
6	2013.04.20.~2013.04.30	52
7	2013.05.01.~2013.05.31	47

표 2의 2년치 뉴스 기사를 분석한 결과, '축제' 관련 기사들 중 매년 2월~5월 사이 각 지방에서 각종 꽃의 개화와 관련된 소식이나 행사가 개최되었고, 그 중에서 트위터에서 사용자 관심이 높았던 주제로는 '벚꽃'이 선정되었다. 실험을 위해 2013년 2월부터 5월 사이, '벚꽃'이 출현한 트윗 1,623개를 수집하였다.

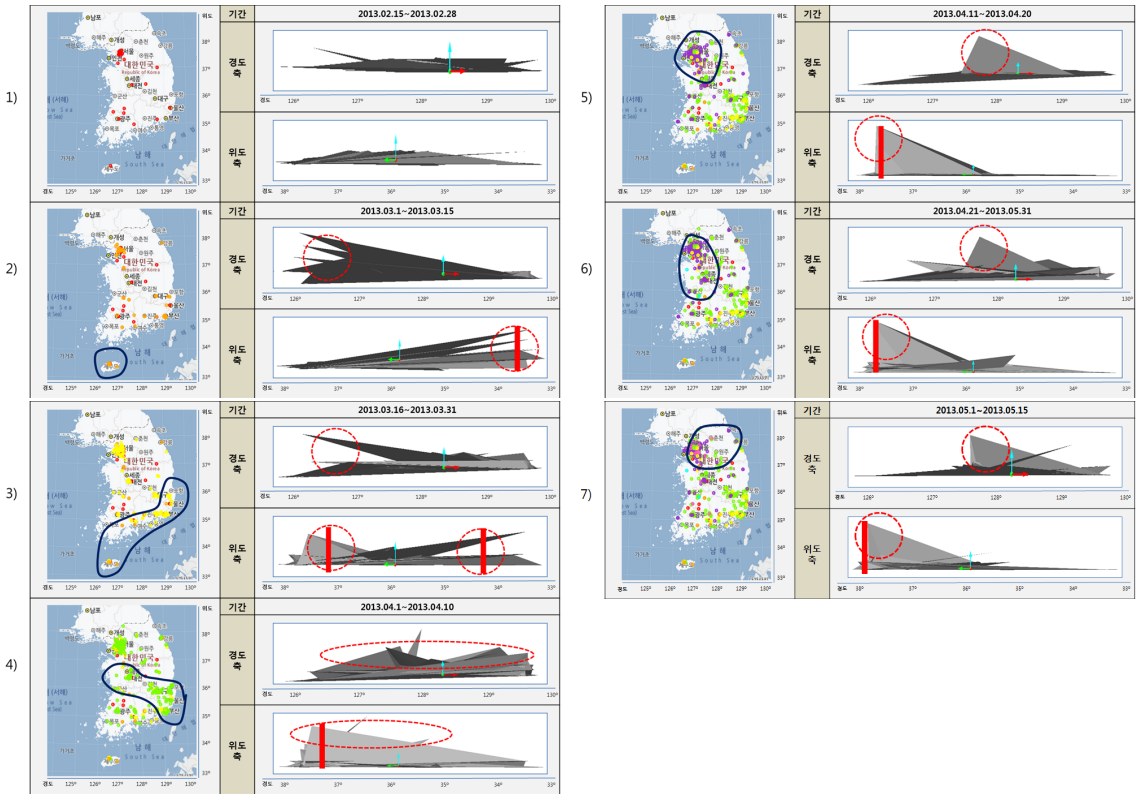
그림 6은 분석 대상으로 수집된 트윗 문서를 예시한 것으로, 실제 트윗 멘션(mention) 내에 시간 및 장소 정보가 명시적으로 표현된 경우도 있고, 문서에 내포된 메타정보(metadata)를 활용해 관련 정보를 추론할 수 있는 경우도 있다. 이벤트 정보 자동 추출 및 시각화 결과의 수작업 검증 및 평가를 위하여 수집된 문서집합 중에서 시간과 장소 정보가 모두 표현된 트윗을 대상으로 실험 집합을 한정하였다. 표 5는 [사례 연구 2]의 실험 대상 집합 트윗 건수를 기간별로 정리한 표이다.

그림 7은 수집된 트윗에서 분석된 '벚꽃' 관련 이벤트 정보를 날짜 구간별로 지도에 경위와 함께 표현한 것이다. 지도 시각화 이외 경위 3차원 표현을 위해 공개 소프트웨어인 SiteTopo(http://www.sitetopo.com)를 사용하였다. 왼쪽 지도에 표현된 점(●)의 크기는 '벚꽃'에 대해 언급한 트윗 수를 의미하며, 오른쪽 경도 및 위도축의 3차원 도형의 높이로 표현되었다. 즉 경위 도표에 표현된 도형의 높이가 높을수록 '벚꽃'에 대한 사용자 언급이 높음을 의미하며, 최고 지점을 원(○)으로 하였다. 특히 벚꽃 추이 정점 위도를 막대그래프 (|)로 표시하였다.

그림 7의 1)은 2013년 2월 한 달 동안 트위터에서 '벚꽃' 이벤트 정보를 추출한 결과이다. 전국 각 지역에서 '벚꽃'에 대한 내용이 간헐적으로 언급됨을 알 수 있다. 그림 7의 2)는 2013년 3월 초의 트위터 분석 결과로, 위도 33도~34도, 경도 126도~128도 사이의 지역에서 많이 언급된 것을 알 수 있다. 이 지역은 제주도 지역으로 실제 벚꽃 개화시기와 축제 기간과 일치함을 알 수 있었다.

그림 7의 3)과 4)를 비교 분석하면, 3월 중순부터 말까지는 위도 34도~37도 사이에 관련 이벤트 정보가 많이 감지되었다가 4월 초에는 북위 35도~38도 사이에 다수 포진된 양상을 볼 수 있다. 이는 벚꽃의 개화가 제주도 지역부터 남부지방을 거쳐 중부지방으로 북상하고 있음을 시사한다. 또한 경도 축의 도형이 126도~130도 사이에 매우 비등한 높이를 나타냄을 알 수 있다. 특히 트위터에서 주로 언급되는 도시는 제주도, 여수, 부산, 진해 등 남해안 지역으로, 이 지역들은 실제 '축제'가 개최된 지역이거나 사용자가 많이 방문한 지역이었다.

그림 7의 5)~7)은 각각 4월 중순과 말, 5월 초순의 연속성 분석 결과로, 위도 37도~38도 사이, 경도 126도~128도 사이, 즉 중부지방과 수도권 지방을 중심으로 '벚꽃'



(그림 7) 트위터에서의 '벚꽃' 이벤트 시각화 결과
(Figure 7) Visualization Result of 'cheery blossoms' Events in Twitter

에 대한 관심이 절정임을 나타낸다. 이후 실제 벚꽃 개화는 5월 말까지 진행되었으나 사용자 관심은 사그라진 것으로 분석되었다. 벚꽃 정점 표시 막대그래프의 이동 경로 역시 초기에 위도 33도~34도에서 시작하여 3월 말에는 34도~37도, 이후에는 37도~38도에 절정을 보인 것으로 분석된다. 그림 7의 결과를 통해 소셜 미디어에 기반한 분석 결과는 단순 개화시기에 의한 이벤트 발생 경로가 아닌 사용자의 관심과 언급이 높은 지역 중심의 이동 경로 분석이 가능함을 알 수 있다.

5. 결 론

본 논문에서는 소셜 미디어에 나타난 이벤트의 정보를 자동으로 추출, 이의 시공간적 시각화를 통해 시간의 흐름과 장소의 이동에 따른 이벤트 추이를 분석하였다.

소셜 미디어에 나타나 시공간적 연속성을 갖는 이벤트들의 특성을 정의하기 위해 2년간의 뉴스 기사 및 트윗 문서를 수집, 관련 도메인 선정을 위한 전수조사를 수행하였으며, 그 결과 연속성을 갖는 이벤트는 사회적으로 관심이 높은 어떤 이벤트가 발생한 후 일정 기간 동안 지리적 변화가 일어난다는 특성을 규명하였다. 특히 본 논문에서는 제안된 방법론의 활용이 높은 이벤트 선정을 위해 사회적 파급력, 전염성, 지속성이라는 기준을 적용하였다.

'구제역' 이벤트의 사례연구를 통해 질병 발생 근원지로부터 초기에는 인접지역으로 확산되다가 마지막에는 전국적으로 확산되는 전파 양상을 확인할 수 있었다. 또한 소셜 미디어간의 연관 관계를 분석하기 위해 뉴스미디어 결과와 트위터 기반 분석 결과를 비교 분석하였다. 뉴스 기반 분석결과와 비교하여 트위터 기반 결과는 최종 확산 지역 분포는 같은 양상을 보이지만 초기 발생 지역이 경북과 수도권이 동시에 나타난 것으로 분석되었다. 이는 실제 최초 발생지역이 '경북' 지역이였으나 사회적인 파장은 수도권에 발생한 이후 커진 것으로 분석

된다. 이와 같이 단순 사실기반의 연속성 분석 결과와 사용자의 관심도가 반영된 소셜 미디어에 기반한 연속성 분석 결과를 비교한 결과, 같은 시기의 이벤트들이라 하더라도 사회적으로 미치는 파장이 큰 장소 이동의 흐름을 파악할 수 있음을 보였다.

또한 본 논문에서 제안한 방법을 '여가/레저' 도메인으로 확장, '벚꽃'에 관한 사례연구를 통해 제안 방법론의 효용성과 소셜 미디어 기반 연속성 분석 결과의 유용성을 검증하였다. 연구 결과, 일반적으로 알려진 벚꽃 개화 진행방향 이외에도 사용자가 관심 있는 지역 중심의 축제가 어떻게 연계하여 개최되는지를 가늠할 수 있었다. 이는 해당 이벤트가 시간과 장소의 변화에 따라 사회적으로 미치는 영향과 이벤트의 변화 방향성을 가늠할 수 있는 중요한 요소가 될 수 있다는 점을 시사한다.

본 논문에서 제안하고 있는 방법론은 근본적으로 소셜 미디어의 텍스트를 분석하여 이벤트 정보를 자동으로 추출하는 이벤트 정보 추출 모듈의 성능에 영향을 받는다. 초기 이벤트 정보 분석에 오류가 포함되면 그 결과가 이후 이벤트 간의 연속성 분석 결과에도 미치게 된다. 따라서 향후 연구 방향으로는 보다 정확한 이벤트 정보 추출에 효과적인 어휘자질을 탐구하고자 한다. 또한 현재 연구는 이벤트의 '발생'과 '확장' 관점에서의 연속성 분석이 진행되었으나, 앞으로는 이벤트 발생 이후의 사용자 반응이나 대응 현상에 대한 연속성 분석으로 확장하고자 한다.

참 고 문 헌 (Reference)

[1] J-H. Do, J-Y. Sim, and J-S, Lee, " Analysis of Spread use of social media and media usage behavior change", Technical Report of Korea Press Foundation, 2010 http://download.kpf.or.kr/MediaPds/PJSFJXLZSKLOXOA.pdf

[2] Jinah Seol, "An Exploratory Study on Social Media Users" Ambivalent Attitude", Journal of Internet Computing and Services(JICS), 15(2), 87-94, 2014

[3] S-Y. Han, M-H. Choi, M-K, Park, and H-J. Lee, "The Effect of SNS Use on the Social Network: Moderating Effect of the Gender", Journal of Korean Society for Information Management, 30(3). 133-156, 2013.

[4] Y. Ho, and S. W. Lee, "Competition and Subsituation of Media: Niche Analysis of Mass Media and Social Media", Korean Journal of Journalism & Communication Studies. 56(4), 29-54, 2012.

[5] S. J Kim, "An Exploratory Study of Undergraduate Students' Health Information Needs and Seeking Behaviors in Social Media", Journal of Korean Biblia Society for Library and Information Science, 23(4), 239-260, 2012.

[6] J. Allan, "Topic detection and tracking: event-based information organization", Kluwer Academic Publishers Norwell (ISBN:0-7923-7664-1), 2002.

[7] R. Swan and J. Allan, "Automatic generation of overview timelines", Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. 49-56, 2000.

[8] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level", Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 314-321, 2003.

[9] Q. He, K. Chang, and E-P, Lim, "Analyzing feature trajectories for event detection", Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 207-214, 2007.

[10] P. Haghani, S. Michel and K. Aberer, "Tracking hot-k items over Web 2.0 streams", Proceedings of the 14. Database Systems for Business, Technology, and Web (BTW 2011), 105-122, 2011.

[11] H. Ma, "Hot topic extraction using Time Window", Proceedings of International Conference on Machine Learning and Cybernetics (ICMS-2011). 56-60, 2007.

[12] A. Mislove, S. Lehmann, et al., "Pulse of the nation: Us mood throughout the day inferred from twitter", Accessed November 22, 2010. http://www.ccs.neu.edu/home/amislove/twittermood/

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors", Proceedings of the 19th international conference on World wide web (WWW '10). 851-860, 2010.

[14] M. Imran, S. Elbassuoni, et al., “Practical extraction of disaster-relevant information from social media”, Proceedings of the 22nd international conference on World wide web (WWW ’13). 1021-1024, 2013.

[15] ©United State Tornades http://www.ustornadoes.com/

[16] H-J. Oh, Y-J. Bae, and B-H. Yun, “Temporal and Spatial Continuity Analysis of Issue Events”, Journal of KIIT. 11(9), 161-167, 2013

[17] H-J. Oh, H-K. Kim, and B-H. Yun, “Continuous Issue Event Analysis in Social Media”, The Journal of Korean association of computer education, 17(2), 31-38, 2014.

[18] B-H. Yun, “Comparison of Sentence-level Event Classification Algorithms”, Journal of KIIT. 10(6), 171-176, 2012.

[19] S. Liy, S. & H. Chang, “A Topic Detection and Tracking System with TF-Density”, Lecture Notes in Electrical Engineering, 156. 115-1203, 2013.

◐ 저 자 소 개 ◑

오 효 정 (Hyo-Jung Oh)
2000년 충남대학교 컴퓨터과학과 (이학석사)
2008년 한국과학기술원 컴퓨터공학과 (공학박사)
2000년 ~ 현재 한국전자통신연구원 지식마이닝연구실 책임연구원
관심분야 : 정보검색, 질의응답, 소셜마이닝, 인공지능, 빅데이터 정보처리, etc.
E-mail : ohj@etri.re.kr

윤 보 현 (Bo-Hyun Yun)
1999년 고려대학교 컴퓨터학과 (이학박사)
1999년 ~ 2002: 한국전자통신연구원 선임연구원 (팀장)
2003년 ~ 현재 : 목원대학교 컴퓨터교육과 교수
관심분야 : 자연어처리, 정보검색, 시맨틱웹, E-Learning, etc.
E-mail : ybh@mokwon.ac.kr

유 철 중 (Cheol-Jung Yoo)
1982년 전북대학교 전산통계학과(이학사)
1985년 전남대학교 대학원 계산통계학과(이학석사)
1994년 전북대학교 대학원 전산통계학과(이학박사)
2012년 1월~2013년 7월 University of California, Irvine(UCI) 국외연구교수
1997년 ~ 현재: 전북대학교 소프트웨어공학과 교수
관심분야 : 소프트웨어품질/메트릭스/테스팅, 임베디드 소프트웨어/테스팅, GIS, 교육공학, 인지과학 etc.
E-mail : cjyoo@jbnu.ac.kr

김 용 (Yong Kim)
1995년 University of North Texas (MS in Information Science)
2000년 충남대학교 컴퓨터과학과 (이학석사)
2006년 연세대학교 문헌정보학과 (문헌정보학 박사)
1996년 ~ 2008년: ㈜ KT 중앙연구소 책임연구원
2008년 ~ 현재: 전북대학교 문헌정보학과 부교수
관심분야 : 정보검색, 디지털도서관, 데이터마이닝, e-Learning, 전자기록물, 전자기록관리시스템 etc.
E-mail : yk9118@jbnu.ac.kr