KIPS Tr. Software and Data Eng.
Vol.3, No.9 pp.341~348 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.9.341

Unspecified Event Detection System Based on Contextual Location Name on Twitter

Pyeonghwa Oh[†] · Junyeob Yim[†] · Jinyoung Yoon[††] · Byung-Yeon Hwang[†††]

ABSTRACT

The advance in web accessibility with dissemination of smart phones gives rise to rapid increment of users on social network platforms. Many research projects are in progress to detect events using Twitter because it has a powerful influence on the dissemination of information with its open networks, and it is the representative service which generates more than 500 million Tweets a day in average; however, existing studies to detect events has been used TFIDF algorithm without any consideration of the various conditions of tweets. In addition, some of them detected predefined events. In this paper, we propose the RTFIDF·VT algorithm which is a modified algorithm of TFIDF by reflecting features of Twitter. We also verified the optimal section of TF and DF for detecting events through the experiment. Finally, we suggest a system that extracts result-sets of places and related keywords at the given specific time using the RTFIDF·VT algorithm and validated section of TF and DF.

Keywords : Twitter, SNS, Event Detect, TFIDF, RTFIDF·VT

트위터에서 문맥상 지역명을 기반으로 한 불특정 이벤트 탐지 시스템

오 평 화[†] · 임 준 엽[†] · 윤 진 영[††] · 황 병 연[†††]

요 약

스마트폰의 확산으로 인한 웹 접근성의 발달은 소셜 네트워크를 기반으로 하는 플랫폼 서비스 이용자의 급격한 증가를 이끌어냈다. 그 중에서도 개방적인 네트워크를 기반으로 빠른 확산과 강력한 영향력을 보이는 트위터(Twitter)는 하루 평균 5억 건이 넘는 트윗(Tweet)이 생산되는 대표적인 서비스이다. 따라서 트위터를 이용하여 이벤트를 탐지하려는 다양한 연구들이 진행되고 있다. 그러나 기존의 연구들은 이벤트 탐지를 위해 트윗을 구성하는 다양한 조건에 대한 고려 없이 일반 문서와 동일하게 일반적인 TFIDF 알고리즘을 적용하였다. 또한 TF와 DF에 대한 언급이 생략된 채, 사전에 지정한 키워드와 관련된 이벤트를 대상으로 탐지하였다. 이에 본 논문에서는 트위터의 특징을 반영한 TFIDF 변형 알고리즘인 RTFIDF·VT를 제안하고, 실험을 통해 이벤트 탐지에 최적인 것으로 검증된 TF와 DF 구간을 밝힌다. 최종 검증된 TF와 DF의 구간과 RTFIDF·VT를 적용하여 특정시점을 입력받아 이벤트로 예상되는 지역명들과 이벤트 관련 키워드의 결과 집합을 추출하는 시스템을 제안한다.

키워드 : 트위터, SNS, 이벤트 탐지, TFIDF, RTFIDF·VT

1. 서 론

최근 한국 갤럽이 전국의 만 19세 이상 남녀 천여 명을 대상으로 실시한 설문조사에 의하면, 2013년 11월 우리나라 성인 스마트폰 사용률은 74.1%인 것으로 나타났다[1]. 이는 조사가 시작된 2012년 1월을 기준으로 불과 2년 사이에 21%라는 급격한 증가세를 보이고 있음을 의미한다. 스마트폰 보급률의 확대로 웹 접근이 편리해지면서 웹 서비스, 소셜 네트워크 서비스(Social Network Service) 등 인터넷 기반의 응용 프로그램들이 점차 PC에서 모바일로 확대되기 시작했다. 그중에서도 스마트폰 사용자의 소셜 네트워크 서비스 사용시간은 하루 평균 13.1시간으로 다른 서비스 이용시간보다 월등히 많은 것으로 기록되었다[2].

※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단 기초연구사업(No. 2011-0009407)의 연구비 지원으로 수행되었음.
† 준 회 원 : 가톨릭대학교 컴퓨터공학과 석사과정
†† 준 회 원 : 가톨릭대학교 컴퓨터공학과 석사
††† 종신회원 : 가톨릭대학교 컴퓨터정보공학부 교수
Manuscript Received : March 10, 2014
First Revision : June 23, 2014; Second Revision : July 14, 2014
Accepted : July 14, 2014
* Corresponding Author : Byung-Yeon Hwang(byhwang@catholic.ac.kr)

트위터는 전 세계 하루 평균 약 1억 명의 사용자가 5억 건이 넘는 트윗을 주고받는 대표적인 소셜 네트워크 서비스 중 하나이다[3]. 140자로 제한된 단문 메시지인 트윗은 개인의 감정이나 일상생활, 사회적 이슈 등 다양한 내용으로 존재하며 일방적인 네트워크 연결과 리트윗(Retweet)이라는 독특한 기능을 통해 빠르게 전파된다. 트위터는 페이스북(Facebook)과 같이 사용자 간 합의에 의해 폐쇄적으로 형성되는 네트워크 기반의 서비스보다 높은 사회적 파급력과 영향력을 갖기 때문에 국내외로 많은 관련 연구가 진행되고 있다.

윤진영 등의 연구에서는 감기 증상에 관한 키워드 세트를 구성하여 수집한 트윗에서 감기 증상에 관한 정보를 추출하고 분석하여 트위터에서 사회적 신호가 탐지 가능함을 보였다[4]. 이들은 기상청에서 제공하는 감기 기상지수와 세 가지 기후 요소인 일교차, 최저기온, 상대습도 데이터의 비교분석을 통해 트윗에서 추출한 감기신호의 신뢰성을 확인했다. 또한 Takeshi 등의 연구에서는 특정 이벤트 키워드를 정해놓고 트위터 이용자 중 몇몇을 센서로 이용하여 그들이 작성한 트윗을 관찰하여 이벤트 탐지를 했다[5]. 이들은 지진과 관련된 “earthquake”, “shake” 등의 단어를 미리 지정해 놓았고 해당 단어가 급증하는 시기를 탐지하여 지진의 경로를 예측했다. 트윗은 온전히 사용자의 주관을 반영한다. 하지만, 많은 사람들의 동일한 주관은 사회적 합의를 이뤄낸다. 트위터를 분석하여 사회적 현상을 탐지할 수 있음을 밝힌 이들의 연구는 사용자로부터 주관적 사실을 이끌어 냈다는 데 의의가 있다. 그러나 이들은 특정 이벤트 키워드를 사전에 정의하고 해당 키워드를 포함하는 트윗을 통해 한정된 분석을 수행했다는 한계를 보인다. 이는 사전에 정의되지 않은 키워드나 의도하지 않게 발생하는 이벤트는 탐지할 수 없음을 의미한다.

본 논문에서는 불특정 이벤트의 예상 지역명과 키워드를 추출하는 방법을 제안하고 이를 탐지하는 시스템을 구현하였다. 여기서 불특정 이벤트란 입력 받은 날짜와 시간의 조건에 따라 추출되는 지역별 재해, 사건, 사고, 행사 등 사전에 정의되지 않은 포괄적인 의미의 이벤트를 의미한다. 이를 위해 지역명 필터링을 적용하고 TFIDF(Term Frequency Inverse Document Frequency) 알고리즘을 트윗 기반 분석에 적합하도록 개선한(Revised) RTFIDF・VT 알고리즘을 제안하였으며 이를 시스템에 적용했다. 또한 RTFIDF・VT 알고리즘을 적용할 때 실험을 통해 검증된 TF 구간과 DF 구간을 제안하여 반영했다.

본 논문의 구성은 다음과 같다. 2절에서 트위터를 이용한 이벤트 탐지 연구들을 살펴보고 3절에서 본 논문에서 제안하는 트위터를 이용한 이벤트 시스템의 구성도와 시스템에 적용된 RTFIDF・VT 알고리즘, 최적의 TF 구간, DF 구간을 제안하는 내용을 차례로 소개한다. 4절에서는 제안한 시스템을 이용하여 이벤트를 추출한 실험결과를 보이고 5절에서는 본 연구의 결론과 향후 연구를 밝힌다.

2. 관련 연구

소셜 네트워크 서비스가 전 세계적인 인기를 누리면서 사용자들이 생산하는 데이터에서 정보를 추출하고 이를 활용하기 위한 연구들이 진행되고 있다. 폐쇄형 네트워크 서비스를 대표하는 페이스북은 주로 마케팅에 활용할 목적으로 분석되어 왔는데 최근에는 사용자들의 의사표시인 ‘좋아요’를 분석하여 개인의 성향을 80% 이상 예측해 낸 연구도 있었다[6]. 페이스북과 달리 개방형 네트워크 서비스를 대표하는 트위터는 대부분의 트윗이 개방되어 있기 때문에 내용 분석을 통해 이벤트를 탐지하려는 단계까지 연구가 진행되었다.

Rui Li 등은 트위터를 기반으로 한 이벤트 탐지기법과 분석 시스템인 TEDAS(a Twitter-based Event Detection and Analysis System)를 제안했다[7]. TEDAS는 사전에 정의한 키워드를 이용하여 새로운 이벤트를 탐지하고 해당 이벤트의 공간과 시간의 패턴을 분석하여 이벤트에 대한 중요도를 식별한다. 이들은 범죄(Crime)와 재난(Disaster)에 대한 키워드들을 중심으로 2011년 7월 지역명 Kentucky와 Missouri에 발생했던 토네이도(tornados)와 관련한 트윗을 분석하였다. TEDAS의 주요 기능은 실시간으로 수집되는 트윗을 통해 새로운 CDEs(Crime Disaster Event)를 탐지하고 위치 정보인 지오태그(Geotag)를 추출하여 지도에 해당 지역명을 표시한다. 또한 ‘Car accident’와 같은 CDEs와 관련된 탐색 키워드, 지리(Geographic) 정보 범위와 시간을 조건으로 함께 입력하여 이벤트에 대한 시각화와 다각적인 분석이 가능하도록 했다. 하지만 범죄와 재난 두 가지 키워드에 제한되어 있다는 점은 의도하지 않은 이벤트에 대해서는 탐지가 불가능하다는 적용범위의 한계를 갖는다.

Meenakshi 등은 공간과 시간, 주제의 집합을 중심으로 사용자가 작성한 트윗을 분석하여 이벤트 안에 숨겨진 지역명이나 세계적 사회 인지에 접근했다[8]. 그들은 일반 사용자를 하나의 센서로 간주하고 이들이 생산하는 트윗에 TFIDF 알고리즘을 적용하여 이벤트와 관련된 키워드들을 추출하였다. 또한 공간, 시간 그리고 의미를 통합한 집합을 통해 지속적인 관측을 수행한 결과 이벤트 안에 숨겨진 시민 지각의 개요를 추출할 수 있음을 밝혀냈다. 실제로 2011년 발생한 뭄바이 테러 당시 뭄바이 테러에 관련된 트윗을 통해 파키스탄과 인도, 미국에서 발생한 키워드들을 탐지한 결과 ‘a.q. khan’, ‘pakistani’ 등의 키워드들을 추출하는 데 성공했다.

이를 위해 트위터로부터 이벤트와 연관된 트윗을 수집하고 처리하여 추출한 키워드 결과를 시각화 모델로 제공하는 시스템으로 Twitris를 제안했다. 그러나 분석에 앞서 대상 이벤트를 미리 정의해야 한다는 점은 미리 정의된 키워드를 통한 분석과 동일한 한계를 갖는다. 뿐만 아니라 국가 단위의 분석을 위한 설계로 인해 지역적 또는 국지적 이벤트의 경우는 고려되지 않았다는 지리적 범위의 한계를 보였다.

국내에서는 TFIDF 알고리즘에서 TF와 DF 구간을 임의로 조정하여 이벤트를 검출한 사례가 있다. 임준엽 등은 트위터에 TFIDF 알고리즘을 적용하여 2013년 3월 9일에 포항

에서 발생한 산불을 성공적으로 탐지하였다[9]. 실험은 수집한 트윗 코퍼스에 형태소 분석을 적용하여 시, 군, 구 행정구역명 등 지역명을 기준으로 필터링하였으며, 1시간 단위로 트윗을 클러스터링하여 TF의 경우 1시간, DF의 경우 5일로 조정된 TFIDF 알고리즘을 적용하였다. 이들은 지오태그 포함 확률이 낮은 국내 트위터 사용자들로부터 트윗 내의 지역명 추출해냈다는 점과 이벤트 탐지에 용이한 TFIDF 알고리즘을 적용했다는 점에서 의의가 있다. 그러나 짧은 단문으로 빠르게 확산되는 트위터에 TFIDF 알고리즘을 바로 적용하기엔 무리가 있다.

Rui Li 등의 연구와 Meenakshi 등의 사례들이 보이는 공통적인 문제점은 지정된 특정 키워드나 이벤트 이외의 대상을 탐지하도록 발전될 수 있는 가능성이 전혀 고려되지 않았다는 것이다. 임준엽 등의 연구에서는 이러한 한계를 고려하여 문맥에서 불특정 이벤트를 탐지하는 방법을 제안하였지만 140자의 짧은 문장에 문서유사도 알고리즘을 그대로 적용하기에는 무리가 따른다. 이에 본 논문에서는 트윗의 문맥에서 판단한 지역명를 기반으로 불특정 이벤트를 탐지하는 방법을 제안한다. 이를 위해 트위터의 특징을 고려한 RTFIDF·VT 알고리즘을 제안하고 실험을 통해 검증된 TF, DF구간을 시스템에 적용했다.

3. 시스템 제안

문서유사도 알고리즘으로 잘 알려진 TFIDF 알고리즘은 특정 단어를 포함한 문서의 검색에서 일반적으로 높은 성능을 보이지만 140자의 단문으로 제한된 트위터에는 적용이 쉽지 않아 특정 시간대를 구간으로 재구성한 트윗 코퍼스에 적용하는 것이 일반적이었다. 그러나 TFIDF 알고리즘을 사용한 대부분의 연구에서는 트위터의 특성을 고려하지 못한 채 해당 알고리즘을 그대로 적용하였을 뿐만 아니라 설정된 구간에 따라서 상이한 결과를 보이는 TF와 DF의 구간을 임의대로 결정하였다. 이에 본 논문에서는 실험을 통해 트위터의 특성에 맞는 보정 TF와 DF를 결정하고, 리트윗에 의해 중복되는 트윗을 제거하여 다양성을 보장하는 계수 VT를 적용한 RTFIDF·VT 알고리즘을 제안한다. 또한 이벤트 검출 순위가 다른 구간보다 높은 TF와 DF 구간을 도출하여 트위터의 특성이 충분히 고려될 수 있도록 설계했다.

실험 데이터는 2013년 9월부터 10월까지 네이버 뉴스를 분석하여 30개의 이벤트를 선정하고 해당 기간의 트윗을 수집하였다. 수집된 트윗은 형태소 분석과 지역명 필터링 후 RTFIDF·VT를 적용하여 '지역명+키워드' 형태로 정의할 수 있는 최종 13개의 이벤트를 대상으로 정의했다.

3.1 시스템 구성

지역명 이벤트 탐지 시스템은 트윗 수집, 형태소 분석, 지역명 필터링, 지역명 순위 도출, 각 지역명의 10개 키워드 추출 순서로 이루어지며 시스템 구성은 Fig. 1.과 같다. 제안하는 시스템은 일반 사용자(End-User)로부터 특정 날짜와 시간을 입력 값으로 받아 RTFIDF·VT를 적용한 후 이벤트가 발생했을 것으로 판단되는 지역명과 키워드를 1위부터 30위까지 추출하며, 각 지역명마다 많이 언급된 상위 10개의 키워드를 출력한다.

Fig. 1의 서버(Server)에서는 트위터 스트리밍(Streaming) API[10]를 이용하여 지속적으로 트윗을 수집한다. 이후 수집된 트윗은 한글 형태소분석을 통해 트윗에 포함된 명사들을 추출하여 코퍼스 형태로 저장되며, 명사로 이루어진 트윗 중 지역명이 포함된 트윗을 추출한다. 여기서 지역명은 시, 군, 구 단위의 행정구역명을 의미하며 정확도 향상을 위해 지하철이 있는 서울특별시와 경기도, 5대 광역시에 대해서는 역명을 포함했다. 본 논문에서 제안하는 시스템은 트윗에 포함될 비율이 낮은 지오태그를 이용하지 않고 트윗 문맥에서 지역명을 필터링함으로써 수집량을 증가시켰다.

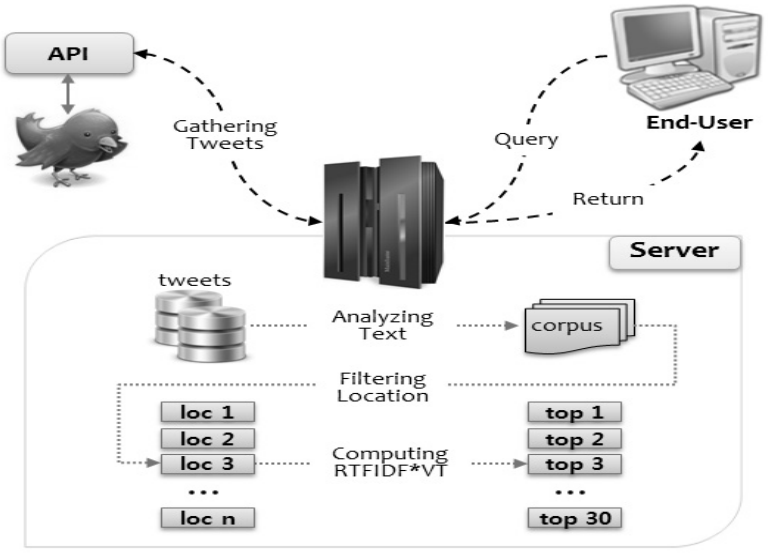


Fig. 1. System configuration

그러나 일단 수집된 트윗에 지오태그가 함께 포함되어 있다면 Yahoo!의 위치-주소 변환 API를 통해 얻은 주소로 문맥의 지역명을 대체하여 보다 정확한 위치정보로 보정하도록 설계하였다. 일반 사용자가 질의를 입력하면 조건 즉, 특정 날짜와 시간에 해당하는 지역명 이벤트를 RTFIDF・VT를 이용하여 도출하고 지역명의 경우 30위까지, 해당 지역명과 연관된 키워드를 언급이 자주 된 순서대로 10개까지 반환한다.

3.2 형태소 분석

수집된 트윗은 아파치 루씬(Apache Lucene) 한글 형태소 분석기[11]를 이용하였다. Table 1은 형태소 분석을 통해 명사로 분석되어 지역명 필터링을 통해 추출된 트윗의 예이다. '지역명+트윗 발생 시간+명사'로 구성된 집합세트는 각각의 지역명에 RTFIDF・VT가 적용될 때 사용되며 이를 통해 가장 많이 언급된 상위 10개의 키워드를 추출한다.

Table 1. Local name filtered tweet sample

지역명	트윗 발생 시간	명사 추출 트윗
서울	2014년 1월 1일 0시 3분	새해/욕/서울/보신각/종치는거/제/제대/안보여준다/채널/역시/예상했던대로/
제주	2014년 1월 1일 0시 4분	왠지/제주/사투리/얘기/복덜/받읍써/되서/반말/보/보이기/하/
김포	2014년 1월 1일 0시 9분	우선/11시/근처/김포/김포로/가/방향/으/으로/할거/한데/이외/이/인천/인천에/때/있/
영덕	2014년 1월 1일 0시 11분	여기/영덕/해돋이/제야/제야의/종소리/개뿔/언제/어디서/어디서든/트위터/트위터를/하/트잉여/트잉여일/경건/

3.3 RTFIDF・VT

기존의 TFIDF 알고리즘[12]은 Table 2와 같다.

Table 2. TFIDF algorithm

TF	$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: 문서 dj에 출현한 단어 ti의 수 $\sum_k n_{k,j}$: 문서 dj에 출현한 모든 단어의 수
IDF	$idf_i = \log \frac{\lvert D\rvert}{\lvert d_j \rvert t_j \in d_j\rvert}$ $\lvert D\rvert$: 문서집합에 포함되어 있는 문서의 수 $\lvert d_j \rvert t_j \in d_j\rvert$: 단어 t_j가 등장하는 문서의 수
TF-IDF	$TFIDF_{i,j} = tf_{i,j} \times idf_i$

Table 3의 RTFIDF・VT 알고리즘은 TFIDF 알고리즘을 구성하는 TF와 DF를 트윗의 특성에 맞게 보정한 RTFIDF에 트윗의 다양성을 고려하도록 고안된 보정계수인 VT(Variety of Tweets)를 반영하여 구성된다. 보정된 TF인 RTF는 평소 해당 지역명이 트윗에 반영되는 빈도를 나타낸다. 즉, 같은 양의 트윗이 발생하여도 평소에 발생되는 트윗의 양에 따라 RTFIDF・VT 값이 보정될 수 있도록 하였다. 보정된 DF인 RIDF는 DF가 log를 취하지 않도록 보정하였는데, 짧은 시간 동안 수집된 트윗은 log를 적용할 만큼 큰 데이터가 아니다. 따라서 DF 값의 차이를 더 명확히 하기 위해 log를 취하지 않았다. VT는 트윗의 다양한 정도를 의미한다. 이는 동일한 내용의 리트윗된 여러 개의 트윗을 하나의 트윗으로 취급한다. 즉, 최대한 리트윗을 지양하고 사용자가 직접 작성한 트윗의 수가 잘 반영되도록 유도했다. 이 밖에도 보정계수 VT는 평소 발생하는 트윗의 양에 따라 지역명별로 RTFIDF 값을 조정하는 역할도 수행한다. 평소 트윗이 빈발하는 지역명은 DF 값이 크기 때문에 이벤트가 발생했더라도 높은 RTFIDF 값을 갖지 못하기 때문에 이벤트가 발생하더라도 30위 이내의 순위를 차지하기 힘들다. 이와 반대의 경우 평소 트윗이 빈발하지 않는 지역명은 낮은 DF 값을 갖기 때문에 높은 RTFIDF 값을 갖는다. 이러한 경우 VT는 두 지역명의 RTFIDF 값의 차이를 완화시킨다. 본 논문에서는 트위터의 특징이 반영된 RTFIDF・VT 알고리즘을 제안하고 시스템에 적용하여 언급된 지역명과 해당 지역명의 주요 키워드에 대한 순위를 도출했다.

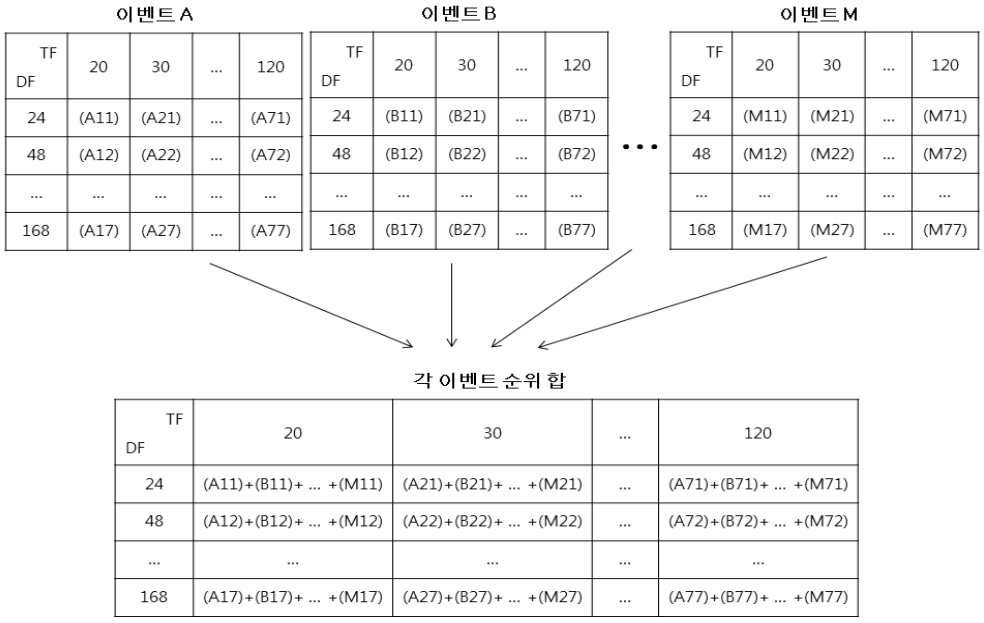
Table 3. RTFIDF・VT algorithm

TF를 보정한 RTF	$rtf_{i,j} = \frac{n_{i,j}}{\sum_j n_{i,j}}$ $n_{i,j}$: 문서 dj에 출현한 단어 ti의 수 $\sum_j n_{i,j}$: **모든** 문서에 출현한 단어 ti의 수
IDF를 보정한 RIDF	$ridf_i = \frac{\lvert D\rvert}{\lvert d_j \rvert t_j \in d_j\rvert}$ $\lvert D\rvert$: 문서집합에 포함되어 있는 문서의 수 $\lvert d_j \rvert t_j \in d_j\rvert$: 단어 t_j가 등장하는 문서의 수
보정 계수 VT (Variety of Tweets)	vt = 리트윗을 제외한 사용자가 직접 작성한 트윗의 수
RTFIDF * VT	$RTFIDF \bullet VT = RTF \bullet RIDF \bullet VT = rtf_{i,j} \times ridf_i \times vt$

3.4 최적의 TF 및 DF의 구간 제안

TFIDF 알고리즘을 적용한 기존의 연구에서는 TF와 DF의 구간 지정에 대한 언급이 생략되었다. 하지만 본 논문에서 RTFIDF・VT를 적용하는 과정에서 TF와 DF의 구간에 따라 값의 변화가 있음을 발견했다. 이에 최적의 TF구간과 DF구간을 제안하는 실험을 추가로 진행했다.

해당 실험을 위해 2013년 9월 1일부터 10월 31일까지의 네이버 뉴스 검색 API[13]를 적용하여 30개의 이벤트를 선정했다. 같은 기간 동안의 트윗을 수집하여 선정한 30개의 이벤트에 RTFIDF・VT를 적용하여 트위터에서도 이벤트로 분류될 수 있는지 알아보았다. 이벤트 탐지를 위해 각 이벤트가 발생한 시간을 기준으로 해당 시간에 발생한 이벤트의 지역명이 30위 이내에 포함되는지를 판별하였다. 이렇게 판별된 이벤트를 통해 TF와 DF 구간을 제안하는 실험을 진행했다. TF와 DF 구간을 제안하기 위한 실험 과정은 다음과 같다.



이벤트 A

TF / DF	20	30	...	120
24	(A11)	(A21)	...	(A71)
48	(A12)	(A22)	...	(A72)
...	...	...	...	...
168	(A17)	(A27)	...	(A77)

이벤트 B

TF / DF	20	30	...	120
24	(B11)	(B21)	...	(B71)
48	(B12)	(B22)	...	(B72)
...	...	...	...	...
168	(B17)	(B27)	...	(B77)

이벤트 M

TF / DF	20	30	...	120
24	(M11)	(M21)	...	(M71)
48	(M12)	(M22)	...	(M72)
...	...	...	...	...
168	(M17)	(M27)	...	(M77)

각 이벤트 순위 합

TF / DF	20	30	...	120
24	(A11)+(B11)+ ... +(M11)	(A21)+(B21)+ ... +(M21)	...	(A71)+(B71)+ ... +(M71)
48	(A12)+(B12)+ ... +(M12)	(A22)+(B22)+ ... +(M22)	...	(A72)+(B72)+ ... +(M72)
...	...	...	...	...
168	(A17)+(B17)+ ... +(M17)	(A27)+(B27)+ ... +(M27)	...	(A77)+(B77)+ ... +(M77)

Fig. 2. Processing of selecting optimal section about TF and DF

① 2013년 9월부터 10월까지 발생한 이벤트이면서 발생과 동시에 트위터에 올라온 13개의 '지역명+이벤트' 집합을 추출한다.
② '지역명+이벤트' 집합 각각의 TF와 DF 구간을 변경하면서 RTFIDF·VT를 적용하여 해당 이벤트의 순위 테이블을 완성한다. 단, 순위가 30위 안에 포함되지 않으면 해당 구간에 대한 순위는 31위로 작성한다.
③ TF는 20, 40, 60, 80, 100, 120분의 총 6개 구간으로 구성한다.
④ DF는 24, 48, 72, 96, 120, 144, 168 시간의 총 7개 구간으로 구성한다.
⑤ 각 이벤트에 대한 13개의 순위테이블이 완성되면 Table 4와 같이 TF 6개 구간과 DF 7개 구간을 조합한 총 42개 구간에 대한 실험결과인 순위테이블을 작성한다. Table 4는 2013년 9월 5일 양주에서 일어난 마트 화재에 관한 TF와 DF의 구간별 순위이다.
⑥ 13개의 이벤트 순위테이블에서 같은 TF와 DF 구간에 속하는 순위를 더하여 하나의 테이블을 만들고 각 이벤트 순위를 더한 테이블 중 최소값이 나온 구간을 도출한다.

Table 4. Example of ranking by region of TF and DF

TF(분) / DF(시간)	20	40	60	80	100	120
24	1	2	6	10	31	16
48	1	1	4	8	31	14
72	1	1	5	6	31	12
96	1	2	5	8	31	14
120	1	2	5	7	13	15
144	1	2	6	10	31	16
168	1	2	6	10	31	16

실험 과정의 전체적인 구조는 Fig. 2와 같다. 13개의 이벤트에 대한 순위테이블이 작성되고 각각의 TF와 DF 구간에 대한 순위를 합하여 하나의 테이블을 작성하였다.

4. 실험 결과

실험에 사용된 환경은 Table 5와 같다.

Table 5. Experimental environment

운영체제	MS Windows 7
CPU	Intel Core2Quad Q8400
RAM	4GB
개발언어	Java 1.7.0
기타	Twitter API 1.1

앞서 언급한 바와 같이 본 논문에서는 최적의 TF와 DF 구간을 제안하기 위한 실험과 제안하는 시스템을 이용하여 실제 이벤트를 탐지하기 위한 실험을 진행했다. 실험 결과는 다음과 같다.

4.1 최적의 TF, DF 구간 제안 결과

실험을 통한 TF와 DF 구간의 최적값은 각각 40분과 48시간이며 결과는 Table 6과 같다. 여기서 TF 구간은 대체로 40분, 60분일 때 순위의 합이 낮아 나머지 구간보다 이벤트 검출에 대체적으로 뛰어난 결과를 보였다. 이는 리트윗에 의한 정보 전파 시간과 관련한다. Kwak 등의 연구에서는 어떠한 사건에 대해 쓰인 리트윗의 50%는 원래 트윗이 작성된 시간으로부터 1시간 이내에 리트윗되고 또한 해당 사건에 대한 75%의 리트윗은 24시간 이내에 확산됨을 밝혔

다[14]. 비록 20분의 경우 표본이 너무 적기 때문에 가장 좋지 않은 결과를 보였지만, 결과적으로 이벤트를 검출하는 실질적인 구간인 TF는 특정 이벤트에 의해 발생한 리트윗의 50%를 차지하는 1시간 이내로 정하는 것이 합당하다고 할 수 있다. 이에 본 실험에서는 순위의 최소값이 나온 40분을 최적의 TF 구간으로 결정했다.

DF 구간은 24시간을 제외하고는 DF 구간이 증가할수록 순위의 합이 높아진다. DF 구간이 증가하면 같은 RTF 값의 분모의 증가량도 매우 커지기 때문에 RTF 값은 감소시키고 RDF 값은 증가시켜 이벤트 검출에 대한 각 지역명의 RTFIDF·VT 값의 차이를 줄인다. 결과적으로 DF 구간은 작을수록 좋다고 판단되므로 실험 결과에 따라 순위의 합이 가장 낮은 48시간을 최적의 DF 구간으로 결정했다. 향후 트위터 관련 연구에서 문서 빈발도에 관한 공식을 이용할 경우 본 실험을 통해 최종 결정된 TF와 DF구간의 길이를 적용한다면 트윗을 통한 다양한 이벤트에 대한 탐지가 가능할 것으로 기대한다.

Table 6. The completed table by adding each event ranking

TF(분) / DF(시간)	20	40	60	80	100	120
24	247	129	149	220	208	184
48	232	120	145	174	211	179
72	243	125	152	168	201	185
96	232	131	157	182	250	190
120	232	129	153	190	170	189
144	232	137	150	188	202	193
168	231	136	150	189	217	189

4.2 본 시스템을 통한 이벤트 탐지 실험 결과

제안하는 시스템은 '지역명+10개 키워드' 집합을 30위까지 이벤트로 탐지해내는 것을 목표로 한다. 특정 날짜와 특정 시간에 대한 이벤트를 추출하고자 할 때 연, 월, 일, 시, 분을 '201312250000'와 같이 형태로 입력하면 해당하는 순간에 발생한 이벤트의 지역명과 키워드에 대한 결과를 볼 수 있다. 시스템의 성능을 검증하기 위해 실험 데이터는 이벤트가 급증할 것으로 예상되는 2014년 1월 1일을 포함, 2013년 12월 한 달 동안 수집한 트윗을 사용했다. 본 실험은 제안하는 시스템의 입력 값을 2013년 12월 1일 자정부터 10분 단위로 변경하며 입력했다. 또한 추출된 30위 내의 이벤트들 중 실제로 발생했을 것으로 예상되는 몇 개의 지역명과 키워드를 선정하여 뉴스와 지역정보를 통해 진위를 판명하였다. 실험 결과는 Table 7과 같다.

Table 7은 시스템을 통해 추출된 이벤트 결과를 보여주고 본 시스템을 통해 탐지된 시간 순으로 정렬되었다. 네이버 뉴스를 통해 각 이벤트의 실제 이벤트 발생 시간을 참고하여 시스템이 탐지한 이벤트 시간과 비교하였다. 부산 남북항 대교붕괴는 실제 사건이 발생한지 1시간 25분 후에 탐지되었으며 서울역 앞 고가도로에서 발생한 분신 관련 사건은 4시간여 후에 탐지되었다. 여기서 주목할 것은 분신 관련 사건의 경우 사건 당시에는 언론의 보도가 많지 않았음에도 탐지가 가능했다는 점이다. 또한 일정 기간 동안 발생하는 지역명의 축제도 탐지되어 남원 눈꽃축제나 양평 빙어축제 등도 순위에 나타났다. 물론 크리스마스나 새해 그리고 이와 관련된 보신각 등 전국단위의 이벤트도 서울, 종로, 인천, 부산 등에서 탐지되었다. 이러한 결과는 기존의 한정된 주제나 키워드에 국한된 이벤트 탐지가 아닌 불특정 이벤트를 추출할 수 있음을 보여준다.

5. 결론 및 향후 연구

소셜 네트워크 서비스가 인기를 얻어가면서 이를 기반으로 의미 있는 정보를 탐지해내고자 하는 다양한 시도들이 이루어졌다. 그러나 기존의 연구들은 특정 키워드나 이벤트 기반의 탐지를 수행하여 미리 정의하지 않은 이벤트를 탐지하기에는 한계가 있었다. 이에 본 논문에서는 트위터를 이용하여 사전에 정의되지 않은 국지적인 이벤트를 탐지하는 방법을 제한하였다. 본 연구는 지오태그를 사용하는 대신 트윗

Table 7. The result of detecting event from suggested system

순번	지역명	예상 이벤트 키워드	실제 이벤트 발생 시간	시스템 탐지 시간	순위
1	부산	남북항 대교 붕괴	2013년 12월 19일 16시 15분	2013년 12월 19일 17시 40분	14위
2	대구	크리스마스	2013년 12월 25일	2013년 12월 25일 0시 0분	11위
3	남원	눈꽃축제	2013년 12월 24일~2014년 02월 09일	2013년 12월 25일 12시 50분	1위
4	대전	크리스마스	2013년 12월 25일	2013년 12월 25일 12시 50분	13위
5	부산	크리스마스	2013년 12월 25일	2013년 12월 25일 13시 30분	13위
6	양평	빙어축제	2013년 12월 27일~2014년 02월 16일	2013년 12월 31일 17시 20분	13위
7	서울	서울역 고가도로 분신자살	2013년 12월 31일 17시 35분	2013년 12월 31일 21시 20분	9위
8	종로	새해	2014년 01월 01일 00시 00분	2014년 01월 01일 0시 0분	23위
9	인천	새해	2014년 01월 01일 00시 00분	2014년 01월 01일 0시 0분	30위
10	서울	보신각	2014년 01월 01일 00시 00분	2014년 01월 01일 0시 0분	8위

자체의 문맥에서 지역명을 탐지한다는 점과 트위터 특성에 맞는 RTFIDF·VT 알고리즘을 사용하고 실험을 통해 유연성을 보장하는 TF 및 DF 구간을 제안했다는 점에서 트위터를 이용한 이벤트 탐지 분야의 연구에 큰 의의가 있다.

향후 연구에서는 '연기', '동안' 등 지역명인 동시에 일상적인 단어로 많이 쓰이는 단어에 대해 보다 명확한 분류가 가능하도록 노이즈 제거 방법이 보완되어야 할 것이다. 제안하는 시스템이 비록 특정 시간뿐만 아니라 현재를 기준으로 가까운 시간에 발생한 이벤트까지 탐지할 수 있다고 하더라도 완벽한 실시간 탐지를 위해서는 해결해야 할 과제가 많다. 향후에는 30위까지 제한된 예상지역명이 아닌, 전국을 대상으로 모니터링 한 예상 지역명을 탐지해낼 수 있어야 할 것이다. 또한 '크리스마스'나 '설날' 등 특정 지역명에 국한되지 않은 키워드 자체가 이벤트가 되는 사례에 대한 추가적인 연구도 진행되어야 할 것이다.

References

[1] S. Cho, Halla Ilbo [Online]. Available : http://www.ihalla.com/read.php3?aid=1386687600449233044, Dec., 11, 2013.

[2] Y. Lee, Digital Daily [Online]. Available : http://www.ddaily.co.kr/news/article.html?no=112017, Dec., 12, 2013.

[3] B. Jung, Bloter [Online]. Availble : http://www.bloter.net/archives/166065, Oct., 4, 2013.

[4] J. Yoon, 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 Korea Multimedia Society, Vol.16, No.2, pp.667-677, 2013.

[5] T. Sakaki, M. Okzaki, and Y. Matsuo, "Earthquake Shakes Twitter Users : Real Time Event Detection by Social Sensors," Proc. of the 19th Conference on World Wide Web, pp.851-860, 2010.

[6] H. Choi, The Kyunghyang Shinmun [Online]. Available : http://news.khan.co.kr/kh_news/khan_art_view.html?artid=201303122154175&code=930401, Mar., 12, 2013

[7] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "TEDAS : A Twitter-based Event Detection and Analysis System," Proc. of the IEEE 28th International Conference on Data Engineering, pp.1273-1276, 2012.

[8] M. Nagarajan, K. Gomadam, A. P. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav, "Spatio Temporal Thematic Analysis of Citizen Sensor Data : Challenges and Experiences," Proc. of the 10th International Conference on Web Information Systems Engineering, pp.539-553, 2009.

[9] J. Yim, S. Kim, J. Yoon, P. Oh, B. Lee, and B.-Y. Hwang, "Detecting Local Events Using Twitter", Proc. of the 40th Korea Computer Congress, pp.248-250, 2013.

[10] Twitter, The Streaming APIs|Twitter Developers [Online]. Available : https://dev.twitter.com/docs/streaming-apis, Sep., 24, 2012.

[11] Apache Lucene Korean Analyzer and dictionary, Source Forge [Online] Available : http://sourceforge.net/projects/lucenekorean, Oct., 26, 2013.

[12] S. Lee, H. Kim, "Keyword Extraction from News Corpus using Modified TF-IDF," Journal of Society for e-Business Studies, Vol.14, No.4, pp.59-73, 2009.

[13] http://developer.naver.com/wiki/pages/News

[14] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a New Media?," Proc. of the 19th international conference on World Wide Web, pp.591-600, 2010.

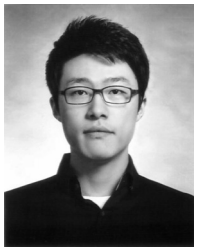
오 평 화

e-mail : oph312@nate.com

2013년 가톨릭대학교 컴퓨터공학과(학사)

2013년~현 재 가톨릭대학교 컴퓨터공학과 석사과정

관심분야 : 소셜네트워크분석, 데이터마이닝, XML, 데이터베이스



임 준 엽

e-mail : junyeob1205@naver.com

2013년 가톨릭대학교 컴퓨터공학과(학사)

2013년~현 재 가톨릭대학교 컴퓨터공학과 석사과정

관심분야 : 소셜네트워크분석, 데이터베이스, 데이터마이닝, 정보검색

윤 진 영

e-mail : my_sk_test@naver.com

2011년 가톨릭대학교 컴퓨터공학과(학사)

2014년 가톨릭대학교 컴퓨터공학과 석사

관심분야 : 소셜네트워크분석, 오피니언마이닝, 데이터마이닝, 정보검색

황 병 연

e-mail : byhwang@catholic.ac.kr

1986년 서울대학교 컴퓨터공학과(학사)

1989년 KAIST 전산학과(석사)

1994년 KAIST 전산학과(박사)

1994년~현 재 가톨릭대학교 컴퓨터정보공학부 교수

1999년~2000년 (美) 미네소타대학교 방문교수

2007년~2008년 (美) 캘리포니아주립대학교 방문교수

관심분야 : 소셜네트워크분석, XML 데이터베이스, 정보검색, 데이터마이닝, 지리정보시스템