RTFIDF·VT: 트윗의 다양성을 고려한 새로운 TF-IDF 알고리즘

오평화, 김석중, 윤진영, 임준엽, 황병연
가톨릭대학교 컴퓨터공학과
e-mail:oph312@catholic.ac.kr

RTFIDF·VT: a New TF-IDF Algorithm considered Variety of Tweets

Pyeonghwa Oh, Seokjung Kim, Jinyoung Yoon, Junyeob Yim, Byung-Yeon Hwang
Dept of Computer Science and Engineering, The Catholic University of Korea

요 약

스마트 폰의 보급으로 웹 접근성이 향상되면서 모바일을 기반으로 성장한 소셜 네트워크 서비스들은 폭발적인 사용자 증가를 이루었다. 그중에서도 트위터는 개방적인 사용자간 네트워크 연결 방식과 강력한 전파능력으로 사용자 개개인이 정보를 생산하고 소비하는 소셜 저널리즘의 형태를 띠며 영향력을 더해가고 있다. 이에 트위터를 이용해 이벤트를 탐지하고자 하는 연구들이 활발히 진행되고 있다. 그러나 이벤트를 탐지할 때 기존의 *TF-IDF* 알고리즘을 적용할 경우 트위터의 특징을 적절히 반영하지 못하는 문제점이 있다. 본 논문에서는 기존의 *TF-IDF* 알고리즘에 트위터의 특징을 반영하도록 가중치를 변형하고 여기에 다시 보정계수를 적용하여 새로운 *TF-IDF* 알고리즘을 제안하였으며 두 번의 이벤트에 적용한 실험을 통해 새로운 알고리즘의 성능향상을 보였다.

1. 서론

트위터(Twitter)는 140자의 단문으로 이루어진 마이크로 블로그로 페이스북과 더불어 대표적인 소셜 네트워크 서비스 중 하나다. 소셜 네트워크 서비스는 스마트 폰의 등장으로 인한 모바일에서의 웹 접근성 향상과 함께 폭발적인 사용자 증가를 이루었다. 트위터는 일방적으로 형성될 수 있는 독특한 네트워크 구조와 이를 기반으로 한 특유의 전파력으로 그 영향력을 더해가고 있으며 트위터 분석을 통한 효과적인 정보검색의 요구가 해마다 높아지고 있다.

최근에는 트위터를 이용하여 실시간으로 이벤트를 탐지하려는 연구들이 활발히 진행되고 있다. 이들 대부분은 간단하면서도 높은 성능을 보이는 *TF-IDF* 알고리즘을 적용하고 있다. 그러나 본래 *TF-IDF* 알고리즘은 문서검색에 높은 성능을 보이는 알고리즘으로, 140글자의 제한과 높은 파급력 등 트위터의 특징은 전혀 고려되지 않았다. 이에 본 논문에서는 트위터를 이용한 실시간 이벤트 탐지를 수행함에 있어 *TF-IDF* 알고리즘을 그대로 적용했을 때 발생하는 문제점을 파악하기 위해 최근 발생한 두 건의 이벤트에 대한 탐지를 시도했다. 그리고 가중치를 보정한 변형 *TF-IDF*와 여기에 특수변수를 추가해 개선된 *TF-IDF* 알고리즘을 제안한다.

논문의 구성은 다음과 같다. 2장에서 *TF-IDF* 알고리즘을 이용하여 이벤트를 탐지했던 연구들을 살펴보고 3장에서 *TF-IDF*와 가중치를 수정한 변형 *TF-IDF*, 보정계수를 추가한 개선된 *TF-IDF* 알고리즘을 차례로 소개한다. 4장에서는 3장에서 언급한 알고리즘들의 실험결과를 분석하고 마지막 5장에서 본 연구의 결론과 향후 연구계획을 밝힌다.

※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2011-0009407).

2. 관련연구

시간과 공간의 제한 없이 어디서나 인터넷에 접속할 수 있는 스마트폰 사용자의 증대는 소셜 네트워크 서비스의 대중화를 이끌어 냈다. 특히 트위터는 위치정보를 포함할 수 있는 140글자의 단문 텍스트를 팔로어(Follower)라는 특수한 관계의 사용자들에게 일괄 전달함으로서 정보전파의 파급력이 높다. 이러한 트위터의 특징을 이용해서 실시간으로 발생하는 이벤트를 탐지하고자 하는 연구들이 진행되어 왔다.

Meenakshi 등은 사용자가 남긴 글인 트윗(Tweet)을 분석할 때 공간과 시간, 주제의 집합을 통해 이벤트 이면에 숨겨진 지역적 혹은 세계적 사회 인지에 쉽게 접근하려 하였다[1]. 그들은 일반 대중들을 센서로 삼고 이들을 관측함으로서 공간, 시간 그리고 의미를 통합한 집합을 통해 이벤트 이면에 존재하는 시민 지각의 개요를 추출할 수 있다는 것을 알아냈다. 예를 들어, 2011년 발생한 뭄바이 테러 당시 해당 이벤트와 관련해서 파키스탄과 인도, 미국에서 발생한 키워드들을 탐지하여 비교 분석하기도 하였다. 하지만 Twitris라는 시스템을 통해 시각화 모델을

제시했음에도 불구하고 분석 전에 대상 이벤트를 미리 정의해야한다는 점과 국가 단위의 분석을 위한 설계로 지역적 또는 국지적 이벤트에 대해서는 언급하지 못한 점에서 한계를 갖는다.

국내에서도 트위터를 이용한 이벤트 탐지를 시도한 사례가 있다. 임준엽 등은 트위터에서 발생하는 트윗을 스트리밍으로 수집하여 국내에서 발생했을 것으로 추정되는 트윗들을 분류하여 형태소 분석과 *TF-IDF* 알고리즘을 통해 미리 정의되지 않은 이벤트를 탐지하고자 하였다[2]. 이 연구는 GPS 좌표를 포함한 비율이 낮은 트윗의 한계를 극복하기 위해 텍스트에서 지역정보를 추출하였다. 이를 위해 행정구역상의 지명을 포함한 트윗들을 선별하여 추출한 후 매 시간마다 추출한 트윗들을 1시간 단위로 클러스터링을 했다. 특정 시간대에서 문서들이 포함하고 있는 지역명과 해당 문서들 간의 유사도 값으로 이벤트가 발생했다고 추정되는 지역을 탐지한 후 해당 지명을 포함한 트윗에서 많이 언급된 단어들을 선별하여 “지역 + 키워드” 쌍의 이벤트를 추출한 결과 2013년 3월 9일 포항시에서 발생한 산불을 탐지해내기도 하였다.

TF-IDF 알고리즘을 사용한 다른 사례도 있다. 이성직 등은 *TF-IDF* 알고리즘의 변형을 통해 전자뉴스에 존재하는 키워드들 추출하였다[3]. 그러나 여기서 사용된 *TF-IDF*는 기존의 텍스트 마이닝 기법으로 비록 정보검색이라는 공통의 목표라 하더라도 트위터에 바로 적용하기에는 무리가 따른다.

두 연구는 이벤트 탐지를 위해 *TF-IDF* 알고리즘을 적극적으로 이용했다는 점에서 공통점을 찾을 수 있다. 그러나 일방적인 네트워크 연결 방식과 더불어 리트윗(Retweet, RT) 기능을 통해 원본 트윗이 전파되는 속도와 범위가 뛰어난 트위터에 일반 문서나 웹 문서에 적용되는 *TF-IDF* 알고리즘을 그대로 적용하는 것은 무리가 있다. 이에 본 논문에서는 트위터의 특징을 기존 *TF-IDF* 알고리즘에 반영하기 위해 가중치를 변형한 *TF-IDF* 알고리즘과 여기에 다시 보정계수를 적용한 알고리즘을 제안하였다.

3. 새로운 *TF-IDF* 알고리즘

기존 *TF-IDF* 알고리즘을 트위터에 그대로 적용할 경우에 발생할 수 있는 문제는 다음과 같다.

<표 1> 평상시 트윗 발생이 활발한 지역의 예

TF구간	1	2	3	4	5	6	7	8	9	10
트윗개수	100	100	100	100	100	100	100	100	100	120

<표 2> 이벤트가 발생할 때만 트윗이 증가하는 지역의 예

TF구간	1	2	3	4	5	6	7	8	9	10
트윗개수	1	1	2	3	1	1	2	2	1	20

표 1은 항상 많은 양의 트윗이 발생하는 지역으로, 평소 지역이 언급된 트윗 발생 수 100개에서 구간 10일 때 120개의 트윗이 발생했다. 표 2는 평소 미비한 양의 트윗이 발생하는 지역으로 평소 지역이 언급된 트윗 발생 수 1~3개에서 구간 10일 때 20 개의 트윗이 발생했다. 두 지역 모두 평소보다 20회나 더 많이 언급되었기 때문에 표 1, 2 두 경우 모두 이벤트라고 추측할 수 있다. 하지만 두 지역의 IDF 값은 같고 10 구간에서의 TF값은 표 1 지역이 높으므로 TF 값이 낮은 표 2 지역은 표 1 지역보다 낮은 감지 성능을 보인다. 이렇듯 기존의 *TF-IDF* 알고리즘은 평소 발생하는 트윗의 양을 반영하지 못하는 문제점이 있다.

트위터에 적합한 형태의 *TF-IDF* 알고리즘으로 변형하기 위한 실험 데이터는 다음과 같이 구성하였다. Apache 재단의 Lucene 한글 형태소분석기[4]로 트위터 Streaming API[5]를 이용해 수집한 트윗에서 명사를 추출한 후 시, 군, 구 단위로 지역명이 포함된 트윗에 *TF-IDF*를 적용하여 가장 빈발하게 발생하는 상위 30개 지역을 도출했다. 또한 지역별로 가장 많이 언급된 상위 10개의 키워드를 나타냈다. 실험은 최근 발생한 두 가지의 이벤트를 대상으로

1) 기존 *TF-IDF*
2) *TF*에 평소 트윗 수를 적용한 *TF* 보정 *TF-IDF*
3) 2)의 결과에 *DF*의 log를 제외시킨 *IDF* 보정 *TF-IDF*
4) 3)의 결과에 보정계수를 반영한 계수 적용 *TF-IDF*

의 4가지 방법을 적용하여 비교하였다. 실험 대상 이벤트는 표 3과 같다.

<표 3> 실험 이벤트

이벤트	내용
KTX 추돌사고	- 2013년 8월 31일 오전 7시 13분경 대구 발생 - 2013년 8월 29일 ~ 2013년 8월 31일 트윗 사용 - TF 30분, DF 48시간 단위로 실험 - 실험 구간은 오전 7시 30분부터 오후 12시까지 30분 간격으로 측정
가스 폭발사고	- 2013월 9월 23일 오후 11시 45분경 대구 발생 - 2013년 9월 21일 ~ 2013년 9월 24일 트윗 사용 - TF 30분, DF 48시간 단위로 실험 - 실험 구간은 오후 11시부터 새벽 3시 30분까지 30분 간격으로 측정

3.1 *TF-IDF* 알고리즘

표 3의 두 이벤트에 적용할 수 있는 기존의 *TF-IDF*의 알고리즘은 표 4와 같다.

3.2 가중치 보정

3.2.1 *TF* 보정

같은 양의 트윗이 발생하더라도 평소에 발생하는 트윗 양에 따른 등락률에는 차이가 존재한다. 그러나 기존의 *TF-IDF*를 그대로 적용할 경우 이를 적절히 반영하지 못한다. 이러한 문제를 개선하기 위해 표 4의 *TF*를 보정한

*RTF(Revised Term Frequence)*는 표 5와 같다.

<표 4> *TF-IDF* 알고리즘

TF	$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: 문서dj에 출현한 단어ti의 수 $\sum_k n_{k,j}$: 문서dj에 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$t_j$가 등장하는 문서의 수
TF− IDF	$TFIDF_{i,j} = tf_{i,j} \times idf_i$

<표 5> *RTF*

*TF*를 보정한 *RTF*	$rtf_{i,j} = \frac{n_{i,j}}{\sum_j n_{i,j}}$ $n_{i,j}$: 문서dj에 출현한 단어ti의 수 $\sum_j n_{i,j}$: 모든 문서에 출현한 단어ti의 수

3.2.2 *IDF* 보정

기존의 *TF-IDF* 알고리즘에서는 *IDF*에 log를 취한다. 하지만 실시간 이벤트 탐지는 *TF*를 30분 혹은 그 이하의 단위시간에서 발생한 데이터에서 분석이 이루어지기 때문에 log를 적용할 만큼 큰 데이터가 아니다. 따라서 *DF* 값의 차이에 의한 분류결과를 더 명확히 구분하기 위해 log를 취하지 않았다. 표 4의 *IDF*를 보정한 *RIDF(Revised Inverse Document Frequency)*는 표 6과 같다.

<표 6> *RIDF*

*IDF*를 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$t_j$가 등장하는 문서의 수

3.2.3 계수 적용

표 5와 같이 *TF* 값은 평소에 발생하는 트윗 수가 적은 지역에서 발생한 이벤트의 순위를 높이고 그렇지 않은 지역에서 발생한 이벤트의 순위는 낮추는 역할을 한다. 그러나 영향력 있는 사용자에 의해 널리 리트윗 된 트윗이 차지하는 비율이 높다면 문제는 달라진다. 즉, 트위터가 가진 막강한 전파기능인 리트윗이 이벤트 탐지에서는 성능을 저하시키는 원인이 된다. 이러한 문제를 해결하기 위해 보정된 *RTF-IDF*에 이벤트 관련 트윗의 다양성을 고려한 계수 *VT(Variety of Tweets)*를 적용하였다. *VT*는 트윗의 종류를 의미하며, 동일한 텍스트를 갖는 리트윗된 복수의 트윗을 하나로 취급한다. 다시 말해 *VT* 값이 높다는 것은 같은 이벤트를 언급한 트윗의 종류가 다양하다는 것을 말하며 이는 더 많은 사용자가 해당 이벤트를 언급했음을 의미한다. 계수 *VT*를 적용한 *RTF-IDF(Revised TF-IDF)*는 표 7과 같다.

<표 7> *VT* 적용한 *TF-IDF* 알고리즘

계수가 적용된 *RTFIDF*	$RTFIDF \bullet VT = RTF \bullet RIDF \bullet VT = rtf_{i,j} \times ridf_i \times vt$

4. 실험 결과

본 실험에 사용된 환경은 표 8과 같다.

<표 8> 실험 환경

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

표 3의 두 이벤트에 변형된 *TF-IDF* 알고리즘을 적용한 실험 결과는 표 9 및 표 10과 같다.

<표 9> 대구 KTX 추돌사고의 순위 비교

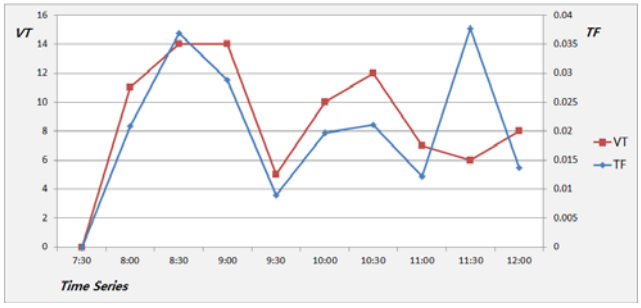
구분 / 시간	TF-IDF	RTF-IDF (RTF)	RTF-IDF (RTF, RIDF)	RTF-IDF·VT
07:30	-	-	-	-
08:00	19	27	21	5
08:30	12	-	22	9
09:00	14	30	22	5
09:30	-	-	-	-
10:00	-	-	-	22
10:30	29	-	29	15
11:00	-	-	-	-
11:30	17	-	25	14
12:00	-	-	-	28

<표 10> 대구 가스 폭발사고의 순위 비교

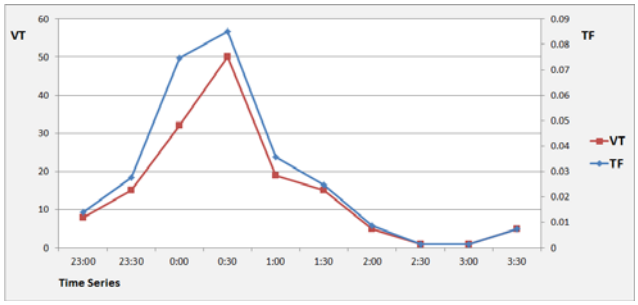
구분 / 시간	TF-IDF	RTF-IDF (RTF)	RTF-IDF (RTF, RIDF)	RTF-IDF·VT
23:00	-	-	-	-
23:30	16	-	-	15
00:00	3	-	28	3
00:30	2	-	24	1
01:00	6	-	26	8
01:30	9	-	27	11
02:00	23	-	-	23
02:30	-	-	-	-
03:00	-	-	-	-
03:30	12	17	16	6

표 9는 대구 KTX 추돌사고, 표 10은 대구 가스 폭발 사고를 TF-IDF, RTF-IDF(RTF), RTF-IDF(RTF, RIDF), RTF-IDF·VT에 적용하여 이벤트 구간 별로 순위를 비교하였다. 순위 비교는 빠른 시간에 순위가 높을수록 좋은 성능을 보인 알고리즘이라고 할 수 있다. *TF-IDF* 알고리즘에 *TF* 값만 보정했을 경우 낮은 성능을 보이는 이유는 두 이벤트가 발생한 대구의 경우 평소 많은 트윗에서 언급되는 대도시이기 때문에 *RTF* 값이 작기 때문이다. *IDF*를 보정한 경우 log 제외로 *RIDF* 값들의 차이가 명확해져 *RTF* 값만 적용했을 경우보다 높은 감지성능을 보였다. 마지막으로 가중치 보정과 계수를 적용한 *RTF-IDF·VT*의 경우 표 9의 대구 KTX 추돌사고에서는 순위 5이고 *TF-IDF*의 경우 순위 19이다. 또한, 표 10의 대구 가스 폭발 사고에서는 *RTF-IDF·VT*의 경우 순위 15이고 *TF-IDF*의 경우 순위 16이다. 그러므로 *RTF-IDF·VT*는 *TF-IDF* 알고리즘을 그대로 적용했을 때보다 뛰어난 감지성능을 보인다.

*VT*를 적용한 이유는 *TF*와 *VT*의 상관관계를 보여주는 그림 1과 그림 2를 통해 유추할 수 있다. 실험에 사용된 두 가지 이벤트가 유사한 결과를 보이는 것을 통해 실제 이벤트일 경우 *TF*가 높게 향상되는 만큼 *VT*도 유사한 패턴을 나타냄을 증명한다. 이로써 소수의 계정에서 작성된 트윗이 리트윗에 의해 높은 *TF*를 형성한다고 하더라도 *VT*에 의해 낮은 *TF-IDF* 값을 갖도록 한다. 다시 말하면 여러 사람에 의해 많이 언급된 이벤트의 순위를 높일 수 있다.



(그림 1) KTX 추돌사고의 *TF* 값과 *VT* 값 비교



(그림 2) 가스폭발 사고의 *TF* 값과 *VT* 값 비교

5. 결론 및 향후 연구

트위터를 구성하고 있는 트윗은 140글자의 제한, 리트윗의 파급력 등 일반적인 문서와는 다른 성향을 갖는다. 이에 본 논문은 트위터를 이용해 이벤트를 감지할 때, 기존 웹 문서에서 사용하던 *TF-IDF* 알고리즘을 그대로 적용할 경우의 문제점을 파악하고 변형된 알고리즘을 제안하였다. 또한 기존의 *TF-IDF* 알고리즘의 가중치를 보정하고 계수를 적용한 변형된 *RTF-IDF·VT* 알고리즘을 두 번의 이벤트에 적용하여 실험한 결과 트위터를 이용한 이벤트 감지의 성능이 향상됨을 보였다. 이는 추후 트위터를 통한 실시간 이벤트 감지 시스템을 설계할 때 중요한 고려사항이 될 수 있다.

향후에는 보다 정확하고 빠른 이벤트 감지 성능을 보일 수 있도록 리트윗 뿐만 아니라 해시태그(Hash Tag) 단위의 클러스터링, 위치정보(Geocode)를 추가한 지역분류 등 다양한 조건을 *VT* 계수에 적용하는 실험을 진행할 필요가 있다.

참고문헌

[1] Meenakshi Nagarajan, Karthik Gomadam, Amit P. Sheth, Ajith Ranabahu, Raghava Mutharaju, and Ashutosh Jadhav, "Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences," In Proceedings of the *10th International Conference on Web Information Systems Engineering*, pp. 539-553, 2009.

[2] 임준엽, 김석중, 윤진영, 오평화, 이범석, 황병연, "트위터를 이용한 지역 이벤트 탐지", 한국정보과학회 한국컴퓨터종합학술대회 논문집, pp. 248-250, 2013.

[3] 이성직, 김한준, "TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법", 한국전자거래학회지, 제14권, 제4호, pp. 59-73, 2009.

[4] Apache Lucene Korean Analyzer and dictionary, http://sourceforge.net/projects/lucenekorean, 2013.

[5] Twitter Streaming API, http://dev.twitter.com/docs, 2013.