특징추출을 이용한 트위터 메시지 주제 분류 방법

송지민* · 김한우* · 김동주* · 정성훈*

*한양대학교, *안양대학교

A Method of Classifying Tweet by subject using features

Ji-min Song* · Han-woo Kim* · Dong-joo Kim* · Sung-hoon Jung*

*Hanyang University, *Anyang University

E-mail : gaejm17@gmail.com

요 약

트위터는 전세계적으로 다양한 정보와 의견을 공유하는 교류의 장으로 이용되고 있다. 트위터에서 생성되는 막대한 양의 데이터를 활용하려는 시도가 이루어지고 있다. 그 중 다양한 주제별 정보를 추출하여 이용하려는 연구가 활발히 진행되고 있다. 트위터는 140자의 짧은 메시지로 정보를 공유하는 서비스이다. 이러한 짧은 메시지는 트윗에서 다양한 주제별 정보를 추출하는 것을 어렵게 한다. 본 논문에서는 트윗의 기능들과 분류할 주제의 특징을 이용하여 트윗 주제별 분류 방법을 제안한다. 이 방법의 유용성을 검증하기 위해, 트윗 API를 사용하여 수집된 10000개의 트윗으로 실험하였다. 그 결과 기존 연구들보다 뛰어난 결과를 얻었다.

ABSTRACT

Twitter is the special place that people in the world can freely share their information and opinion. There are tries to utilize a vast amount of information made from twitter. The study on classification of tweets by subject is actively conducted. Twitter is a service for sharing information with short 140-characters text message. The short message including brief content makes extracting a variety of information hard. In the paper, we suggests the method to classify tweet by subject. The method uses both tweet and subject features. In order to conduct experiments to verify the proposed method, we collected 10,000 tweet messages with the Twitter API. Through the experimental results, we will show that the performance of our proposed method is better than those of previous methods.

키워드

트위터, 트윗, 분류, 특징

Ⅰ. 서 론

트위터는 트윗이라고 부르는 140자로 제한된 메시지를 통해 정보와 자신의 의견을 다른 사람들과 공유하는 마이크로블로그 중 하나이다. 전세계적으로 트위터의 이용자 수는 5억명 이상이고, 매일 방대한 정보들을 생성하고 있다. 이러한 정보를 활용하기 위한 시도가 이루어지고 있다. 특히 특정 주제에 대한 정보를 추출하여 활용하려는 연구가 활발히 진행 되고 있다. 그러나 트윗을 분석하여 주제별로 분류하는 것은 어렵다. 트윗의 짧은 길이는 기존의 문서 분류방법인 Bag-Of-Words 방법을 사용하기에는 부족한 단어들의 수를 가지고 있을 뿐만 아니라 짧은 길이의 메시지와 자유로운 양식은 사람들로 하여금 줄인말과 비속어를 사용하게 만든다[1]. 이러한 문제점을 해결하기 위해 본 논문에서는 트위터의 기능들과 분류하려는 특정 주제의 특징을 이용하여 분류하는 방법을 제안한다.

Ⅱ. 관련 연구

트위터에서 특정 주제의 트윗을 분류하기 위한

다양한 방법들이 제시 되었다. Jagan[2]은 뉴스 트윗만을 분류하여 제공하는 TwitterStand 시스템을 개발 하였다. [2]는 뉴스 트윗을 분류하기 위해 Seeder를 사용하였다. Seeder는 200개의 방송국, 신문사에 속한 사람 또는 뉴스 관련 블로거들의 계정으로 구성되어 있다. Bharath[3]은 트윗들을 뉴스, 이벤트, 의견, 거래, 개인 메시지와 같은 5가지의 주제로 분류를 하였다. [3]은 저자 정보와 각각의 주제별 트윗이 가지고 있는 특징을 이용하여 분류를 하였다. Perdana[4]는 트위터의 트렌딩 토픽을 단어 빈도를 이용하여 주제별로 분류 하였다. Takeshi[5]는 트위터에서 지진에 관련된 정보를 분류하여 지진 시간과 위치를 예측하는 시스템을 개발하였다. Earthquake 또는 shaking과 같은 지진에 관련된 단어가 포함된 트윗이 특정 시간 동안 빈번히 탐지되면, 트윗들의 위치를 분석하여 예상 지역에 있는 사용자들에게 정보를 보내주는 시스템을 구축하였다.

III. 본론

본 논문에서는 뉴스, 의견, 좋은 글귀와 같은 3개의 주제로 트윗들을 분류하겠다. 여기서 좋은 글귀란 위인들이 남긴 명언이나 영화 또는 책에서 나온 글들로써, 감동을 주는 글귀를 말한다. 트윗들을 각각의 주제로 분류하기 위해 트위터의 기능들과 각 주제별 트윗의 특징을 이용하여 분류 하였다.

3.1 트윗 주제 분류를 위해 사용한 트위터의 기능들

트윗들을 주제별로 분류하기 위해 사용한 트위터의 기능들은 Retweet, URL 첨부기능, Mention, Favorites를 사용하였다.

트위터의 Retweet 기능은 이메일의 포워드 기능과 같이 다른 사용자가 작성한 트윗을 내 계정을 통해 다시 작성하는 기능이다 트위터 사용자들은 Retweet을 주로 유용한 정보 등을 팔로워들에게 알리기 위해 사용한다. 이러한 Retweet은 빠른 정보 전파는 물론, 트위터 사용자들의 관계를 넓히는 효과까지 가져다 준다. 트위터 유저들이 Retweet을 가장 많이 하는 트윗의 주제는 뉴스이고 두 번째로 많이 하는 주제는 사용자들의 의견이다[6]. 이러한 특징을 뉴스와 의견 트윗을 분류하는데 사용했다.

또한 트위터는 URL첨부를 통해 140자로 제한된 트윗의 단점을 완화해 준다. 또한 URL의 길이를 줄여주는 짧게 줄인 URL 알고리즘(Shortened URL Algorithm)을 제공하고 있다. 대부분의 뉴스 트윗은 140자로 정보를 다 제공할 수가 없기에 부족한 정보를 위해 URL을 첨부한다.

트위터는 트윗에서 특정 트위터 사용자를 언급하는 Mention기능이 있다. Mention을 통해, 특정 사용자와의 의견을 나눌 수 있다. 사용 방법은 특수문자 @ 뒤에 호출하고 싶은 트위터 사용자의 아이디를 쓰면 된다.

트위터에는 Favorites 기능이 있다. 이 기능은 트윗들 중 다시 보기를 원하는 트윗들을 따로 보관하는 기능이다. 이 기능을 통해 트윗들을 에 볼 수 있다. 좋은 글귀 트윗은 다른 주제의 트윗보다 높은 Favorites 수를 가지고 있다.

3.2 각 주제의 트윗이 가지고 있는 특징들

본 논문에서는 트윗을 뉴스, 좋은 글귀 의견 트윗으로 분류하기 위해 트위터의 기능적인 면뿐만 아니라, 각 분류할 주제의 특징을 이용한다.

트위터에서 대부분의 좋은 글귀 트윗은 글귀를 말한 사람의 이름을 포함하고 있다. 아래의 그림 1은 좋은 글귀 트윗의 예이다.

```
God never sends us more than we
can handle - Mother Theresa
```

그림 1. 좋은 글귀 트윗의 예

이러한 특징을 이용하기 위해 좋은 글귀에서 자주 나오는 2만명 이상의 사람들의 이름으로 사전을 구축하였다. 이 사전에 포함된 사람들은 정치가, 뮤지션, 배우, 과학자와 같은 직업들로 이루어져 있다. 표 1은 사전에 포함된 사람들의 예이다.

표 1. 좋은 글귀 사람사전의 예

이름	직업
Walt Disney	만화가
Jesus Christ	종교인
Steve Jobs	사업가
Albert Einstein	물리학자
Bob Marley	음악가

또 다른 좋은 글귀의 특징은 일반 대화와는 다르게 비속어, 이모티콘 등이 포함되어 있지 않다.

뉴스는 감정을 배제하고 객관성을 유지하며 정보를 제공해야 한다. 이러한 특성을 위해 우리는 감정어를 뉴스 트윗을 분류하는데 사용하였다. 또한 뉴스들은 육하원칙에 따라 정보를 제공한다. 이를 이용하여 고유명사인 유명인과 도시와 나라 이름을 뉴스 트윗을 분류하는데 사용하였다. 육하원칙 중 '누가'에는 유명인이 '어디'에는 도시이름과 나라이름이 포함된다[7].

반면에 의견 트윗은 어떤 대상에 대하여 가지는 생각과 감정을 표현한 것이다. 이것은 뉴스 트윗과는 반대되는 특징으로 감정어를 이용하여 의

견 트윗을 분류하였다. 트위터에서는 이모티콘으로도 의견을 표현 할 수 있다. 또한 비속어, 줄인 말을 통해서도 의견과 감정을 나타낼 수 있다.

Ⅳ. 실험 및 결과

실험데이터는 Twitter API를 이용하여 약 10000개의 트윗을 사용하였다. 수집된 트윗 중, 소수의 단어로만 이루어진 트윗과 인사, 안부의 말은 제거하였다. 그 후, 트윗들을 3000개씩 직접 3가지의 주제로 나누었다. 실험을 위해 WEKA를 이용하였고, 분류기로는 Naive Bayes를 사용하였다. 10-fold cross validation을 사용하였다. 표 2는 각 주제 별 정확률과 재현율이다.

표 2. 각 주제별 정확률, 재현율

	좋은글귀	뉴스	의견
정확률	86.5%	78.6%	78.4%
재현율	85.6%	75.6%	78.0%

실험은 정확률과 재현율로 평가 하였다. 정확률은 분류기에 의해 분류된 각 주제별 트윗에서 실제 주제 트윗의 포함된 비율이다. 재현율은 실제 주제별 분류한 트윗에서 분류기에 의해 분류된 주제별 트윗의 비율이다. 본 논문에서 제안한 방법은 기존의 분류 방법[4]보다 더 우수한 성능을 보였다.

Ⅴ. 결론 및 향후 연구 방향

본 논문에서는 트위터에서 트윗을 각 주제별로 분류하기 위해 트위터의 기능과 각 주제별 특징을 이용하는 방법을 제안하였다.

그러나 향후 더 분별력 있는 특징정보들을 고안할 필요가 있다.

참고문헌

[1] RITTER, Alan, et al, Named entity recognition in tweets: an experimental study, Proceedings of the Conference on Empirical Methods in Natural Language Processing, p. 1524-1534, July 2011.

[2] SANKARANARAYANAN, Jagan, et al, Twitterstand: news in tweets, Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 42-51, November 2009.

[3] SRIRAM, Bharath, et al, Short text classification in twitter to improve information filtering, Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, p. 841-842, July 2010.

[4] Adhitama, Perdana et al, 나이브 베이지안 분류기를 이용한 트위터 트랜딩 주제 분류, 한국정보과학회, p. 879-991, 11월 2013.

[5] SAKAKI, Takeshi et al, Earthquake shakes Twitter users: real-time event detection by social sensors, Proceedings of the 19th international conference on World wide web, p. 851-860, April 2010.

[6] 이성윤 et al, 사용자의 활동과 영향력을 이용한 트위터의 URL 추천 시스템, 정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 17권 제8호, p. 447-456, 8월 2011.

[7] LI, Juanzi, et al, Keyword extraction based on tf/idf for Chinese news document, Wuhan University Journal of Natural Sciences, Vol.12, No.5, p. 917-921, September 2007.