트위터의 실시간 트렌드에서 해시태그 추천을 개선하기 위한 해시태그 분류기법

(A Hashtag Classification Method for Improving Hashtag Recommendation in Twitter Trending Topic)

김 혜 원 † 김 형 주 †† 김 기 성 †
(Hyewon Kim) (Hyoung-Joo Kim) (Kisung Kim)

요 약 본 논문에서는 트위터의 해시태그 추천을 개선하기 위한 해시태그 분류 기법을 제안한다. 트위터의 해시태그에는 그 특성과 사용된 의도가 분명하게 다른 두 가지 유형이 있다. 하나는 사용자들에게 정보를 주기 위해 사용되는 정보 제공형 해시태그이고 다른 하나는 사용자들이 해시태그를 사용하게끔 유도하는데 사용되는 참여 유도형 해시태그이다. 우리가 해시태그의 유형을 분류할 수 있다면 현재 트위터의 실시간 트렌드에서 사용되는 해시태그 추천 결과를 풍성하게 해줄 수 있을 것이다. 이에 본 논문에서는 나이브 베이지안 분류 기법을 이용해 해시태그를 분류하는 방법을 제안한다. 이를 위해 해시태그를 포함하는 트윗들을 분석하여 분류 조건을 만들었다. 또한, 실제 트위터 데이터를 이용한 실험을 통해 본 논문에서 제안한 분류 기법의 정확도를 보였고 각 분류 조건들의 효과에 대해 분석했다.

키워드 : 트위터, 해시태그, 분류, 추천

Abstract In this paper, we suggest a hashtag classification method for improving hashtag recommendation in Twitter. There are two types of hashtags with different characteristics and intentions. One is an informative hashtag used to provide users with information and the other is a meme hashtag used to induce users to participate in micro-meme. If we could classify the types of hashtags automatically, we can improve the result of hashtag recommendation in trending topics of Twitter. To address this concern, we propose a method for classification of the hashtags using Bayesian classification approach. We propose several measures for classification through analysis of tweets having hashtags. Also, we show the effectiveness of our approach through experiments using real-life twitter data, and analyze the effects of our measures.

Key words : Twitter, Hashtag, Classification, Recommendation

1. 서 론

마이크로블로깅 서비스 중에 하나인 트위터는 2006년 3월에 서비스를 시작한 이래로 현재까지 많은 사용자들에게 인기를 얻고 있다. 2012년 2월을 기준으로 트위터에 등록된 사용자 수는 4억 6천 5백만명이 넘고 이들은 하루에 1억 7천 5백만개 이상의 트윗을 발행한다.[1)]

다른 마이크로블로깅 서비스와 구별되는 트위터 고유의 특징으로 해시태그라는 것이 있다. 해시태그는 트위터 사용자들이 트윗을 주제별로 분류하기 위해 자발적으로 만들어서 사용하고 있는 트위터의 태그이다. 사용자들은 트윗에서 특정 키워드나 주제어를 표시하기 위해 해시태그를 사용한다. 해시태그는 특정 문자열 앞에 해시기호(#)를 붙이는 형태로 사용되며 작성한 트윗의

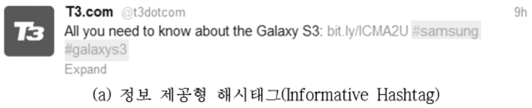
· 본 연구는 BK-21 정보기술 사업단의 연구결과로 수행되었음
· 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 20120009186)

† 비 회 원 : 서울대학교 컴퓨터공학과
hyewonkim@idb.snu.ac.kr
(Corresponding author임)
kskim@idb.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학과 교수
hjk@snu.ac.kr

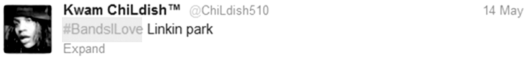
논문접수 : 2012년 7월 17일
심사완료 : 2012년 8월 8일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제18권 제11호(2012.11)

1) http://infographiclabs.com/news/twitter-2012



(a) 정보 제공형 해시태그(Informative Hashtag)



(b) 참여 유도형 해시태그(Meme Hashtag)

그림 1 해시태그 예제

어디에나 위치할 수 있다. 사용자가 해시태그와 함께 트윗을 작성하면 해시태그에 자동으로 하이퍼링크가 생성되는데, 이 링크를 클릭하면 같은 해시태그가 사용된 트윗들을 모아서 볼 수 있다.

해시태그에는 그 특성과 사용된 의도가 분명하게 다른 두 가지 유형이 있다. 기존 연구[1,2]에서는 해시태그의 특징을 분석하여 해시태그의 유형을 암시적으로 언급했으나 해시태그의 유형에 대한 구체적인 언급은 없었다. 이에 본 논문에서는 해시태그의 특성에 따른 유형을 정의하고자 한다. 첫 번째 유형은 정보 제공형 해시태그(Informative Hashtag)이다. 정보 제공형 해시태그는 그림 1(a)에서 사용된 #samsung 이나 #galaxys3처럼 트윗 내용을 요약하거나 트윗 내용과 관련된 정보를 주기 위해 사용된다. 또한 기존의 Delicious[2]와 같은 Web2.0사이트에서 사용되는 태그와 유사한 목적과 특징을 가지며, 사용자들 사이에서 꾸준히 이용되는 경향을 가진다. 두 번째 유형은 참여 유도형 해시태그(Meme Hashtag)이다. 참여 유도형 해시태그의 예로는 그림 1(b)에서 사용된 #BandsILove 가 있다. 이 태그는 자신이 좋아하는 밴드를 다른 사람들에게 알리기 위해 사용하는 태그이다. 태그 자체에 특별한 의미는 없지만 이 태그를 포함하는 트윗을 작성하는 것에 의미가 있다. 주로 사용자들의 참여를 유도하기 위해 사용되고, 사용자들 사이에서 갑자기 인기를 얻어 많이 사용되다가 금방 사라지는 경향이 있다.

해시태그는 트위터의 '실시간 트렌드'에서 트렌딩 토픽으로 많이 추천된다. 실시간 트렌드는 현재 트위터에서 화제가 되고 있는 토픽을 제공하는 시스템이다. 이를 통해 전세계적으로 지금 어떤 일들이 일어나고 있는지, 어떤 주제들이 이목을 끌고 있는지 쉽게 발견할 수 있다.[3] 그러나 실시간 트렌드는 짧은 기간에 사용 빈도수가 급증하는 주제어를 토픽으로 추천하기 때문에[4] 대체적으로 참여 유도형 해시태그만 트렌딩 토픽으로 추천된다. 이런 실시간 트렌드 시스템의 추천 방식의 한계 때문에 정보 제공형 해시태그는 트렌딩 토픽으로 추천되지 못한다. 따라서 사용자들은 한쪽으로 치우친 주제만을 추천 받는다. 트위터에서 화제가 되고 있는 토픽을 편향 없이 추천해 주기 위해, 해시태그의 유형을 자동으로 분류하여 모든 해시태그의 유형을 트렌딩 토픽으로 추천해야 한다. 해시태그의 유형 분류 결과에 따라 다른 방식의 추천 기법을 적용하면 더욱 더 풍성한 추천 결과를 얻을 수 있을 것이다.

이에 본 논문에서는 트위터의 해시태그 추천을 개선하기 위한 해시태그 분류 기법과 해시태그 분류에 이용되는 분류 조건들을 제안한다. 이미 해시태그의 성격과 역할에 대해 언급한 논문들은 있었으나, 해시태그의 유형을 명확히 정의하고 분류한 논문은 없었다. 본 논문에서는 해시태그의 유형을 두 가지로 정의하고 나이브 베이지안 분류 기법을 적용해 해시태그의 유형을 자동으로 분류하는 기법을 제안한다. 해시태그를 분류하기 위해 해시태그가 사용된 트윗의 여러 특성(URL 포함여부, 해시태그의 개수, 사용된 시간, 해시태그 길이 등)들을 분류 조건으로 선택하고, 이를 이산화된 값으로 변환하기 위해 불순도를 측정하는 방법인 엔트로피를 적용했다. 또한 실제 트위터의 데이터를 이용해 본 논문에서 제안한 방법이 효과적으로 해시태그를 분류할 수 있음을 보인다. 본 논문에서는 해시태그의 유형 정의와 분류 방법에 초점을 맞추고, 이에 따른 추천 알고리즘의 개선은 다루지 않으며 향후 연구로 남긴다.

2. 관련 연구

소셜네트워크 서비스가 전세계적으로 급속한 성장세를 보임에 따라 트위터를 비롯한 소셜네트워크 서비스에 대한 분석이 계속해서 활발하게 이루어지고 있다.

트위터에 대한 많은 연구들이 해시태그의 유용성을 밝혔다. [3]에서는 트윗 메시지를 군집화 할 때 해시태그가 주제를 나타내는 좋은 지표로 사용될 수 있음을 실험적으로 증명했다. 사용자들이 매일 많은 트윗들을 발행하기 때문에 우리가 모든 트윗들을 읽어보고 이해하기에는 큰 어려움이 요구된다. 이 많은 정보들을 효율적으로 받아들이기 위하여 트윗들을 같은 주제로 분류하는 기술이 필요해졌다. 이 때 해시태그를 지표로 삼아 감독분류(Supervised classification)방법으로 분류했더니 괄목할 만한 결과를 얻을 수 있었다. [4]에서는 해시태그가 특정 관심사에 대한 커뮤니티를 생성하거나 사용자들의 대화 스레드를 생성하기 위한 수단으로 사용됨을 언급했다. 이처럼 해시태그는 트위터에서 하나의 식별자로 사용될 수 있음을 지적하고 이를 평가하는 방법을 제안했다. 또한 [5]에서는 2010년 미국의 선거운동으로 사용된 트윗을 분석하여 해시태그가 하나의 커뮤니케이션 그룹을 찾

2) http://www.delicious.com
3) https://support.twitter.com/articles/101125-about-twitter-trends
4) http://blog.twitter.com/2010/12/to-trend-or-not-to-trend.html

아낼 수 있는 식별자로 사용됨을 밝혔다.

기존 해시태그 관련 연구들 중에서 해시태그의 유형에 대해 암시적으로 언급한 논문들이 있다. [1]은 트위터의 해시태그와 웹에서 사용되는 태그를 비교했다. [1]의 연구 결과에 따르면, 웹에서 사용되는 전통적인 태그는 나중에 이루어질 정보검색을 효율적으로 하기 위한 목적으로 사용되고 리소스를 요약하는 단어들로 구성된다. 반면 트위터의 해시태그는 사용자로 하여금 트윗 내용을 상기시키기 위한 목적으로 사용되기 보다는 해시태그와 관련된 트윗들을 모아서 보기 위한 목적으로 사용된다. 또한 해시태그의 사용으로 마이크로-미미(micro-meme)라는 트위터만의 독특한 문화가 생겨났고 이에 사용자들은 마이크로-미미와 관련된 해시태그를 발견하면 그 해시태그를 포함하는 트윗을 작성하기를 즐긴다는 것을 밝혔다. [2]는 트위터에서 해시태그는 두 가지 역할을 동시에 하고 있다고 제안했다. 첫째로 해시태그는 트위터의 소셜 북마크 역할을 한다. 해시태그는 컨텐츠에 주석을 붙이는 것처럼 사용되고 많은 사람들에게 공유되어 트윗을 분류하는 폭소노미로 활용된다. 둘째로 해시태그는 트위터에서 커뮤니티를 나타내는 심볼 역할을 한다. 해시태그는 트위터에서 이야기가 오가는 커뮤니티를 나타내는 지표로 사용되어 사용자들이 이 커뮤니티를 쉽게 발견하게 하고 참여하게 하는 역할을 한다. 본 연구에서는 이 두 논문의 아이디어를 기반으로 해시태그의 두 가지 유형을 정의하고 각각을 정보 제공형 해시태그(Informative hashtag)와 참여 유도형 해시태그(Meme hashtag)로 명명했다.

사용자들이 발행하는 트윗들이 많아짐에 따라 트윗을 분류하는 연구도 활발히 이루어졌다. 트윗의 내용이 140자 이하로 제한되기 때문에 기존의 분류 기법을 트윗 분류에 적용하기에는 무리가 있다. 이를 극복하기 위해서 [6]에서는 트윗이 분류 될 카테고리와 트윗이 각 카테고리에 분류되기 위한 조건들을 미리 정의해 두고 트윗을 분류했다. 분류 조건으로는 트윗을 작성한 저자에 대한 정보와 트윗의 내용이 사용됐다. 또한 [7]에서는 트렌딩 토픽을 미리 정의해 둔 카테고리로 분류했다. 분류 조건으로는 트윗 내용과 리트윗, 리플라이 등의 트윗 자체의 정보를 이용했다. [6,7]의 연구 목적에 따르면 트윗의 유형이 명시된다면 사용자는 필요에 따라 정보를 선택하여 볼 수 있기 때문에 사용자의 정보 처리 능력이 향상된다고 한다. 이와 같은 맥락으로 본 논문에서는 해시태그와 함께 해시태그의 유형을 명시하여 사용자의 정보 처리 능력 향상을 기대한다.

3. 두 종류의 해시태그

이번 절에서는 두 가지 유형의 해시태그를 정의하고, 각 유형의 특성에 대해 설명한다.

3.1 정보 제공형 해시태그(Informative hashtag)

정보 제공형 해시태그는 트윗을 나중에 다시 검색할 때 이용할 목적으로 사용되며 주로 트윗을 대표할 수 있는 한 단어로 구성된다. 정보 제공형 해시태그의 예로는 #job, #green, #android, #apple 등이 있다. 정보 제공형 해시태그는 트윗을 주제별로 분류하는 폭소노미로 활용되기도 한다. 따라서 정보 제공형 해시태그는 많은 사용자들에게 널리 공유되어 꾸준히 사용되는 특성을 갖는다.

정보 제공형 해시태그를 포함하고 있는 트윗들은 대체로 유용한 정보를 담고 있다. 그러나 트윗은 140자 이내로 작성해야 하는 길이의 제약이 있기 때문에, 정보 제공형 해시태그를 포함하는 트윗들은 여러개의 해시태그들과 URL 링크를 포함하는 경향이 있다. 다른 해시태그들을 이용하면 몇 글자의 해시태그로 트윗의 주제를 제공할 수 있다. 또 짧은 텍스트로 표현하기 힘든 뉴스 기사나 사진, 동영상 등의 정보는 URL링크를 이용해서 보여줄 수 있다.

이처럼 정보 제공형 해시태그가 사용된 트윗은 내용 자체뿐만 아니라, URL 링크나 다른 해시태그들을 통해 유용한 정보를 제공한다. 또한 이런 트윗들은 리트윗을 통해 여러 사용자에게 전달되는 경향이 있으며, 따라서 사용자들에게 정보 제공형 해시태그를 추천해줄 때에 리트윗 하도록 제안한다면 정보를 효과적으로 확산할 수 있을 것이다.

3.2 참여 유도형 해시태그(Meme hashtag)

참여 유도형 해시태그는 사용자들로 하여금 이 해시태그를 포함하는 트윗을 작성하도록 유도한다. 참여 유도형 해시태그는 다른 사람들이 사용한 태그를 보고 그 태그에 어울리는 트윗을 작성하는 방식으로 사용하며, 이런 현상을 마이크로-미미(micro-meme)라고 한다. 참여 유도형 해시태그는 태그를 사용한 트윗을 작성하여 마이크로-미미에 참여하는 것에 의미가 있다.

참여 유도형 해시태그의 예로는 자신이 좋아하는 사람에 대한 내용을 나누는 #15PeopleILove 처럼 자신의 경험이나 생각을 공유하거나, #HappyBirthdayIU처럼 자신이 좋아하는 연예인의 생일을 알리고 함께 축하하기 위한 해시태그, #WeWantBigbangInItaly처럼 자신이 좋아하는 연예인을 응원하는 해시태그도 있다. 이처럼 참여 유도형 해시태그는 주로 상황을 설명하는 문장으로 구성된다.

참여 유도형 해시태그는 사용자들 사이에서 갑자기 인기를 얻어 많이 사용되다가 금방 사라지는 특징이 있다. 특히 참여 유도형 해시태그는 실시간 트렌드에 노출되었을 때 사용량이 폭발적으로 증가한다. 따라서 참여

유도형 해시태그를 추천할 때에는 그 순간 가장 인기 있는 해시태그를 추천해주는 것이 좋다. 또한 정보 제공형 해시태그와는 달리, 참여 유도형 해시태그는 리트윗보다는 새로운 트윗을 작성하는 것이 의미가 있다. 따라서 사용자들에게 참여 유도형 해시태그를 포함한 새로운 트윗 작성을 제안함으로써 사용자의 자발적인 참여를 유도할 수 있을 것이다.

4. 해시태그의 분류

이번 절에서는 앞서 정의한 해시태그의 유형을 분류하기 위한 모델을 제안한다.

4.1 해시태그 분류 모델

해시태그를 분류하는 과정은 해시태그를 학습하는 단계와 해시태그를 분류하는 단계로 나누어 볼 수 있다. 해시태그를 학습하는 단계에서는 각 해시태그가 사용된 트윗들을 분석해서 분류기를 만든다. 분류기를 만들 때 트윗들의 내용은 고려하지 않고 트윗이 가지고 있는 트윗 자체의 특징들만 이용한다. 이렇게 만들어진 분류기에 새로 분류할 해시태그를 적용하면 해시태그의 타입이 정보형 해시태그인지 참여 유도형 해시태그인지를 알 수 있다. 이 모든 과정을 그림 2에 흐름도로 나타내었다.

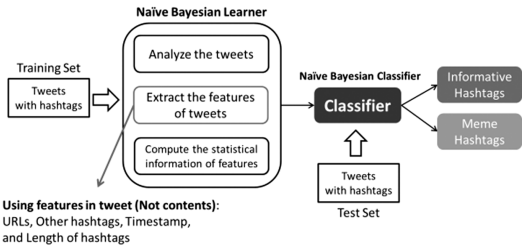


그림 2 해시태그 분류 흐름도

4.2 해시태그 분류 방법

본 논문에서는 해시태그를 분류하기 위한 방법으로 나이브 베이지안 분류 방법[8]을 사용했다. 나이브 베이지안 분류 방법은 베이지안 정리를 기반으로 한 확률적인 분류 방법이다. 각 인스턴스는 분류 조건들의 결합으로 표현되며 인스턴스의 분류 결과는 분류 조건 값들에 따라 계산되는 확률로 결정된다.

해시태그는 분류 조건 $a_1, a_2, \cdots, a_k$ 의 결합으로 표현되고 해시태그의 분류 결과 c는 해시태그 유형의 집합 C에서 하나를 취한다. 새로운 해시태그가 들어왔을 때 이 해시태그의 분류 결과는 식 (1)을 계산하여 가장 높은 확률을 만드는 해시태그 유형으로 정해진다.

$$c = \arg\max_{c_i \in C} P(c_i \mid a_1, a_2, \cdots, a_k) \quad (1)$$

여기에 베이지안 정리를 적용한다면 다음과 같이 표현된다.

$$c = \underset{c_i \in C}{\arg\max} \frac{P(a_1, a_2, \cdots, a_k \mid c_i) P(c_i)}{P(a_1, a_2, \cdots, a_k)} = \arg\max_{c_i \in C} P(a_1, a_2, \cdots, a_k \mid c_i) P(c_i) \quad (2)$$

나이브 베이지안 분류 방법은 어떤 해시태그에 대하여 분류 조건들끼리 서로 조건부 독립이라는 가정을 가지고 있다. 따라서 분류 조건들의 결합의 확률은 각각의 분류 조건들의 확률의 곱으로 표현될 수 있다.

$$P(a_1, a_2, \cdots, a_k \mid c_i) P(c_i) = \Pi_j P(a_j \mid c_i) P(c_i) \quad (3)$$

식 (3)을 식 (2)에 적용한다면 해시태그의 분류 결과 c는 식 (4)로 계산된다.

$$c = \arg\max_{c_i \in C} \Pi_j P(a_j \mid c_i) P(c_i) \quad (4)$$

식 (4)를 계산하기 위해서 해시태그를 표현하는 분류 조건들이 필요하다. 해시태그 분류에 사용되는 분류 조건들로는 해시태그 h를 포함하는 트윗에 사용된 URL 링크 수, h와 함께 사용된 다른 해시태그들의 수, h의 타임스탬프의 표준편차, h의 길이를 사용했으며 이를 표 1에 정리했다. 해시태그 h_1 을 포함하는 트윗에 URL링크와 다른 해시태그들이 많다면 이 트윗에 많은 정보가 있을 것으로 판단되므로 h_1 은 정보를 주는 해시태그일 확률이 높다. 반면 해시태그 h_2 의 타임스탬프의 표준편차가 낮다면 해시태그 h_2 가 포함된 트윗들이 h_2 의 타임스탬프의 평균과 가까운 시기에 많이 작성되었을 것이므로 h_2 는 참여를 이끄는 해시태그일 확률이 높다. 또한 해시태그 h_3 의 길이가 길다면 h_3 이 문장형 태그로 사용되었을 확률이 높음으로 h_3 를 참여를 이끄는 해시태그로 판단할 수 있을 것이다.

표 1의 분류 조건들이 모두 정수나 실수로 연속적인 값을 가지기 때문에 이것들을 바로 나이브 베이지안 분류 방법에 적용시킬 수 없다. 계산에 필요한 분류 결과에 따른 분류 조건들의 확률을 구할 수 없기 때문이다. 따라서 연속적인 값들을 몇 개의 범주로 나누는 과정이 필요하다.

연속적인 값들을 범주화 할 때는 범주 내에 분류결과가 많이 섞이지 않게 레코드들이 분배 되는 것이 중요하다. 이를 계산하기 위해 범주 내 분류 결과의 불순도를 측정하는 엔트로피[9]를 사용한다. 엔트로피는 식 (5)

표 1 해시태그 분류 조건

	정보 제공형 해시태그	참여 유도형 해시태그
URL 링크	많다	적다
함께 사용된 해시태그	많다	적다
타임스탬프의 표준편차	높다	낮다
해시태그의 길이	짧다	길다

에 의해 계산되며, 엔트로피 값이 가장 적게 계산되는 범주를 채택한다.

$$\text{Entropy(t)} = -\sum_{i=0}^{c-1} p(i|t)\log_2 p(i|t) \quad (5)$$

4.3 해시태그 분류와 해시태그 추천

본 논문에서 제안한 해시태그 분류 기법을 해시태그 추천에 적용하면 실시간 트렌드의 해시태그 추천 결과를 향상시킬 수 있다. 현재 실시간 트렌드는 대부분 참여 유도형 해시태그만을 추천해준다. 이는 최근에 사용 빈도수가 급등한 토픽을 추천하기 때문에 나타나는 결과다. 이런 상황을 개선하기 위해 실시간 트렌드에서 해시태그의 유형을 인지하고, 이에 따라 추천 알고리즘을 달리한다면 현재보다 더 풍성한 해시태그 추천 결과를 제공할 수 있을 것으로 기대된다. 또한 해시태그 추천에만 그치는 것이 아니라 해시태그 유형에 따른 적절한 사용자 행동(리트윗 혹은 새로운 트윗 작성)을 명시적으로 제안해서 사용자들의 적극적인 참여를 유도할 수 있다. 그러나 본 논문에서는 해시태그 분류에 초점을 맞추고 추천 방법의 개선은 향후 연구로 남긴다.

5. 실험

5.1 실험 데이터 수집

본 실험에서 사용한 트위터 데이터는 트위터 스트리밍 API[5]를 이용해서 수집했다. 전체 트윗의 약 1%를 샘플링한 데이터를 받아 그 트윗들에 사용된 해시태그들을 모았다. 이렇게 모은 해시태그들 중에서 알파벳으로 만들어진 해시태그들만 추려냈으며 그 중에서도 영어 단어로 구성된 해시태그들만 실험 데이터로 사용했다.

실험 데이터로 모은 해시태그들 가운데 많이 사용된 해시태그들을 중심으로 각 해시태그를 포함하는 트윗들을 모았다. 하나의 해시태그당 1500개의 트윗들을 모았으며 분류 조건으로 이용되는 URL개수의 평균과 중간 값, 함께 사용된 해시태그 개수의 평균과 중간 값, 타임스탬프의 표준편차, 해시태그의 길이를 계산했다. 또한 해시태그 분류에서 정답으로 사용될 분류 결과 값을 미리 정의하기 위해 해시태그가 사용된 트윗들의 내용을 읽고 이 해시태그가 정보 제공형 해시태그인지 참여 유도형 해시태그인지를 판단했다. 마지막으로 각 분류조건에 엔트로피를 적용하여 적게는 두 개, 많게는 네 개의 구간으로 나누었다.

이렇게 실험에 사용할 데이터로 1000개의 해시태그 레코드들을 만들었으며 그 중에 395개는 정보 제공형 해시태그이고 695개는 참여 유도형 해시태그이다.

5.2 실험 방법 및 결과

실험을 통해 본 논문에서 제안한 해시태그 분류가 얼마나 높은 정확도를 가지는지 알아보았다. 그리고 해시태그 분류에 이용된 분류 조건들의 효과와 분류 조건들의 조합에 따른 분류 결과를 살펴보았다.

표 2 실험 결과

실험	정답	오답	정확도	실험	정답	오답	정확도
1	382	18	0.955	6	375	25	0.938
2	382	18	0.955	7	378	22	0.945
3	381	19	0.953	8	384	16	0.96
4	382	18	0.955	9	385	15	0.963
5	373	27	0.933	10	379	21	0.975
평균	380.1	19.9	0.950				

실험을 위해 600개의 트레이닝 데이터와 400개의 테스트 데이터를 사용했다. 먼저 트레이닝 데이터를 분석해서 분류기를 만들었고 여기에 테스트 데이터를 적용한 결과와 미리 정의해 둔 정답을 비교했다. 임의의 트레이닝 데이터와 테스트 데이터를 구성해서 10번 실험했고 그 평균 값을 실험의 결과로 사용했다. 실험의 결과는 표 2에 정리했다.

해시태그 분류의 평균 정확도는 95%로 높은 결과가 나왔다. 이는 해시태그 사이에 그 특성이 명확히 구분되는 해시태그 유형이 존재함을 증명한다고 볼 수 있다.

해시태그 분류에 이용되는 분류조건들의 효과를 살펴보기 위하여 각각의 분류조건들을 가지고 해시태그를 분류했다. 실험 결과 해시태그를 분류하는데 가장 효과적인 분류 조건은 URL이고 타임스탬프의 표준편차가 가장 비효율적인 분류 조건인 것으로 나타났다. 자세한 실험 결과는 그림 3, 표 3에 정리했다.

트위터에서 URL은 대체로 정보를 제공하기 위한 목적으로 사용되기 때문에 어떤 해시태그를 포함하는 트

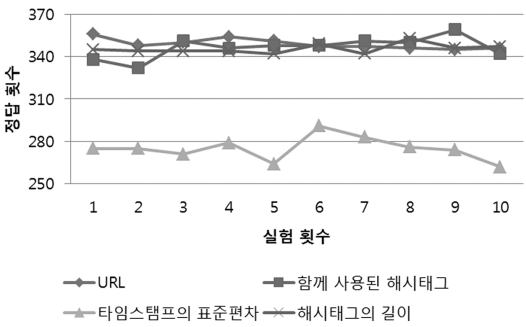


그림 3 해시태그 분류 조건의 효과 - 정답 횟수

표 3 해시태그 분류 조건의 효과 - 정답 평균

	URL	함께 사용된 해시태그	타임스탬프의 표준편차	해시태그의 길이
정답 평균	349	346.5	275	345.6

5) https://dev.twitter.com/docs/streaming-apis

윗들이 URL을 많이 가지고 있다면 그 해시태그는 정보 제공형 해시태그일 확률이 높다. 따라서 URL을 사용했을 때의 분류 결과가 가장 좋게 나타난 것으로 보인다.

타임스탬프의 표준편차는 해시태그 유형의 특성을 분명하게 반영하지 못했다. 정보 제공형 해시태그 타임스탬프의 표준편차는 대체로 높은 반면 참여 유도형 해시태그 타임스탬프의 표준편차는 넓은 범위의 값을 가지고 있었다. 그러므로 참여 유도형 해시태그는 실시간 트렌드에 등장했을 때 폭발적으로 사용되었다가 실시간 트렌드에서 사라지고 나면 사용량이 현저히 줄어들며 경우에 따라 지속적으로 사용되는 해시태그도 있고 금방 사용하지 않게 되는 해시태그도 있다고 보아야 할 것이다.

함께 사용된 해시태그와 해시태그의 길이는 비슷한 효과를 가지는 것처럼 보인다. 그러나 함께 사용된 해시태그는 데이터 구성에 따라 성능이 고르지 못한 반면 해시태그의 길이는 안정적인 성능을 보인다. 따라서 함께 사용된 해시태그 보다 해시태그의 길이가 더 좋은 분류 조건이라고 할 수 있다.

결과적으로, 해시태그 유형을 분류하기 위한 가장 효과적인 분류 조건은 URL이며 그 다음으로는 해시태그의 길이, 함께 사용된 해시태그 순이고 타임스탬프의 표준편차가 가장 비효율적인 분류 조건인 것으로 나타났다.

마지막으로 분류 조건들의 결합 효과를 보기 위해 모든 조합에 대한 분류 결과를 살펴보았다. 그림 4에서처럼 URL이 포함된 분류 조건의 조합이 대체로 좋은 결과를 나타내었고 그 중에서도 [URL, 함께 사용된 해시태그, 해시태그의 길이] 조합이 가장 우수한 성능을 보였으며 이는 모든 분류 조건을 다 사용한 조합의 실험과 비슷한 결과를 보였다. 그리고 해시태그 분류 조건의 효과를 입증한 실험결과와 동일하게 [함께 사용된 해시태그, 타임스탬프의 표준편차] 조합이 가장 좋지 않은 분류 결과를 보였다.

[URL, 해시태그의 길이] 조합에 [함께 사용된 해시태그]가 사용된 것과 [URL, 해시태그의 길이] 조합에 [타임스탬프의 표준편차]가 사용된 것의 성능 향상을 비교해보면 [함께 사용된 해시태그]가 [타임스탬프의 표준편차]보다 조금 더 효과적인 분류 조건이라고 할 수 있다. 그러나 [URL, 해시태그의 길이] 조합에 [함께 사용된 해시태그, 타임스탬프의 표준편차]를 사용하는 것의 성능 향상을 비교해보면 [함께 사용된 해시태그]와 [타임스탬프의 표준편차]의 결합은 좋은 시너지 효과를 일으키지 못한다는 것을 알 수 있다.

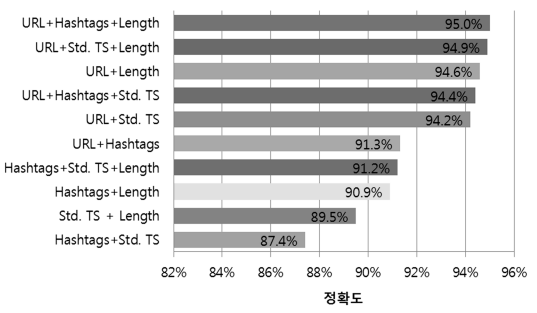


그림 4 모든 조합에 대한 실험결과

이렇게 다소 아쉬운 실험 결과를 보완하기 위해서 타임스탬프의 표준편차를 대신할 좋은 분류 조건을 찾아야 할 것으로 보인다. 이는 참여 유도형 해시태그의 사용 추세를 반영하는 통계적 수치를 이용하면 좋은 효과를 얻을 수 있을 것으로 기대된다.

6. 결론 및 향후 연구

본 논문에서는 트위터의 해시태그 추천을 개선하기 위한 해시태그 분류 기법을 제안했다. 해시태그에는 그 특성과 사용된 의도가 분명하게 다른 두 가지 유형이 있다. 하나는 사용자들에게 정보를 주기 위해 사용되는 정보 제공형 해시태그(Informative Hashtag)이고 다른 하나는 사용자들로 하여금 마이크로-미미(Micro-meme)에 참여하도록 하는 참여 유도형 해시태그(Meme Hashtag)이다. 해시태그를 유형에 따라 분류하기 위하여 나이브 베이지안 분류 방법을 적용했으며 분류 조건들로는 해시태그가 포함된 트윗들이 가지고 있는 URL의 개수, 함께 사용된 다른 해시태그들의 개수, 해시태그 타임스탬프의 표준편차, 해시태그의 길이를 사용했다. 실험을 통해 본 논문에서 제안한 분류 조건을 이용해서 해시태그를 분류했을 때 95%의 높은 정확도를 얻었고 해시태그 분류에 가장 큰 영향을 미치는 분류 조건은 URL의 개수와 해시태그의 길이라는 것을 밝혔다.

향후 연구로는 해시태그의 유형을 다양하게 정의해서 해시태그의 분류를 좀 더 풍성하게 개선하고 해시태그 유형간의 관계를 분석하여 해시태그를 더 잘 분류할 수 있는 분류조건을 찾아야 할 것이다. 또한 해시태그 유형별로 다르게 적용되어야 할 추천 알고리즘에 대한 연구가 필요하다.

참 고 문 헌

[1] J. Huang, K. M. Thornton, and E. N. Efthimiadis, "Conversational tagging in twitter," *Proc. of the 21st ACM conference on Hypertext and hypermedia HT 10*, pp.173-178, 2010.

[2] L. Yang, T. Sun, M. Zhang, and Q. Mei, "We Know What @ You # Tag : Does the Dual Role Affect Hashtag Adoption?," *Proc. of the 21st international conference on World Wide Web*, pp.261-270, 2012.

[3] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical Clustering of Tweets," *Proc. of the 3rd Workshop on Social Web Search and Mining (SWSM)*, 2011.
[4] D. Laniado and P. Mika, "Making sense of Twitter," *Proc. of the Semantic Web - ISWC*, vol.6496, pp.470-485, 2010.
[5] A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic, "The Party is Over Here : Structure and Content in the 2010 Election," *Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, vol.161, no.3, pp.201-208, 2011.
[6] B. Sriram, D. Fuhry, E. Demir, and H. Ferhatosmanoglu, "Short Text Classification in Twitter to Improve Information Filtering," *Proc. of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp.841-842, 2010.
[7] A. Zubiaga, D. Spina, and R. Martínez, "Classifying Trending Topics : A Typology of Conversation Triggers on Twitter," *Proc. of the 20th ACM international conference on Information and knowledge managemen*, pp.2461-2464, 2011.
[8] T. Mitchell, Machine Learning, pp.177-178, McGraw Hill, 1997.
[9] P. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley, pp.158-164, 2006.

김 혜 원
2011년 숙명여자대학교 컴퓨터과학과 학사. 2011년~현재 서울대학교 컴퓨터공학부 석사과정 재학 중. 관심분야는 소셜 네트워크, 빅데이터 처리

김 형 주
1982년 서울대학교 전산학과(학사). 1985년 Univ. of Texas at Austin(석사) 1988년 Univ. of Texas at Austin(박사) 1988년~1990년 Georgia Institute of Technology(조교수). 1991년~현재 서울대학교 컴퓨터공학부 교수. 관심분야는 데이타베이스, XML, 시맨틱 웹, 온톨로지

김 기 성
2003년 서울대학교 응용화학부 학사. 2006년~2009년 티맥스소프트. 2003년~현재 서울대학교 컴퓨터공학부 박사과정 재학 중. 관심분야는 데이터베이스, 시맨틱 웹, 분산 병렬 데이터 처리