토픽 기반의 트윗 분류를 위한 해시태그 분석 기법

김용성°, 전상훈, 유제혁, 황인준
고려대학교 전기전자공학과
{kys1001, ysbhjun, rjh1026, ehwang04}@korea.ac.kr

Hashtag Analysis Scheme for Topic based Tweet Categorization

Yongsung Kim°, Sanghoon Jun, Jehyeok Rew, Eenjun Hwang
School of Electrical Engineering, Korea University

요 약

최근 SNS 사용자가 급증하면서 매우 다양하고 방대한 양의 글이 여러 종류의 SNS를 통해 생성되고 있다. 그중 트위터는 정보의 전달 및 확산에 상당히 유용한 도구로 사용되고 있다. 이러한 트위터의 사용자 트윗은 뉴스, 음악, 사진, 여행 등 다양한 형태로 등장한다. 또한 트위터는 해시태그라는 사용자 정의 태그를 사용하는데 이는 트윗의 키워드 및 핵심을 쉽게 표현할 수 있도록 해주는 효과적인 수단이다. 최근 상당히 많은 양의 트윗의 생성에도 불구하고 이를 다양한 카테고리별로 분류할 수 있는 연구가 많이 진행되지 않았다. 따라서 본 논문에서는 해시태그를 이용해 트윗의 핵심을 파악하고 수많은 트윗을 다양한 토픽별로 분류할 수 있는 기법을 제안한다. 우선 다양한 카테고리의 인기 해시태그가 포함된 트윗을 수집하고 수집한 트윗에서 해시태그별 키워드를 추출한다. 그리고 코사인 유사도를 통해 해시태그별 내용 유사도를 파악하여 각 카테고리 내의 해시태그가 얼마나 유사한 내용을 지니고 있는지 파악한다. 마지막으로 사용자 트윗이 입력되면 모든 카테고리와 유사도를 비교하여 가장 유사도가 높은 카테고리를 찾아 추천해준다. 제안된 기법을 바탕으로 프로토타입을 구현하고 실험을 통해 성능을 평가한다.

1. 서론

최근 SNS(Social Network Service) 사용자가 급증하면서 매우 다양하고 방대한 양의 글들이 Facebook, Twitter, Instagram 등의 SNS를 통해 생성되고 있다. 그중 Twitter는 대표적인 SNS서비스로서 정보의 전달 및 확산에 상당히 유용한 도구로 사용되고 있다. 트위터는 뉴스의 전달, 좋아하는 음악 및 영상의 공유, 좋아하는 여행지의 공유 등 다양한 형태의 정보 전달의 도구로 사용된다.

트위터의 해시태그는 트위터의 핵심적인 기능 중 하나로 사용자 정의 태그를 의미한다. 트위터는 140자 이내의 트윗을 작성해야 하므로 짧은 문장에 트윗의 핵심을 표현해야한다. 이를 위해 해시태그를 사용하는데 이는 트윗의 핵심 및 키워드를 쉽게 표현할 수 있도록 하는 효과적인 수단으로 사용되고 있다. 이러한 해시태그를 분석하게 되면 여러 트윗의 주제를 쉽게 파악하고 이에 따라 분류할 수 있다.

위에서 언급한 것과 같이 트위터에는 뉴스, 음악, 영상, 여행 등과 같은 매우 다양한 유형의 컨텐츠들이 등장한다. 하지만 상당히 많은 양의 트윗을 다양한 토픽별로 분류할 수 있는 연구가 많지 않다. 사용자 트윗과 가장 유사한 정보를 가진 카테고리를 파악하게 되면 많은 양의 트윗을 카테고리별로 쉽게 분류할 수 있어 트위터를 이용한 마케팅 및 광고, 팔로워 추천 등 다양한 분야에 사용될 수 있다.

따라서 본 논문에서는 해시태그를 이용해 트윗의 핵심을 찾아 수많은 트윗을 다양한 토픽별로 분류할 수 있는 기법을 제안한다. 우선 다양한 카테고리의 인기 해시태그가 포함된 트윗을 수집하고 해시태그별로 수집한 트윗에서 키워드를 추출한다. 그리고 코사인 유사도를 기반으로 해시태그별 내용 유사도를 파악한다. 마지막으로 사용자 트윗이 입력되면 모든 카테고리와 유사도를 비교하여 가장 유사도가 높은 카테고리를 찾아 추천해준다.

2. 관련 연구

기존에 트위터 해시태그와 카테고리 분류 기법에 관한 연구가 진행되어 왔는데 [1]에서는 가장 많이 사용되는 해시태그를 기반으로 클러스터링 된 4개의 주요 토픽을 기반으로 정보 확산 특성을 파악하였고, [2]에서는 트위터 내용에 대하여 해시태그와 일반 용어 각각에 대하여 k-means clustering을 활용하고 상관관계를 분석하였다. [3]에서는 다양한 회사명(브랜드명)과 관련된 트윗 검색을 할 때 Apple과 같은 모호한 단어를 분류하여 회사명(브랜

드명)이 언급된 트윗만 효율적으로 분류해내는 기법을 제안하였다. [4]에서는 다양한 언어로 된 트윗을 대표 4개의 카테고리로 분류할 수 있는 기법을 제안하였다.

3. 토픽 기반의 트윗 분류를 위한 해시태그 분석 기법

토픽 기반의 트윗 분류를 위한 해시태그 분석 기법을 위해 본 논문에서 제시하는 전체적인 시스템 구성은 그림 1과 같다.

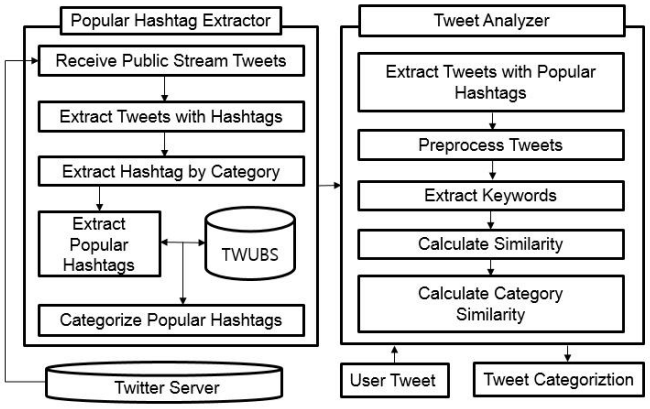

그림 1 시스템 구성

3.1 인기 해시태그 추출

'TWUBS'[4]라는 웹사이트는 다양한 해시태그를 카테고리별로 분류해주며 사용자가 직접 특정 카테고리에 해시태그를 등록할 수 있다. 본 논문에서는 위에서 분류된 카테고리 중 유사한 항목이 있는 카테고리를 제외하고 9개의 카테고리를 이용하며 자세한 내용은 표 1과 같다. 그리고 해당 웹사이트의 각 카테고리에 있는 해시태그를 모두 크롤링하여 수집한다.

인기 해시태그를 추출하기 위해 먼저 Twitter API[1]를 활용하여 무작위 트윗을 수집한다. 수집된 트윗에서 각 카테고리명으로 된 해시태그가 포함된 트윗을 분석한다. 예를 들면 Art카테고리의 경우 #Art라는 해시태그가 포함된 트윗을 수집한다. 수집된 트윗에서 해시태그만을 추출하고 가장 많은 빈도수를 차지하는 해시태그를 선별한다. 선별된 내용이 해당 카테고리에 적합한지 여부는 'TWUBS' 사이트에 카테고리별로 분류되어 있는 해시태그와 비교를 하여 해당 카테고리에 포함된 해시태그라면 인기 해시태그로 추출하고 그렇지 않으면 제외한다. 이렇게 카테고리별 총 40개의 해시태그를 추출한다. 예를 들면 #Art가 포함된 트윗에서 추출한 해시태그가 #artist, #painting, #design 등이라면 웹사이트에 미리 정의된 총 8656개의 해시태그와 비교하여 분류에 포함되어 있으면 인기 해시태그로 가정하고 표 1과 같이 이를 추출한다.

Category	추출된 인기 해시태그 예시
Art	#fashion, #art, #photography, #photo, #design, #painting, #fineart, #handmade
Game	#gameinsight, #xbox, #androidgames, #iphonegames, #ps4, gameinsightbr
Health	#health, #food, #run, #family, #fitness, #beauty, #gym, #healthcare, #medical
Movie	#video, #music, #movie, #tv, #youtube, #watch, #film, #drama, #action, #dvd
Music	#nowplaying, #np, #dj, #itunes, #pop, #soundcloud, #edm, #indie, #jazz, #mix
News	#un, #job, #israel, #news, #tech, #gaza, #usa, #isis, #stocks, #ebola, #iphone
Politics	#obama, #usa, #palestain, #iran, #pjnet, #news, #gaza, #hamas, #obamacare
Sports	#football, #nba, #fitness, #athlete, #mlb, #workout, #baseball, #football, #nfl
Technology	#android, #ipad, #iphone, #tech, #app, #update, #science, #mobileapps, #mac

표 1 인기 해시태그 추출 예시

3.2 인기 해시태그 관련 트윗 수집 및 분석

3.2절에서 수집한 각 카테고리별 Top40 해시태그가 포함된 트윗을 모두 수집한다. 예를 들면, 음악 관련 해시태그인 #nowplaying이 포함된 트윗을 모두 수집하여 하나의 문서로 가정한다. 그리고 해당 문서의 키워드를 추출한다. 키워드를 추출하기 전에 수집한 트윗에 대한 전처리 과정이 필요하다. 가장 먼저 영어권 트윗을 대상으로 실험을 진행하기 위해 한국어, 중국어, 일본어 등의 기타 언어는 모두 제거하고 영어로 구성된 트윗만 추출한다. 그리고 대소문자로 인해 같은 단어가 다르게 인식되는 경우를 제외하기 위해 모든 트윗을 소문자로 변경한다. 마지막으로 각 문서별로 tf-idf(3)를 적용하여 표 2와 같이 키워드를 추출한다. 아래의 식(1)에서 f는 term frequency를 나타내며, f는 특정 용어의 빈도수, d는 document, t는 term을 나타낸다. 식(2)의 idf의 경우는 inverse document frequency를 나타내며, 은 총 document의 개수, $|d \in D : t \in d|$는 term t가 등장하는 document의 개수를 의미한다. 식(3)은 식(1)과 (2)의 곱의 형태로 나타내며 이를 통해 문서의 키워드를 손쉽게 추출할 수 있다.

$$tf_{,d} = \log(1 + f_{t,d}) \quad (1)$$

$$idf_{t,} = \log \frac{|N|}{|d \in D : t \in d|} \quad (2)$$

$$tfidf_{t,d,D} = tf_{t,d} \times idf_{t,D} \quad (3)$$

각 대표 해시태그별로 선택된 키워드 예시는 표 2와 같다.

해시태그	해시태그별 키워드 예시
#game	iphonegames, tribez, castlez, android gameinsight, ipadgames, gameinsightbr
#now playing	listenlive, tune, radio, listen, bruno, gaga, itunes, breakingnews, love, live, remix
#hiphop	reverbnation, nowplaying, rap, graffiti, indie, itunes, artise, rnb, radio, djmarcd
#fashion	ap_fashion, ldnfashion, timesfashion, img, hexagon, milano, moda, jewelry, models
#photo	photography, photo, celebrityphotos, art, cute, travel, instagood, smile, camera
#hot	movie, iphone, clip, travel, click, video, hot, party, hottest, model, promotion
#tv	tvtag, sticker, wactching, tv, voice, love, join, fans, vote, ng, episode, watch, tvd
#mac	itunes, leather, air, app, video, iphone, macbook, mac, paid, retina, machine

표 2 해시태그별 트윗 문서의 키워드 예시

3.3 트윗 기반 유사도 판별 및 평균 계산

3.2절에서 얻은 Term Frequency 값을 기반으로 모든 문서간 코사인 유사도(4)를 계산한다. 식(4)에서 A, B의 값은 각 document의 tf(term frequency)를 사용한다. 또한 유사도의 값은 0에서 1사이의 값이 나오며 값이 클수록 문서간의 유사도가 높은 것이다. 코사인 유사도를 적용하여 각 해시태그별 문서의 유사도를 파악한다.

$$Similality \quad \cos\theta = \frac{A \quad B}{\| \quad \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sum_{=1}^{n} (A_i)^2 \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (4)$$

사용자가 트윗을 입력하게 되면 트윗 내용을 분석하여 해시태그별 문서와 코사인 유사도 기법을 적용한다. 또한 각 카테고리별 유사도 평균을 구한다. 예를 들면, 사용자 입력 트윗과 각 카테고리별 40개의 문서, 즉 총 360개의 문서와 유사도 비교를 진행한다. 그리고 식(5)와 같이 각 각의 유사도를 모두 더해 평균을 낸다. 여기서 S는 유사도, n은 총 문서의 개수, a는 특정 카테고리를 나타낸다.

$$average_a = \frac{\sum_{i=1}^{n} S_{i,a}}{n_a} \quad (5)$$

각 카테고리별로 유사도 평균을 내게 되면 사용자 입력 트윗과 가장 유사한 카테고리가 무엇인지 쉽게 파악할 수 있다. 그러므로 사용자 트윗만을 가지고 해당 트윗이 어떤 카테고리에 속해 있는지 쉽게 유추할 수 있다. 이를 통해 트윗 내용 및 분류 파악이 매우 쉬워질 뿐 아니라 사용자 트윗을 기존의 트윗들과 비교하여 가장 유사한 컨텐츠 분류를 찾아낼 수 있어 유사 성향의 트위터 팔로워를 찾는 데도 많은 도움을 줄 것이다.

4. 실험 결과

Twitter API[5]에서 Twitter4J[6] 라이브러리를 활용해 2014년 7월 18일부터 8월 1일까지 총 64,340,000개의 트윗을 수집하였고 해시태그가 하나라도 포함되어 있는 트윗 5,975,700개를 대상으로 실험을 진행하였다. 수집된 트윗을 기반으로 3절의 그림 1과 같이 시스템 구성을 하였다. 사용자가 트윗을 입력하게 되면 트윗을 용어 단위로 나누어 유사도를 분석하고 가장 유사도가 높은 카테고리를 추천하여 준다. 표 3에서와 같이 사용자가 트윗을 입력하게 되면 기존에 분류된 문서와 유사도를 비교하게 되고 각 카테고리별 유사도의 평균을 내어 가장 높은 값을 갖는 카테고리를 선택할 수 있도록 해준다. 그림 2에는 사용자 입력 트윗과 모든 카테고리별 유사도 평균값을 나타내었다. 실험은 음악과 관련된 내용의 사용자 트윗이 입력되면 다른 해시태그별 문서와 비교를 하고 유사도 평균을 낸다. 그리고 사용자 트윗의 입력 내용과 유사도 평균값이 가장 높은 카테고리인 음악 카테고리를 추천하여 준다.

사용자 입력 트윗				
#nowplaying cornelis gerard - stop being so hot #ska #punk #pop #music on the #1 #variety #radio station #bwdradio #np				
해시태그 기반 문서와 유사도 비교				
Art	Game	health	Movie	Music
0.129117	0.147729	0.152603	0.422868	0.208857
0	0.136519	0.249577	0.241582	0.129117
0.12483	0.179856	0.228809	0.129117	0.28068
0.071351	0.212089	0.166775	0.22471	0.262696
0.185344	0.090198	0.074392	0.153259	0.314337
0.206426	0.111356	0.235862	0.188663	0.148038
0.169536	0.119172	0.177188	0.319119	0.301802
0.029973	0.178695	0.148038	0.226335	0.290717
0.165577	0.128472	0.203533	0.199327	0.263915
0.139265	0.026024	0.145382	0.235177	0.167161
0.123723	0.119103	0.215989	0.195908	0.268135
0.178688	0.148919	0.189317	0.207436	0.373809
0.15087	0.130395	0.242595	0.267426	0.155254
0.136048	0.121053	0.186219	0.188186	0.350836
0.109974	0.177213	0.198467	0.1869	0.236782
0.148131	0.11565	0.194607	0.215336	0.254792
0.172393	0.207398	0.208798	0.172378	0.371125
0.126219	0.14832	0.219246	0.207398	0.248444
0.147889	0.134557	0.157942	0.167161	0.225305
0.116948	0.112658	0.185992	0.268135	0.244276
0.033117	0.167161	0.184882	0.254792	0.268198
0.148038	0.113659	0.192796	0.152759	0.450546
0.116697	0.187143	0.211827	0.225305	0
0.129196	0	0.222293	0.308299	0.039464
0.11196	0.254792	0.176479	0.218437	0.439
...	...	...	...	...
유사도 평균				
0.123441	0.139218	0.189693	0.221699	0.265314
추천 카테고리 : Music				

표 3 사용자 입력 트윗과의 유사도 및 유사도 평균

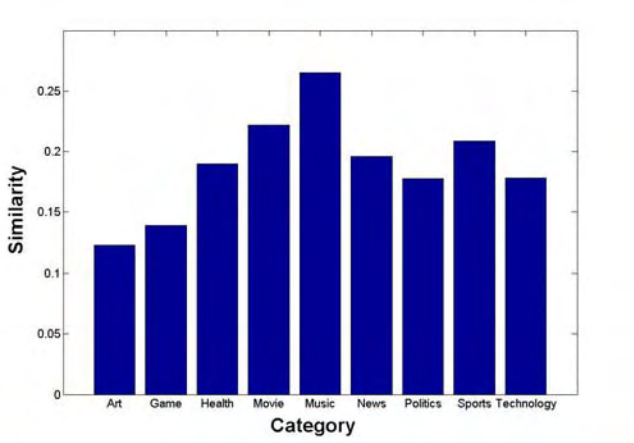

그림 2 카테고리별 유사도 평균

5. 결론 및 향후 연구 방향

본 논문에서는 토픽 기반의 트윗 분류를 위한 해시태그 분석 기법을 제안하였다. 트위터에서 카테고리별 인기 해시태그가 포함된 트윗을 수집하고 이를 분석하여 사용자가 남긴 트윗과 가장 유사한 카테고리를 유추할 수 있었다. 이를 통해 많은 양의 트윗을 카테고리별로 쉽게 분류할 수 있어 트위터를 이용한 마케팅 및 광고, 유사 성향의 팔로워 추천, 유사 트윗 파악 등 다양한 분야에 사용될 수 있을 것이다.

향후 연구에서는 트위터 해시태그 기반 토픽을 더욱 세분화하고 트윗 유사도 비교에 사용되는 해시태그 기반 문서를 트위터 뿐만 아니라 다양한 웹사이트 문서 분석을 통해 트윗 분류 정확도를 향상시키는 연구를 진행할 예정이다.

6. Acknowledge

본 연구는 미래창조과학부 및 정보통신산업진흥원의 대학 ICT연구센터 육성·지원사업(NIPA-2014-H0301-14-1001)과 2013년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2013R1A1A2012627)

Reference

[1] Geerajit Rattanaritnont, et al, "Characterizing topic-specific hashtag cascade in twitter based on distributions of user influence", 14th Asia-Pacific Web Conference, p735-742, 2012.

[2] CI Muntean, et al, "Exploring the Meaning behind Twitter Hashtags through Clustering", BIS 2012 International Workshops and Future Internet Symposium, p231-242, 2012.

[3] Minoru Yoshida, et al, "ITC-UT: Tweet Categorization by Query Categorization", CLEF 2010 Labs and Workshops Notebook Papers, 2010.

[4] K Nishida, et al, "Tweet classification by data compression", international workshop on DETecting and Exploiting Cultural diversity on the social web, p29-34, 2010.

[4] http://twubs.com/

[5] Twitter API, http://dev.twitter.com/docs/using-search/

[6] Twitter4J, http://twitter4j.org/