트위터 분석을 이용한 카테고리별 실시간 트렌드 추출 기법

나병진, 김용성, 황인준
고려대학교 전기전자공학과
e-mail:nbjin0208@korea.ac.kr

Real-time Category Trend Extraction Scheme based on Twitter Analysis

ByeongJin Na, YongSung Kim, EenJun Hwang
School of Electrical Engineering, Korea University

요 약

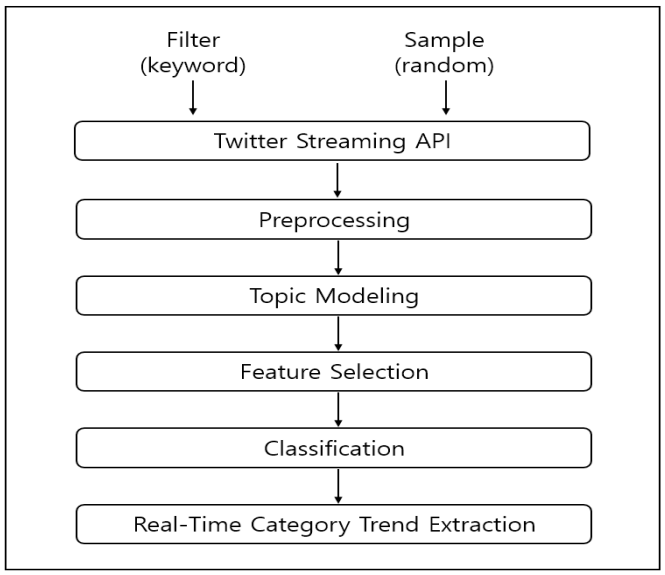
최근 소셜 네트워크 서비스상의 데이터를 실시간으로 분석하여 의미있는 정보를 찾아내기 위한 연구가 활발하게 진행되고 있다. 특히, 스마트폰과 같은 스마트 디바이스를 이용하는 많은 사용자들이 실시간으로 발생하는 이벤트를 소셜 네트워크상에 게재하고 서로 공유하면서, 대중들이 관심을 가지는 토픽의 경우 굉장히 빠르게 확산되는 경향을 보이고 있다. 본 논문에서는 이러한 SNS의 특성을 토대로 트위터상의 트윗을 분석하여 여러 분야의 토픽들을 카테고리별로 분류하고, 카테고리별 트렌드를 추출하여 실시간으로 시각화하는 기법을 제안한다. 이를 위해, 트위터를 기반으로 SVM 분류 알고리즘과 Twitter-LDA를 통하여 트윗을 분야별로 분류하고, 각각의 트렌드를 이루는 대표적인 키워드를 선출하여 이를 기반으로 실시간 트렌드를 추출한다. 제안하는 기법의 성능을 평가하기 위해, 분류 특징 선택의 신뢰도를 측정한다.

1. 서론

최근 소셜 네트워크 서비스(SNS)의 발달과 더불어 많은 사람들이 SNS를 활용하여 일상을 게시하고 의견과 정보를 공유하고 있다. 일반적으로 SNS에서 공유되는 게시글은 분야를 막론하고 광범위한 내용을 내포하고 있다. 특히, 스마트폰과 같은 스마트 디바이스를 이용하는 대다수의 사용자는 이벤트가 발생하면 실시간으로 SNS에 글을 게시한다. 사용자들의 이러한 자발적인 활동으로 정보를 일방적으로 전달하는 기존의 언론매체들보다 훨씬 빠르고 효율적으로 실시간 이벤트를 공유할 수 있게 된다.

현재까지 다양한 SNS가 등장했고, 그 중에서도 글로벌 사용자 수가 가장 많은 SNS 중 하나인 트위터[1]는 사용자가 트윗(Tweet)이라고 하는 메시지를 트위터에 게시하고, 다른 사용자의 트윗을 리트윗(Retweet)하여 트윗의 정보를 확산시킨다. 트윗은 140자 이내의 짧은 메시지만을 작성해야 하는 한계로 인해 효과적인 정보 전달을 위해 핵심 및 키워드 위주로 작성된다. 이러한 특징은 키워드 필터를 통하여 수월하게 관련된 내용의 트윗을 수집할 수 있게 한다.

따라서 본 논문에서는 SNS의 실시간 특징과 핵심 및 키워드 위주의 트윗 작성을 이용하여 실시간으로 여러 분야에서의 트렌드를 분석하는 카테고리별 실시간 트렌드를 추출하는 기법을 제안한다. 전체적인 시스템의 구조는 (그림1)과 같이 구성된다. 먼저 키워드 기반의 트윗 수집을 통하여 카테고리를 분류할 수 있는 특징을 추출하고, SVM 분류기를 이용해 카테고리별 트윗을 훈련시킨다. 그 후에 지도 학습된 분류기를 이용하여 무작위 수집을 통한 트윗의 카테고리를 분류하고, 실시간 트렌드를 차트에 그린다. 2장에서는 관련 연구를, 3장에서는 제안하는 기법을 상세히 기술한다. 4장에서는 실험 결과 및 분석을 기술하며, 5장에서 결론 및 향후 계획을 기술한다.



<그림 1> 시스템 구조

2. 관련 연구

일반적으로 Latent Dirichlet Allocation (LDA)[2]는 토픽 모델링 분야에서 가장 보편적이고 효과적인 도구로 알려져 있다. LDA를 확장시킨 기법들은 다방면에서 활용되고 있으며, SNS 분석에도 활용되어지고 있다. 그 중에서도 [3]에서는 트위터에서 기존의 전통적인 LDA 기법이 트윗과 같은 단문의 특성을 고려하지 않는 점을 보완한 Twitter-LDA를 기술하였다. Twitter-LDA가 소개된 이후로 트위터를 분석하는 많은 연구에서 사용되었으며, [4]의 경우에서도 Twitter-LDA를 통하여 주제적인 핵심문장을 추출하는 연구를 진행하였다. 본 연구에서 제안하는 시스템에서도 Twitter-LDA를 이용하여 분류 특성을 추출하는 기법을 제안한다.

실시간 SNS 분석을 통한 트렌드 검출 연구로는 [5]에서 급상승하는 키워드를 감지하는 실시간 트위터 모니터링 시스템이 있다. [6]는 스트림 데이터에 TF-IDF 기법을 적용하여 트렌드를 분석했으며, [7]은 트위터 사용자들이 일종의 센서가 되어 실시간으로 지진을 감지를 해내는 시스템을 제안하였다. 또한, [8]는 트위터 내의 뉴스 토픽의 분석에 초점을 맞춰 트렌드를 분석하는 연구를 진행하였다.

3. 제안 방법

본 연구는 Twitter-LDA와 SVM 분류기를 이용하여 특징 추출 및 카테고리별 트윗을 지도 학습시키고, 무작위의 실시간 트윗이 주어질 때 카테고리별 실시간 트렌드를 추출하여 부분 및 전체적인 트렌드를 분석하는 것을 목표로 한다.

3.1 트윗 수집 및 전처리

트렌드 추출에 앞서, 무작위로 수집되는 트윗을 각각의 카테고리로 분류해주는 작업이 필요하다. 분류 성능을 높이기 위해 특징 추출에 용이하게끔 전처리 과정을 거친 뒤, SVM 알고리즘을 활용한 지도 학습을 통해 트윗을 분류한다. 본 논문이 제안하는 시스템에서는 Java 기반의 트위터 스트림 API 라이브러리인 Twitter4J[9]를 사용하여 분류기의 트레이닝 셋을 구성하기 위한 트윗을 수집하였다. 총 10개의 카테고리를 표 1과 같이 미리 선정하였으며, 각 카테고리별로 키워드 기반의 수집을 통해 집중된 주제를 가지는 트윗을 수집하였다. 10개의 카테고리는 구글 트렌드[10]의 25개의 카테고리 중에서 유사한 카테고리는 결합하고, 여러 카테고리에 전반적으로 나타날 수 있는 카테고리는 제거하여 10개의 카테고리로 압축하였다. 각 카테고리별 해당 키워드들은 2006년부터 모든 트윗을 저장하고 있는 Topcy의 API[11]를 이용하여 수집하였다. Topcy search API로 카테고리명이 포함된 트윗을 검색하여 Topcy내 자체 순위에서 가장 높은 순위의 상위 1000개 트윗 중 빈도수가 높은 단어 위주로 선택하였다. 모든 트윗은 영문 트윗이며, 2015년 9월 6일 오전 0시부터 10일 오전 0시까지 총 100만 개의 트윗을 수집하였다.

<표 1> 분류 카테고리 및 키워드

Categories	Keywords
Art&Beauty	art, beauty, style, hair, natural, skin, fashion, face, daily, care, makeup, model, jewelry, nail
Book	book, literature, storytelling, review, series, kindle, page, wattpad, copy, comic, covers
Business	business, industry, market, job, marketing, insider, customer, development, build, sale, owner
Computer & Science	computer, electronic, pc, windows, laptop, core, wifi, digital, driver, phone, battery, smart, iot, it
Entertainment & game	entertainment, music, movie, comedy, theater, dance, performance, play, video game, online game, mobile game
Food & Shopping	food, drink, eat, plate, wine, cook, diet, junk, water, coffee, beer, tea, alcohol, juice, restaurant, shopping
Health & Sports	health, fitness, nutrition, body, weight, exercise, muscle, benefit, insurance, sports, athletic, stadium, team
Law	law, government, enforcement, court, federal, president, vote, police, gun, legal, sign, judge, abortion, public, drug
Home & Pet	home, garden, squar, grow, plant, green, house, flower, diy, vegetable, indoor, furniture, homemade, pet, animal
Travel	travel, hotel, ttot, trip, adventure, tourism, tourist, travelnews, traveling, historic site

수집된 트윗들은 자연어로 이루어진 단문들이기 때문에 특징 추출과 토픽 모델링의 적용을 위해서는 전처리 과정이 필요하다. 이를 위해 이모티콘과 같은 무의미한 표현과 불용어, 외부로 연결되는 링크를 제거하는 토큰화를 수행하였다.

3.2 특징 추출 및 카테고리 분류

문서의 특징 추출에는 단어의 빈도수를 이용한 기법을 일반적으로 사용한다. 자주 등장하는 단어들이 문서 내에 포함되는지의 여부를 판단하여 0과 1의 값을 가지는 이진 특징 벡터로 표현이 된다. 하지만 본 논문에서는 특징 추출에 Twitter-LDA를 사용하여 각각의 카테고리별 주제를 나타낼 확률이 높은 상위 150개 단어들을 카테고리별 특징 후보로 선정하였다. 선정된 후보 집합을 통합하고, 중복 제거를 통해 하나의 특징 벡터 집합을 생성하였다.

카테고리 분류를 위한 기법은 대표적인 분류 알고리즘인 Support Vector Machine(SVM)[12]을 사용한다. SVM은 지도 기계 학습 (Supervised machine learning) 기법의 하나로, 주로 분류와 회귀 분석에 사용된다. SVM의 구현은 자바 라이브러리인 libSVM[13]을 이용하였다.

보다 좋은 트렌드를 추출하기 위해서 필요한 요소 중 하나는 적절한 시간 간격의 설정이다. 실시간으로 트렌드를 검출하려는 목적을 벗어나지 않으면서 다수의 주제를

선정할 수 있는 시간 간격을 설정할 필요가 있다. 일반적으로 5분동안 5천개 이상의 트윗의 수집이 이루어지기 때문에 본 연구에서는 시간 간격을 5분으로 설정하였다. 5분 간격으로 나누어진 트윗은 Twitter-LDA를 통해 10개의 주제를 정하고, 각각의 주제는 학습된 분류기를 통해 카테고리의 라벨링이 이루어진다. 카테고리로 분류된 주제들은 시간대비 해당 주제에 포함된 트윗의 개수를 포인트로 하여 시각화한다.

3.3 트렌드 추출 및 대표 키워드 생성

매 5분마다 수집한 트윗을 바로 나타낼 경우 주제의 변화가 심하여 안정적인 트렌드를 추출할 수 없다. 그렇기 때문에 시각화하는 단계에서의 시간 간격은 1시간으로 설정하여 유의미한 트렌드를 추출할 수 있게 한다. 이 때에 시간은 x축, 분류된 트윗들의 개수는 각각의 카테고리별 y값이 된다. 하나의 평면에 10개 카테고리를 나타내어 카테고리별 흐름을 한눈에 볼 수 있도록 하였다.

카테고리별로 분류된 트윗들은 Twitter-LDA를 통해 주제를 정하고, 주제일 확률이 가장 높은 상위 5개의 단어의 집합을 대표 키워드로 선출하였다. 선출된 대표 키워드는 시각화된 트렌드에 부가적인 정보로 포함된다.

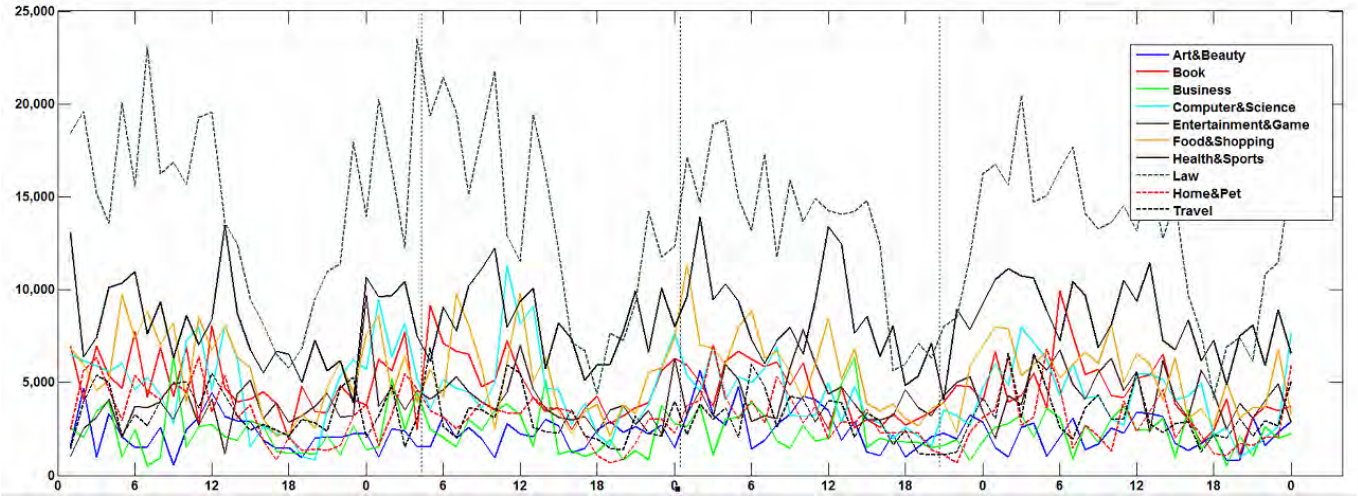
4. 실험 결과

필터 기반으로 수집된 트윗은 카테고리별로 10만개를 사용해 933개의 특징을 추출하였고, 추출된 특징을 이용해 트레이닝을 진행하였다. 분류에 쓰일 무작위 트윗은 트위터 스트림 API 라이브러리를 이용하여 전세계 공개 트윗 중 1%의 무작위 트윗 샘플을 수집하였다. 분류에 앞서, 분류의 정확도 향상 및 Over-fitting을 방지하기 위하여 k-fold cross validation[14]을 사용하였다. k-fold cross validation은 수집된 샘플을 k개의 서브샘플로 나누어 하나의 서브샘플이 검증을 위한 테스트 셋, 나머지 서브샘플들이 트레이닝 셋으로 사용된다. 모든 서브샘플이 테스트 셋으로 한 번씩 사용될 때까지 k번 반복된다. 본 연구에서는 총 필터 기반으로 수집된 19990개의 트윗을 이용하여 k값을 10으로 두고 검증을 진행하였고, 그 결과는 표 2와 같다. 총 10개 중 9개의 카테고리에서 높은 수준의 정확도와 재현율을 보였다. 다만 Travel 분야에서 정확도가 절반정도의 수준을 보이는 것을 확인할 수 있다. 이는 다른 카테고리에서 전반적으로 나올 수 있는 단어들이 특징으로 선택되어 분류의 정확도가 낮아진 것으로 판단된다.

<표 2> k-fold cross validation 결과

Category	Precision	Recall
Art&Beauty	0.907	0.801
Book	0.907	0.837
Business	0.876	0.816
Computer	0.968	0.872
Entertainment&Game	0.971	0.925
Food&Shopping	0.846	0.857
Health&Sports	0.858	0.704
Law	0.878	0.895
Pet	0.965	0.887
Travel	0.563	0.92

그림 2는 2015년 9월 3일 0시부터 6일 24시까지 수집된 무작위 트윗을 한 시간 간격을 기준으로 생성한 실시간 트렌드이다. 그림 2에서 볼 수 있듯이 전체적으로 Law 분야의 주제를 다루는 트윗이 많음을 알 수 있다. 또한, 첫 번째 날 많은 트윗이 발생했던 6시에서 12시 사이의 대표 키워드는 damage, heard, shootings, north, police 등이었다. 실제로 9월 1일 미국 일리노이주 Fox Lake에서 경찰관이 총에 맞아 사망한 사건이 발생하였으며 일주일 넘게 많은 언론에서 기사로 다뤄졌다. 이를 통해 본 연구에서 추출한 트렌드가 실제 사건 발생과 연관성 있는 결과를 도출해내는 것을 확인할 수 있다.



<그림 2> 4일동안의 카테고리별 트렌드

5. 결론

본 논문에서는 트위터 트윗을 분석하여 사전 정의된 카테고리별 실시간 트렌드를 추출하는 기법을 제안하였다. 토픽 모델링과 기계 학습을 통하여 무작위 트윗 샘플을 카테고리별로 분류하였으며, 분류된 트윗을 빈도수 기반으로 실시간 트렌드를 추출하였다. 또한 트렌드를 이루는 대표 키워드를 선정하여 어떠한 키워드로 트렌드가 이루어지는 지를 확인하였다. 이를 통해 SNS 정보를 이용하여 실시간으로 사람들이 관심있어 하는 트렌드의 흐름을 분야별로 확인하는데 도움을 줄 수 있을 것으로 예상된다.

향후 연구에서는 트윗에 포함된 해시태그를 이용한 카테고리 분류의 특징 추출이 분류의 정확성을 높이는 지에 대해 확인하고, 트렌드를 이루는 키워드와 관련있는 뉴스 등의 컨텐츠를 추천해주는 시스템에 대한 연구를 진행할 예정이다.

5. Acknowledgements

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업(IITP-2015-H8501-15-1004)과 2014년 대한민국 교육부와 한국연구재단의 기초연구사업의 지원을 받아 수행된 연구임(NRF-2013R1A1A2012627)

참고문헌

[1] Twitter, http://twitter.com/

[2] DM Blei, AY Ng, MI Jordan. “Latent dirichlet allocation”, The Journal of Machine Learning Research, Vol. 3, pp.993 - 1022, 2003.

[3] WX Zhao, J Jiang, J Weng, J He, EP Lim. “Comparing twitter and traditional media using topic models”, Advances in Information Retrieval, Vol. 6611, LNCS, pp.338-349, 2011.

[4] WX Zhao, J Jiang, J He, Y Song. “Topical keyphrase extraction from twitter”, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp.379-388, 2011.

[5] M Mathioudakis, N Koudas. “TwitterMonitor: trend detection over the twitter stream”, Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp.1155-1158, 2010.

[6] J Benhardus, J Kalita. “Streaming trend detection in twitter”, International Journal of Web Based Communities, Vol. 9, No. 1, pp.122-139, 2013.

[7] T Sakaki, M Okazaki, Y Matsuo. “Earthquake shakes Twitter users: real-time event detection by social sensors.” Proceedings of the 19th international conference on World wide web. ACM, pp.851-860, 2010.

[8] O Phelan, K McCarthy, B Smyth. "Using twitter to recommend real-time topical news.", Proceedings of the third ACM conference on Recommender systems, pp.385-388, 2009.

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

[10] Google Trends, http://google.com/trends

[11] Topcy, http://Topcy.com/[12] C Cortes, V Vapnik. “Support-vector networks.", Machine learning, Vol. 20, Issue 3, pp.273-297, 1995.

[13] libSVM, https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[14] R Kohavi. “A study of cross-validation and bootstrap for accuracy estimation and model selection.", Ijcai, Vol. 14, No. 2, 1995.