계산량 감소를 위한 LDA 단어 분포를 이용한 트위터 사용자 성향 분류 기법

이호경○, 양선, 고영중
동아대학교 컴퓨터공학과
{hogay88, seony.yang, youngjoong.ko}@gmail.com

A Twitter Users' Tendency Clustering Method using LDA Word Distributions for Complexity Reduction

Ho-Kyung Lee○, Seon Yang, Youngjoong Ko
Department of Computer Engineering, Dong-A University

요 약

트위터 데이터는 대표적인 빅 데이터 중 하나로, 트위터 데이터 활용에 대한 잠재 가치는 많은 주목을 받고 있다. 본 연구 또한 트위터 데이터의 가치에 주목한다. 본 연구에서는 트위터 사용자 성향을 자동 군집화하여 분류하는 기법을 제안하는데, 사용자 분류 정보는 마케팅 등 다양한 분야에서 유용하게 활용될 수 있다. 본 실험에서는 사용자 성향 분류를 위해 사용자와 출현 단어의 관계를 벡터로 표현 후 비지도 군집화 기법인 K-means를 적용한다. 이 때 주의할 점은 트위터 데이터의 방대한 양 때문에 벡터의 차원이 매우 높아지게 되고, 이로 인해 계산량이 엄청나게 커질 수 있다는 점이다. 이 문제를 해결하기 위해 본 연구에서는 토픽 모델을 이용한다. 즉, 토픽 모델을 통해 토픽별 단어 분포를 추출하고, 각 토픽별로 상위에 랭크된 단어들만을 이용하여 벡터 차원을 줄임으로써 계산량을 현저히 감소시킨다. 실험 결과, 전체 단어들 중 5% 이내의 단어만 사용해서도 전체 단어를 사용했을 때에 거의 근접한 결과를 얻을 수 있었다.

1. 서 론

트위터는 단시간 내에 특정한 이슈를 많은 사람들과 공유할 수 있는 최적의 서비스로 활성화되어 있으며, 정치인, 운동선수 등 많은 유명인들이 트위터를 사용자들과 소통하는 장으로 이용하고 있으며, 수많은 개인 사용자들이 유명인을 팔로우하여 커뮤니티를 구성하고 있다. 이로 인해 트위터는 사용자 수, 데이터 양, 증가 속도 등 모든 면에서 대표적인 빅데이터(big data) 중 하나로 자리매김하였으며, 이 데이터의 가치가 큰 주목을 받고 있다.

본 연구에서는 빅데이터 분석의 일환으로 트위터 사용자들의 성향을 자동 분류한다. 각 사용자의 성향을 여러 토픽으로 분류할 수 있다면 이 정보는 매우 유용하게 활용될 수 있을 것이다. 예를 들어 기업의 마케팅 경우 각 고객(혹은 잠재 고객)의 성향을 미리 알 수 있다면 훨씬 효과적인 마케팅이 가능하다. 사용자들을 대상으로 성향을 분류할 수 있는 기준 중에 하나는 어떤 유명인을 팔로우 하느냐 여부라고 볼 수 있다. 이 방법은 어떤 사용자(follower)가 특정 유명인(followee)을 팔로우한다면 그 사용자를 유명인이 속한 토픽으로 분류해도 된다고 가정한 경우에 해당되는데, 예를 들어 스포츠인을 팔로우한다면 스포츠 성향의 사용자라고 가정하는 경우이다. 하지만, 이러한 기준만으로 분류하는 것이 정말 타당한지를 확인해 볼 필요가 있다.

따라서 본 연구에서는 팔로우한 유명인이 누군지에 대한 정보를 배제한 체 순수하게 사용자들이 트위터에 올린 게시물만을 대상으로 사용자 성향 분류를 시도한다. 비지도(unsupervised) 분류 기법인 K-means를 이용하여 사용자 성향을 군집화(clustering)하는데, 자질(feature)로는 글에 출현한 단어를, 가중치 계산 기법으로는 TFIDF를 이용한다.

여기서 특히 주의할 점은, 트위터 사용자 수 자체도 워낙 많지만 사용된 단어 종류도 워낙 많기 때문에, 사용자와 단어를 벡터로 표현했을 때 차원이 너무 높아져서 계산량이 엄청나게 증가한다는 점이다. 만약 차원의 수를 획기적으로 줄이고도 전체를 사용할 때와 유사한 성능을 유지할 수 있다면, 효율적인 군집화를 수행할 수 있을 것이다.

벡터 차원을 줄이기 위해 본 연구에서는 적은 수의 단어만으로도 최대의 성능을 내는 것을 목표로 한다. 이를 위해 토픽 모델(topic model)을 이용하는데, 먼저 LDA(Latent Dirichlet Allocation)[1]에서 토픽별 단어 분포를 추출 후, 각 토픽별로 상위에 랭크된 단어들만 사용해서 전제 사용자 분류를 수행한다.

요약하자면 본 연구는 다음과 같다.

1. 트위터 사용자의 성향을 분류 및 분석한다.
2. 유명인의 토픽과 그를 팔로우한 사용자들을 특정

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아수행된 기초연구사업임 (No. NRF-2013R1A1A2009937)

토픽으로 분류한 결과가 어떤 연관관계가 있는지를 확인한다.

3. 적은 수의 자질만을 이용하는 경우에도 전체 자질을 모두 사용하는 경우와 유사한 성능을 내기 위해 LDA 단어 분포를 이용한다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구를 소개하며, 3장에서는 제안 기법을 상세히 기술한다. 4장에서는 실험 결과를 분석하며, 5장에서 결론 및 향후 계획을 기술한다.

2. 관련 연구

트위터 유명인의 정보를 이용한 다양한 트위터 연구가 활발히 진행 되어 오고 있다. 황유선[2]은 유명인과 일반 사용자의 매개 상호작용 특성 탐색 및 감정 반응을 탐색하였다. Zhang외[3], Lim외[4]는 트위터 유명인의 팔로워 정보, 관심사를 이용하여 일반 사용자들의 공통관심사를 공유하는 커뮤니티를 감지하는 연구를 하였다.

트위터 토픽모델링에 관한 선행연구는 다음과 같다. Zhao외[5]는 일반 미디어 토픽 모델링과의 차이를 기술하였고, Rosa외[6]와 Muntean외[7]는 해시태그 정보를 이용한 토픽 모델링을 제안하였다. 배정환외[8]는 토픽 모델링을 이용하여 트위터 트랜드를 추출하였으며, 류우종외[9]는 토픽모델링을 이용하여 이슈를 추출하고 이를 웹상에 시각화 하는 시스템을 설계하고 구축하였다.

트위터 게시물을 분석하는 연구로는 김석중외[10]가 있는데, 게시물의 감정을 추출하여 트위터 사용자의 정치적 성향을 분석 하였다.

3. 제안 방법

본 연구는 LDA 단어분포를 이용하여 자질 수를 획기적으로 감소시킨 후 K-means에 적용하여 사용자 성향을 군집화 하는 것을 목표로 한다. 먼저 데이터 수집을 위해 twtKr(http://twtKr.com)에서 5개 토픽별로 유명인 5명씩을 [표 1]과 같이 설정하였다.

표 1. 데이터에 포함된 토픽별 유명인 리스트

토픽	유명인(25명)				
정치인	박근혜 (대통령)	문재인 (국회의원)	박원순 (서울시장)	안철수 (국회의원)	유시민 (정당인)
연예인	하하 (가수)	노홍철 (방송인)	정준하 (방송인)	김수로 (영화배우)	최강희 (영화배우)
스포츠	이영표 (스포츠 해설가)	손연재 (체조 선수)	차두리 (축구 선수)	양준혁 (스포츠 해설가)	박문성 (스포츠 해설가)
기업인/ CEO	박용만 (두산 CEO)	이찬진 (드림위즈 CEO)	김영세 (이노디지인 CEO)	임정욱 (Lycos CEO)	허진호 (크레이지피쉬 CEO)
작가	이외수 (소설가)	공지영 (소설가)	황석영 (소설가)	강풀 (만화가)	정다정 (만화가)

총 25명 유명인들의 팔로워 62,500명(각 5,000명씩)의 트위터 게시물을 twitter api를 이용하여 2014년 8월 한 달간 수집하였다. 이 중 트위터 활동량이 지나치게 적거나 사행성 광고만을 게시하는 사용자를 제거하여 최종 13,309명 사용자의 게시물을 수집하였다.

3.1 K-means 군집화

13,309명의 트위터 게시물을 하나의 형태소 분석기를 이용하여 분석한 후 명사, 대명사, 동사, 형용사 등을 추출하였으며, 불필요한 단어들을 제거하자 최종적으로 단어 종류의 수는 5,481개였다. 한 사용자의 게시물 집합을 하나의 문서로 간주하였으며, 사용자에 대한 벡터, 즉 문서 벡터가 5,481차원이 되므로, 차원 수를 줄이기 위해 추출된 단어들을 전체문서에서 빈도수(TF) 순으로 단어 종류 수를 250(5%), 500(9%), 1000(18%), 2000(36%), 3000(55%), 4000(73%), 전체(100%)로 나누어 각각 코사인 유사도 기반의 K-means 군집화를 수행 하였다. 토픽 수 결정은 Rosa외[5]에서와 마찬가지로 수집된 게시물 그룹 수에 맞추어 K를 5로 설정하였다.

3.2 LDA 단어 분포를 이용한 벡터 차원 감소

벡터 차원 감소를 위한 다른 방법으로 LDA 단어 분포를 이용하였다. 즉, 토픽수 5개(α=0.01, β=50/토픽수)를 이용한 Mallet toolkit[11]의 오픈소스를 이용하여 사용자 게시물의 토픽별 단어 분포를 계산 한 뒤 토픽별로 스코어가 높은 단어 50개씩(Zhao외[4]) 총 250개의 단어를 선정하여 K-means에 적용하였다. [표 2]는 각 토픽별 스코어가 상위인 단어 예를 보여주고 있다.

표 2. LDA 토픽별 상위 단어 리스트

토픽1	토픽2	토픽3	토픽4	토픽5
세월호	기업	오빠	생각	선수
국민	사회	팬	사랑	운동
새누리	회사	컴백	마음	경기
교황	시장	배우	세상	응원
박근혜	삼성	엑소	행복	훈련
...	...	...	...	...

4. 실험 결과 및 분석

트위터 게시물 분류 결과 정치 기사, 연예 기사, 희망글, 자기개발 글과 같은 게시물이 많은 비중을 차지하였고, 그 외에는 종교, 사회, 경제 등의 토픽을 나타내는 글을 발견할 수 있었다. 성능 평가를 위해서 K-means 결과 중 토픽별로 상위 50개의 트위터 사용자 게시물을 보며 해당 토픽을 정한 후 실제 그 토픽에 맞는 게시물인지를 4명의 어노테이터가 확인 하였으며(Kappa=0.86), P@50을 이용하여 성능을 평가하였다. [그림 1]과 [표 3]은 단어 빈도순으로 벡터 차원을 조절하였을 때의 결과 및 LDA 단어 분포를 이용한 경우의 성능을 나타낸다. 단어 빈도순으로 차원을 줄인 경우에도 5% 사용 시 0.86의 성능을 나타내어, 단어 빈도수에 의한 벡터 차원 감소도 어느 정도 효과는 있음을 알 수 있다. 하지만 제안 방법인 LDA 단어 분포를 이용한 성능 결과는 동일하게 5% 이내인 250개의 단어만 사용하여서도 0.94의 높은 성능을 나타내었다.

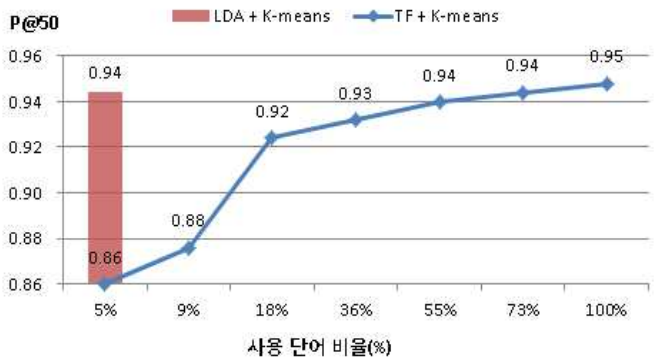

그림 1. 사용 단어 비율별 성능 변화

표 3. 최종 성능 결과

실험	사용된 단어 종류	P@50
전체	5,481 (100%)	0.95
TF 상위 250	250 (5%)	0.86
LDA 상위 250	**250 (5%)**	**0.94**

즉, LDA 단어 분포를 이용하면 단어 수, 즉 벡터 차원 수를 획기적으로 줄이고도 우수한 성능을 얻을 수 있음을 확인할 수 있다. [표 3]에서 최종 정리하였듯이 전체 단어의 5% 이내만 사용해서도 전체 단어를 다 사용한 경우인 성능 0.95에 거의 근접한 0.94라는 우수한 성능을 산출하였으며, 이는 현재 화두가 되고 있는 빅 데이터의 엄청난 계산량 문제의 해결 가능성을 보여주는 한 예라고 볼 수 있다.

그리고 어느 유명인을 팔로우 하느냐에 따라 사용자들의 성향을 분류하는 것이 타당한지를 확인해 보았는데, 유명인 자체의 성향(즉, 토픽)과 팔로워 성향 간의 관계가 전혀 없지는 않았지만, 그렇다고 오직 팔로우 여부만으로 사용자를 분류하는 것에는 무리가 있음을 확인할 수 있었다. 실제로 정치인을 팔로우한 사용자들의 트위터 게시물을 분석하면 기업, 스포츠, 연예, 희망 글 등 다양한 성향을 보임을 알 수 있었다. (이에 대한 자료는 후속 연구를 통해 상세히 보고할 예정이다.)

또한 본 연구에서는 한 달 동안의 게시물만을 수집하였는데 그 달의 이슈화 된 뉴스에 따라 트위터 게시물의 주제가 큰 영향을 받음을 알 수 있었다. 8월에는 세월호, 단식투쟁, 유민아빠, 박근혜, 새누리당에 관련된 뉴스가 이슈화가 되었다. 이로 인해 위와 관련된 토픽의 게시물들이 많은 비중을 차지하였다. 따라서 향후 더 긴 기간의 게시물을 수집하여 심도 있는 분석을 시도할 필요가 있다고 파악된다.

5. 결론 및 향후 연구

본 논문에서는 트위터 사용자에 대한 성향 분류 기법을 제안하였다. K-means를 이용한 군집화를 수행하였으며 LDA 단어 분포를 통하여 상위 랭크된 단어만을 이용함으로써 사용자 벡터의 차원을 현저히 감소시켰다. 실험 결과 전체 단어의 약 5%만 사용해서도 전체 단어 모두를 사용한 경우에 거의 근접한 결과를 얻을 수 있었다. 제안 방법을 이용하면 데이터 계산량을 현저히 줄일 수 있으므로, 앞으로 빅 데이터 분석에 유용하게 활용될 수 있을 것으로 기대된다. 향후 연구로는 짧은 기간의 데이터를 수집하기 보다는 긴 기간 동안 데이터를 수집하여 폭 넓은 실험을 수행할 예정이며, 유명인들의 성향과 팔로워들의 성향간의 세밀한 분석 자료도 보고할 계획이다.

참고문헌

[1] David M. Blei, Andrew Y. Ng and Michael I.Jordan, "Latent dirichlet allocation." The Journal of Machine Learning research, 3, pp. 993. 2003.

[2] 황유선, “유명인과의 트위터 매개 상호작용 특성 탐색”, 제13권 제8호, pp. 72-82, 2013.

[3] Yang Zhang, Yao Wu, Qing Yang, “Community Discovery in Twitter Based on User Interests”, Journal of Computational Information Systems, Vol. 8 (3), pp. 991-1000, 2012.

[4] Kwan Hui Lim, Amitava Datta, “Finding Twitter Communities with Common Interests using Following Links of Celebrities”, MSM '12 Proceedings of the 3rd international workshop on Modeling social media, pp. 25-32 , 2012.

[5] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan and Xiaoming Li, “Comparing Twitter and Traditional Media using Topic Models”, ECIR'11 Proceedings of the 33rd European conference on Advances in information retrieval, pp. 338-349, 2011.

[6] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, Robert Frederking, “Topical Clustering of Tweets”, WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining, pp. 223-232, 2012.

[7] Cristina Ioana Muntean, Gabriela Andreea Morar, and Darie Moldovan, “Exploring the Meaning behind Twitter Hashtags through Clustering”, BIS 2012 - Business Information Systems Workshops. Revised papers, pp. 231 - 242. 2012.

[8] 배정환, 한남기, 송민, “토픽모델링을 이용한 트위터 이슈 트래킹 시스템”, 지능정보연구 제20권 제2호, pp. 109-122, 2014.

[9] 류우종, 하종우, Md. Hijbul Alam, 이상근, “토픽 모델링 기법을 이용한 트위터 트렌드 추출”, 2013 한국정보과학회 제40회 정기총회 및 추계학술발표회, pp. 191-193, 2013.

[10] 김석중, 황병연, “타임라인의 감정 추출을 통한 트위터 사용자의 정치적 성향 분석”, 멀티미디어학회논문지 제17권 제1호, pp. 43-51, 2014.

[11] Mallet toolkit, http://mallet.cs.umass.edu/download.php