트위터 이고-네트워크상의 사용자 친밀도 연관 특징 분석

박창욱*, 홍지원*, 김상욱*†
*한양대학교 컴퓨터소프트웨어학과
*{ukletter, nowiz, wook}@agape.hanyang.ac.kr

On Analyzing Affinity-Related Features of Users in Twitter Ego-Networks

Chang-Uk Park*, Ji-Won Hong*, Sang-Wook Kim*†
*Dept of Computer and Software, Hanyang University

요 약

소셜 네트워크 서비스(SNS)에서는 사용자들의 친한 관계를 나타내는 여러 가지 특징을 발견할 수 있다. 본 논문에서는 트위터 이고-네트워크(ego-network) 데이터를 이용한 분석 실험을 통해 유저 간 친밀한 정도를 나타내는 여러 특징들과 관심사 유사도의 상관관계를 밝힌다.

1. 서론

트위터와 같은 소셜 네트워크 서비스들은 사용자들을 서로 여러 가지 방식으로 연결하여 사용자들이 서로 의사소통을 하거나 친분을 나타낼 수 있는 환경을 제공한다. 사용자들은 자신의 현재 상태나 특정 주제에 대한 생각을 표현하고 다른 사람의 글에 관심을 나타낼 수 있기 때문에 다양한 분석 기법을 이용한다면 사용자 간 관심사 유사도를 유추해낼 수 있다. 이러한 소셜 네트워크에서 사용자들 간 친밀한 정도를 나타내는 여러 특징들을 이용하여 관심사의 유사도를 판단할 수 있다면 소셜 네트워크를 이용하는 여러 응용 분야에 사용자들의 친밀도를 적극 활용할 수 있을 것이다.

소셜 네트워크 전체를 모두 분석하는 것은 심각한 계산 비용이 발생하기 때문에 자기 자신과 그 친구들까지만 포함하는 이고-네트워크(ego-network)를 대상으로 한 연구가 최근 활발히 수행되고 있다[3][4][5][6]. 사용자가 관심 있을만한 트윗을 추천함에 있어 이고-네트워크만 이용한 방법이 전체 네트워크를 이용하는 것보다 계산 비용에 있어서 큰 이득이 있으면서도 정확도는 크게 차이나지 않는다는 연구 결과도 발표된 바 있다[2].

본 연구에서는 이고-네트워크를 이용하여 사용자 간의 친밀한 정도를 나타내는 몇 가지 특징들과 사용자 간 관심사 유사도의 상관관계를 밝히고, 여러 응용 분야에 사용자 친밀도 관련 특징들이 이용될 수 있음을 확인한다.

2. 데이터 수집

본 연구에서는 실험을 위해 트위터 데이터 수집기[8]를 제작하고 CNN의 트위터 계정을 팔로우하는 사용자들을 대상으로 이고-네트워크 데이터를 직접 수집하였다. 수집 대상은 미국에 거주하고 영어를 구사하는 사용자 계정으로 한정하였고, 비공개 계정과 공인 및 공식 계정, 팔로잉 또는 팔로워 수가 5,000회 이상이거나 작성한 트윗이 100개 이하인 계정들은 제외하였다. 본 논문에서는 상호 팔로우하는 관계를 친구 관계로 정의하고, 536명의 사용자들을 중심으로 해당 사용자와 친구들을 포함하는 536개의 이고-네트워크 데이터를 수집하였고, 이 중 네트워크의 크기가 100 이하인 경우를 제외하여 최종적으로 89개의 이고-네트워크를 선별하였다. 89개의 이고-네트워크의 총 사용자 수는 13,629명이고, 사용자들의 타임라인에서 사용된 총 어휘의 종류는 1,379,986개이다. 어휘는 NLTK 형태소 분석기[7]를 이용하여 명사, 고유명사, 형용사, 동사에 대해서만 추출하였다. 분석 실험은 각 이고-네트워크 별로 독립적으로 수행하였다.

3. 분석 실험

본 분석 실험에서 사용자 간의 연락 빈도(mention count), 상대방이 작성한 트윗을 좋아한 횟수(like count), 두 사용자가 공통으로 친구 관계를 형성하고 있는 사용자의 수(mutual friends count)를 사용자 간 친밀도를 나타내는 특징 요소로 사용하였고, 사용자의 관심사를 유추할 수 있는 측도로는 토픽 유사도(topic similarity)와 트윗 선호 유사도(like similarity)를 이용하였다.

토픽 모델은 사용자가 글을 작성할 때 사용한 어휘와 다른 사용자들이 사용한 어휘를 비교 분석하여 해당 사용자가 어느 토픽(예를 들어, 정치, 경제, 문화 등)에 더 관심이 많고 적은지를 수치적으로 구할 수 있는 통계 모델이다. 토픽 유사도를 측정하는 것은 사용자가 작성한 글의 내용을 고려하는 내용 기반의 관심사 비교 방법이다.

본 논문에서는 트윗을 리트윗하고, 공유하고, 관심글에 등

† 교신저자

본 연구는 (1) 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원(NRF-2014R1A2A1A10054151)과 (2) 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원(IITP-2015-H8501-15-1013)을 받아 수행됨.

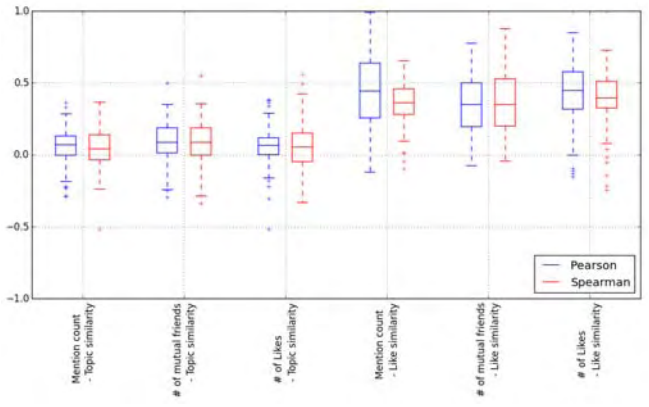
록하는 행위 모두를 해당 트윗을 “좋아한다”고 간주하는데, 트윗 선호 유사도는 두 유저가 함께 좋아한 트윗이 얼마나 많은지를 나타낸다. 동일한 트윗을 많이 좋아하는 두 사용자는 관심사가 유사하다고 생각할 수 있다.

이와 같이 토픽 유사도와 트윗 선호 유사도는 두 사용자의 관심사가 얼마나 유사한지를 나타내는 지표로 이용될 수 있다. 본 논문에서 트윗 선호 유사도는 Jaccard 유사도 계산 방법을 이용하여 계산하였다. 토픽 유사도는 수집한 89개의 이고-네트워크의 모든 사용자에 대해서 타임라인 글에 쓰인 어휘들을 이용하여 Latent Dirichlet Allocation(LDA) 알고리즘[1]을 수행하고 각 사용자 쌍별로 cosine 유사도 계산 방식으로 계산하였다. 토픽의 수는 30개로 정하고, collapsed Gibbs sampling[9]으로 모델을 학습하였다.

우선 각 이고사용자(ego user)마다 친구들과의 연락 빈도를 구한 후 연락 빈도에 따른 토픽 유사도와의 상관관계와 트윗 선호 유사도와의 상관관계를 구한다. 상관관계는 Pearson 방식과 Spearman 방식 두 가지에 대해서 측정하였다. 사용자마다 소셜 네트워크 서비스를 이용하는 성향이 각기 다르기 때문에 각 빈도 값은 min-max scaling을 통해 0과 1사이의 범위로 사상하였다. 연락 빈도의 경우처럼 상대방 글을 좋아하는 횟수와 공통 친구들의 수에 대해서도 동일한 방법으로 측정하였다.

<표 1> 상관관계의 평균값

		Mention count	Mutual friends count	Like count
Topic similarity	P	0.0581	0.0929	0.0568
	S	0.0459	0.0875	0.0625
Like similarity	P	0.4374	0.3417	0.4198
	S	0.3512	0.3656	0.3901



(그림 1) 특징 요소와 관심사 유사도의 상관관계

<표 1>과 (그림 1)에서 알 수 있듯이 모든 특징 요소가 서로에 대해 평균적으로 양의 상관관계를 가짐을 확인할 수 있었다. 서로 연락을 많이 하고, 상대방이 작성한 글을 많이 좋아하는 사용자일수록 자신이 좋아하는 글을 상대방도 좋아하는 경우가 많음을 알 수 있다. 그러나 토픽 유사도에 대해서는 뚜렷한 상관관계가 나타나지 않았다. 이는 트위터에서 사용자들이 사용하는 어휘 중 오타, 의성어, 의태어 등 비표준어의 비율이 정형화된 문서보다 훨씬 많고, 문법을 무시한 트윗이 많으며, 장시간에 걸쳐 다양한 주제에 대해서 트윗을 작성하기 때문에 토픽 모델링을 바로 적용하는 것이 트위터 환경에서는 적절치 않았던 것으로 해석될 수 있다.

4. 결론

본 논문에서는 사용자 간 친밀도를 나타내는 특징 요소인 연락 빈도, 상대방 글을 좋아하는 횟수, 공통 친구들의 수와 관심사의 유사도를 나타내는 토픽 유사도, 트윗 선호 유사도를 서로 비교하여 각 측도 간의 관계를 살펴보았다. 서로 연락하는 횟수가 많거나 상대방의 글을 많이 좋아하면 비슷한 관심사를 가질 가능성이 높다는 사실을 확인하였다. 추천 시스템과 같이 사용자 관심사를 이용하는 응용 분야에 사용자 간 친밀도 관련 특징을 추가로 이용한다면 더 높은 추천 정확도를 기대할 수 있을 것이다. 사용자들은 소셜 네트워크 서비스를 이용하면서 작성자와의 친한 관계로 인해 글에 대한 만족도와는 무관하게 좋아하는 경우도 많기 때문이다.

향후 연구에서는 사용자 간의 친밀도를 더 잘 파악할 수 있는 추가적인 특징 요소들을 찾고, 고려할 수 있는 모든 특징 요소를 실제 여러 응용 분야에 적용 후 성능을 검증해 볼 수 있을 것이다.

참고문헌

[1] D. M. Blei et al. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993-1022, 2003.

[2] A. Sharma et al. Friends, Strangers, and the Value of Ego Networks for Recommendation. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, pp. 721-724, 2013.

[3] K. Chen et al. Collaborative Personalized Tweet Recommendation. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 661-670, 2012.

[4] F. Liua et al. Use of social network information to enhance collaborative filtering performance. Expert Systems with Applications, 37(7):4772-4778, 2010.

[5] Y. Yang et al. Automatic Social Circle Detection Using Multi-View Clustering. Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pp. 1019-1028, 2014.

[6] J. McAuley et al. Discovering Social Circles in Ego Networks. ACM Transactions on Knowledge Discovery from Data, vol. 8, no. 1, pp. 4:1-4:28, 2014.

[7] http://www.nltk.org

[8] https://github.com/ChangUk/TwitterCrawler

[9] https://github.com/ChangUk/pyGibbsLDA