소셜 데이터에서 재난 사건 추출을 위한 사용자 행동 및 시간 분석을 반영한 토픽 모델

촐몽 바야르, 이경순
전북대학교

요 약

본고에서는 소셜 빅데이터에서 공공안전에 위협되고 사회적으로 이슈가 되는 재난사건을 추출하기 위한 방법으로 소셜 네트워크상에서 사용자 행동 분석과 시간분석을 반영한 토픽 모델링 기법을 알아본다. 소셜 사용자의 글 수, 리트윗 반응, 활동주기, 팔로워 수, 팔로잉 수 등 사용자의 행동 분석을 통하여 활동적이고 신뢰성 있는 사용자를 분류함으로써 트윗에서 스팸성과 광고성을 제외하고 이슈에 대해 신뢰성 높은 사용자가 쓴 트윗을 중요하게 반영한다. 또한, 트위터 데이터에서 새로운 이슈가 발생한 것을 탐지하기 위해 시간별 핵심어휘 빈도의 분포 변화를 측정하고, 이슈 트윗에 대해 감성 표현 분석을 통해 핵심 이슈에 대해 사건 어휘를 추출한다. 소셜 빅데이터의 특성상 같은 날짜에 여러 이슈에 대한 트윗이 많이 생성될 수 있기 때문에, 트윗들을 토픽별로 그룹핑하는 것이 필요하므로, 최근 많이 사용되고 있는 LDA 토픽모델링 기법에 시간 특성과 사용자 특성을 분석한 시간상에서의 중요한 사건 어휘를 반영하고, 해당 이슈에 대한 신뢰성 있는 사용자가 쓴 트윗을 중요시 반영하도록 토픽모델링 기법을 개선한 소셜 사건 탐지 방법에 대해 알아본다.

Ⅰ. 서 론

사람들은 자신의 의견, 생각, 경험을 서로 공유하기 위해 페이스북, 유튜브, 인스타그램, 트위터, 링크드인 등과 같은 소셜 네트워크 서비스(Social Network Service)를 이용한다. 트위터(twitter) 는 블로그의 인터페이스에 미니홈피의 인적 네트워크 형성, 메신저의 신속성을 한데 모아놓은 구조로, 트위터의 사용자가 급증하면서 기존의 언론 미디어보다 더 빠르게 정보를 전달하는 파급 효과를 가지고 있다[1]. 실례로 뉴욕 허드슨강 여객기 불시착 사건, 강남 파이낸스센터 화재사건 등은 트위터가 언론보다 더 빠르고 정확하게 정보를 전달한 사례가 있다. 트위터에서 재난 사건 및 이슈 추출에 관한 연구가 활발하다[2][3].

트위터를 이용하는 사용자 수가 지속적으로 늘어나 2012년 기준 1억명 이상의 활동적인 사용자가 매일 평균적으로 3억4천개의 트윗(tweet) 글을 올리며, 하루에 16억 번의 질의어 검색이 이루어지고 있다[4]. 2016년 미국 대선일 하룻동안 트위터에는 4천만건 이상의 선거관련 트윗이 게재되는 등 가장 주요한 정보 전달 매체가 되고 있다[4].

이와 같은 소셜 빅데이터에서 사용자는 관심 있는 주제에 대하여 신뢰할 만한 트윗에서 중요한 정보를 얻는데 시간이 많이 걸리는 문제점이 있다. 특히 사용자가 입력으로 준 질의에 대한 트윗 내용을 살펴보면 스팸과 광고 같은 트윗이 많이 있다. 수많은 트위터 데이터에서 사용자가 필요하고 신뢰성 높은 트윗 또는 그 질의(이슈 또는 사건)에 대해서 잘 아는 전문가와 같은 사용자를 알아보기 위한 트위터 사용자 분류 시스템이 요구되고 있다.

본고에서는 소셜 빅데이터에서 사회적으로 이슈가 되는 재난사건을 추출하기 위한 방법으로 사용자 행동 분석과 시간 분석을 반영한 LDA(Latent dirichlet allocation)[5] 기반 사건 토픽모델에 대해 알아본다.

소셜 사용자 네트워크에서 사용자 행동 분석은 사용자가 게재한 트윗에 대한 신뢰도를 판단하는 근거로 이용할 수 있다. 재난에 대한 글을 올렸을 때, 사용자의 신뢰도 정도에 기반해서 그 글을 중요하게 다룰지 스팸으로 다룰지 결정할 수 있다. 사용자들의 행동은 사용자가 쓴 총 트윗 수, 그 글들이 리트윗된 횟수, 주기적으로 트위터에서 활동하는지 등에 기반하여 그 사용자가 쓴 글에 대한 신뢰성을 측정할 수 있다. 또한 사용자 관계 네트워크에서 팔로워(follower)와 팔로잉(following) 비율을 이용하여 활동성이 높은 사용자와 활동적이지 않은 사용자를 분류한다.

트위터 데이터의 시간분석을 통해서 시간에 따른 어휘 분포에서의 변화를 분석함으로써 새로운 이슈가 발생했는지를 탐지할 수 있다. 이는 날짜별로 어떤 이슈에 대해 트위터에 게재된 트윗 개수의 변화, 사용자들이 이슈에 대해 긍정 또는 부정 감정 표현 인식, 그리고 리트윗 (Retweet/RT) 개수 등의 변화를 분

석한다.

〈그림 1〉의 위 그래프에서는 경주지진 사건이 발생한 날인 2016년 9월 13일을 기준으로 하루전날과 그 다음날들에서의 트위터 글에서 '경주 지진' 어휘가 몇번 나타났는지의 분포 변화를 그래프화로 표현한 것이다. 아래 그래프는 갤럭시 노트7 폭발 사건이 발생한 날인 2016년 8월 24일을 기준으로 '노트7 폭발', '갤노트7 폭발', '갤럭시 폭발' 등의 어휘들이 트윗 개수가 갑자기 높게 발생한다. 트윗 데이터에서 특정한 사건이 발생했을 때, 그 어휘에 대한 트윗 개수가 급격히 증가하는 것을 볼 수 있다. 시간상에서의 어휘 분포 변화를 탐지함으로써 새로운 이슈를 탐지할 수 있다.

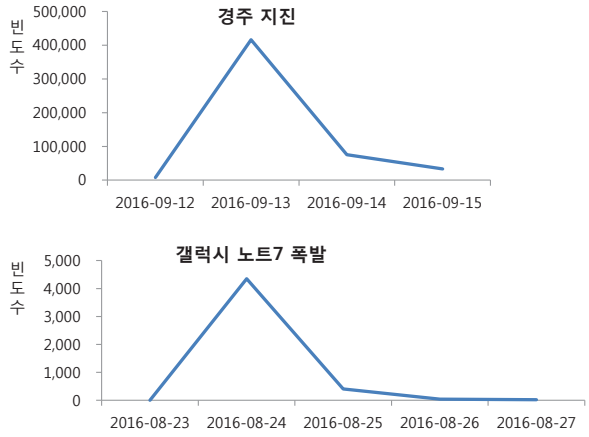


그림 1. 사건 발생일에 트윗 빈도 변화 그래프

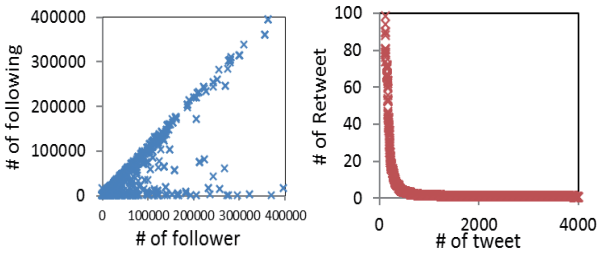
소셜 데이터의 특성상 하루에도 여러가지 사건 사고가 발생할 수 있기 때문에 그에 따른 트윗 내용들도 다양하다. 따라서 같은 사건에 대한 트윗 글들을 그룹핑하여 분석할 필요가 있다. 여러 사건을 추출하기 위해 자연언어처리 분야와 기계학습분야에서 최근 많이 이용하고 있는 LDA 토픽모델링 기법을 이용하여 트윗 문서들과 어휘들을 주제별로 분류한다. 본고에서는 LDA 모델에 신뢰성 높은 소셜 사용자와 시간상에서의 어휘변화 분석 정보를 반영함으로써 더 효율적인 사건 추출 방법에 대해서 알아보겠다.

Ⅱ. 소셜 데이터에서 사용자 행동 분석

트위터 데이터는 사용자의 목적에 따라 스팸성 트윗과 광고성 트윗도 많이 포함되어 있다. 수많은 트위터 데이터에서 신뢰성 높은 트윗 또는 어떤 주제에 대해서 잘 아는 전문가와 같은 사용자를 추출하기 위한 트위터 사용자 분류 시스템이 필요하다[6].

사용자의 신뢰도를 측정하기 위해 본고에서는 트위터 사용자를 사회적으로 잘 알려진 사용자, 신뢰성과 활동성이 높은 사용자, 일반 사용자, 활동력이 낮은 사용자로 분류한다.

사용자 행동분석을 위해 트위터 사용자의 팔로워와 팔로잉 수, 트윗 수, 리트윗 횟수, 주기별 활동량 정보를 이용하여 분석한다.



사용자간 팔로잉과 팔로워 수 트윗 수와 리트윗 수

그림 2. 사용자 팔로잉 관계 및 트윗과 리트윗 수

사용자 행동 분석을 효율적으로 하기 위해, 사용자간의 팔로잉과 팔로워 관계 정보를 그래프로 표현하여 그래프 분석 기법인 HITS(Hyperlink-Induced Topic Search) 알고리즘을 이용하여 소셜 네트워크 사용자 분석에 적합하도록 개선하여 활동력과 신뢰성이 높은 사용자를 분류하는 방법을 알아본다.

1. 트위터 사용자의 행동분석 및 분류

신뢰성 있는 트윗 또는 중요한 정보를 포함하고 있는 트윗을 쓰는 사용자를 수식 (1)를 통해 사용자의 신뢰성(RTRatio)을 측정한다.

$$RTRatio(u) = \frac{1}{N}\sum_{i=1}^{N} RT(u, Di) \quad (1)$$

여기서 N은 사용자 u가 트위터 데이터를 수집한 기간 내에 쓴 모든 트윗 개수 이다. RT는 사용자 u가 쓴 트윗 Di 가 얼마나 많이 리트윗(retweet) 되었는지 나타낸다. RTRatio(u) 값은 사용자가 트윗 하나를 작성하였을 때 그 트윗이 리트윗 될 평균값을 의미한다. 리트윗은 정보 전달 횟수로 트윗 내용에 대한 유용성을 표현한다고 볼 수 있다. 따라서, RTRatio 값이 높을수록 그 사용자의 글에 대한 반응이 높고, 유용한 정보를 전달하고 있다고 볼 수 있다.

트위터 사용자의 팔로잉과 팔로워 수를 이용하여 사용자가 얼마나 트위터에서 여러 사람과 관계를 갖고 활동을 하는지를 나

타내는 사용자 활동성(FFRatio)을 측정한다.

$$FFRatio(u) = \frac{\#of\ Follower(u)}{\#of\ Following(u)} \quad (2)$$

여기서 Follower(u)는 사용자 u를 팔로잉하는 사용자 수이고, Following(u)는 사용자 u가 팔로잉하는 사용자 수를 나타낸다. 즉, FFRatio 값은 사용자 u가 트위터에서 얼마나 많은 사람들에게 알려져 있는지를 의미한다.

RTRatio 값과 FFRatio 값을 조합하여 Follower(u) 수가 Following(u) 보다 아주 큰 경우 이 사용자를 팔로워하는 수가 많으므로 트위터에서 잘 알려진 유명한 사람으로 볼 수 있다. 사용자의 팔로잉과 팔로워 비율이 비슷한 경우 일반 사용자로 분류할 수 있다. 어떤 사용자를 팔로워 하는 사용자 거의 없는데 팔로잉 수가 높으면 광고 목적을 가진 스팸 사용자로 분류할 수 있다.

2. 활동성이 높은 사용자 분류

소셜 사건이 일어나면 그 사건과 관련된 트윗이 폭발적으로 증가한다. 그 사건과 관련된 트윗을 쓴 사용자들은 그 주제에 대해서 관심 있다고 볼 수 있다. 사용자의 팔로워 수와 리트윗 수를 이용하여 그 사용자가 사건과 직접적으로 관련 있다고 판단하기가 어렵다. 앞서 이용한 사용자의 팔로워와 팔로잉 수 정보는 고정된 값이기 때문에 이슈에 대해서 동적인 관계를 나타내지 못한다.

웹 페이지들 간의 동적인 관계를 분석하는데 많이 이용되고 있는 HITS 알고리즘[7]을 본 목적에 맞게 개선하여, 트위터 사용자간의 네트워크의 링크를 분석하는데 적용한다[8].

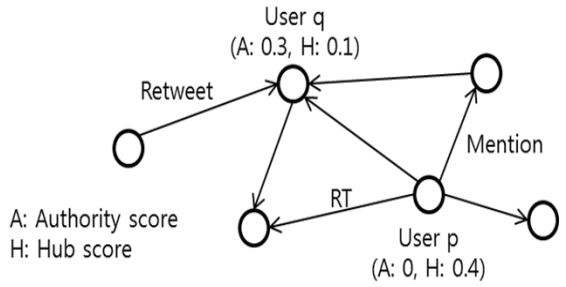


그림 3. 소셜 사용자간의 네트워크 그래프

소셜 사용자간의 네트워크에서 사용자를 하나의 노드(p)로 표현한다. 각 노드는 권위(Authority) 점수 및 허브(Hub) 점수를 가진다. 높은 권위값을 갖는 사용자로부터 링크되어 있을수록 좋은 허브값을 갖게 되고, 높은 허브값을 갖는 사용자들로부터 링크되어 있을수록 높은 권위값을 갖게 된다. 방향 간선(edge, p->q)는 사용자 p의 트윗에서 사용자 q를 언급했을 때 생성한다. 즉, 사용자 p가 작성한 트윗을 사용자 q가 리트윗(retweet 또는 RT) 했을 때 사용자 q에서 p로 나가는 간선을 생성하고 또는 사용자 p가 q의 이름을 언급(Mention)하여 트윗을 작성했을 때 간선을 생성하여 사용자 p에서 사용자 q로 간선이 몇 번 나가는지를 계산하여 간선 가중치 w_{pq}를 부여한다.

트위터 사용자간의 네트워크를 분석하여 사용자를 분류하기 위해 기존의 HITS 알고리즘에 추가로 간선 가중치 w_{pq}를 반영하여 권위 값과 허브 값을 계산하는 수식은 다음과 같다.

$$HubScore^{(T+1)}(p) = \sum_{p \to q} w_{pq} \times AuthScore^{T}(q) \quad (5)$$

$$AuthScore^{(T+1)}(p) = \sum_{q \to p} w_{pq} \times HubScore^{T}(q) \quad (6)$$

각 노드 p의 초기값 $HubScore^{(0)}(p)$과 $AuthScore^{(0)}(p)$은 사용자 p의 FFRatio와 RTRatio 값으로 한다. 반복 횟수 T는 HITS 알고리즘을 몇 번 반복할 것인지를 나타내며 T과 T+1번째의 사용자 AuthScore값의 차이가 0.0001 보다 작으면 반복을 멈춘다. 간선 가중치인 w_{pq}에 트위터의 특성을 반영한 수식은 다음과 같다.

$$w_{pq} = \sum_{p \to q} FreqRT(p,q) + \sum_{p \to q} Mention(p,q) \quad (7)$$

여기서 수식 5과 6에서 나오는 w_{pq}는 기존 HITS 알고리즘을 개선하여 적용한 것이다. w_{pq}의 요소에는 사용자 p가 사용자 q의 트윗을 리트윗한 횟수를 반영한 FreqRT(p,q) 값과 사용자 p가 사용자 q의 이름을 언급하여 작성한 트윗 수를 나타내는 Mention(p,q) 값이 반영된다.

신뢰성이 높고 활동성이 높은 사용자는 AuthScore 값이 높은 상위 100명의 사용자를 선택하여 그들이 쓴 글을 중요하게 반영한다.

3. 이슈에 대해 신뢰성이 높은 사용자 분류

앞서 HITS 알고리즘을 이용하여 활동성이 높은 사용자를 추출할 때 사용자간의 네트워크 구조를 분석하였지만 시간 속성이 중요하게 반영되지는 않았다.

트위터에서 아주 활발한 사용자는 아니지만 어떤 이슈에 대해 사건이 일어날 때마다 트윗을 쓰는 사용자가 존재한다. 이러한 사용자는 지속적으로 이슈와 관련된 정보를 주기 때문에 신뢰성이 높은 사용자로 볼 수 있다.

따라서 이슈에 대해 지속적으로 트윗을 쓰며 중요한 정보를 주는 사용자를 추출하기 위한 주별 평균 활동력 값을 측정한다.

$$Activity(u) = \frac{1}{W}\sum_{i=1}^{W} twFreq(u,d_i) \times RTFreq(u,d_i) \quad (8)$$

여기서 W는 총 주(week)의 수이고, twFreq(u,d_i)은 사용자 u가 i번째 주의 이슈에 대해 작성한 모든 트윗의 수 d_i를 나타난다. RTFreq(u, d_i)은 사용자 u가 i번째 주에 이슈에 대해 작성한 모든 트윗의 수 d_i이 리트윗된 횟수를 나타낸다. 각 사용자에 대한 활동성이 높을수록 신뢰성이 높은 사용자가 된다.

'지진' 이슈에 대해 직접적인 관계를 가진 사용자의 트윗 분포 및 지진이 발생한 날짜를 표현한 그래프는 다음과 같다.

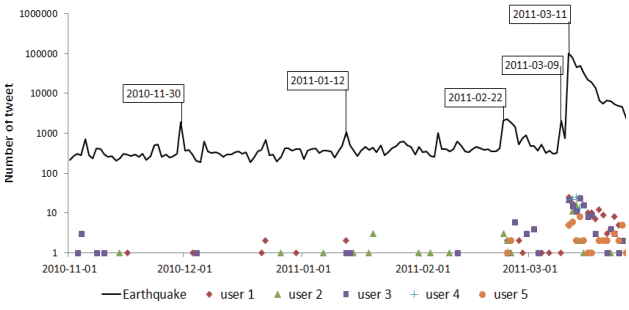


그림 4. 지진 사건에 대한 날짜별 트윗 수 및 신뢰성 있는 사용자의 트윗 분포

위 사용자 중에서 user 1은 지진 발생한 지역에 한국 사람이 피해를 당했을 때 직접적으로 도움을 주기위해 트윗을 작성하였다. user 2와 user 3은 공식적으로 지진이 발생했다는 것을 알려주는 트윗을 많이 썼다. user 5는 해외에서 발생한 지진에 대해 알려주고 지진 발생한 지역에 있는 한국 대사관에 대한 안내를 해주고 있었다. 사용자 행동 분석을 통해서 사건에 대해 중요한 사용자를 탐지함으로써 신뢰성 있는 정보를 추출할 수 있음을 알 수 있다.

실험에서는 사회적으로 잘 알려진 활동성이 높은 사용자를 추출하는 것에 대해 AuthScore가 78%, 신뢰성과 활동성이 높은 사용자를 추출하는 것에 대해 Activity 값에 기반한 것이 91%의 성능을 보였다. 그 이슈 사건에 대해 트윗을 작성한 사용자를 대상으로 실험한 결과 높은 성능을 보인 것을 알 수 있었다.

III. 시간 분석을 통한 사건 어휘 추출

본고에서는 트위터 자료의 시간별 분석을 통해서 새로운 이슈가 발생했는지를 인식하고 보다 효율적으로 사건을 나타내주는 어휘를 선택하기 위해 날짜별로 사건 어휘 후보들의 빈도수, 감성 자질 표현들을 추출한다. 그리고 시간상에서의 어휘들의 분포 변화를 반영하기 위해 통계적인 기법인 카이제곱(Chi-Square)을 사용하여 이슈에 대한 핵심 사건을 추출한다.

1. 시간별 어휘 빈도수를 이용한 기본 자질 추출

사건 어휘를 표현하는 기본 자질 추출을 위해 수집된 트위터 데이터를 형태소 분석을 한 뒤, 불용어와 불필요한 URL정보를 제거하였다. 정제된 자질들에 대해 하나의 트윗 내에서 거리가 3이내에 있는 두 어휘들의 조합인 바이그램(Bigram)으로 핵심 사건 어휘 후보로 추출한다. 재난재해 사건을 추출하기 위해서는 재난에 관한 키워드 어휘들을 중심으로 같이 발생한 어휘들을 추출한다. 예를 들어, '지진', '폭발', '홍수', '테러' 등의 키워드들과 함께 나오는 어휘들의 조합을 사건 후보 어휘로 한다.

각 시간별로 바이그램으로 추출된 자질들을 어휘 빈도수를 계산하여 큰 값부터 순위화하였다. 빈도수를 이용하여 기본 자질들을 추출한다. 어휘 빈도수에 의한 기본 자질의 값을 Freq(w, t_0)라 하겠다.

$$Freq(w,t_0) = \sum_{D \in t_0} tf(w,D) \quad (9)$$

여기서 w는 사건 어휘를 나타내고, t_0는 각 날짜를 나타내고, D는 시간 t_0에 속하는 트윗 문서를 나타낸다. 그날 t_0에 생성된 트윗들에서 어휘 w가 발생한 총 개수를 계산한 것이다.

2. 감성 자질을 반영한 자질 추출

수집된 트위터 데이터를 관찰해보면 어느 한 이슈에 대해 특정 사건이 발생하게 되면 그 사건에 해당하는 사건 어휘와 함께 감성 자질이 함께 출현하는 경우를 자주 볼 수가 있다.

감성 어휘는 한국어 감성 표현으로 강한 긍정과 강한 부정을 나타내는 감성 자질 1192개를 구축하였다. 긍정표현으로는 '찬사', '칭찬', '격려', '동의' 등의 490개의 어휘를 포함하고, 부정표현으로는 '비난', '반대', '분노' 등 702개의 표현을 포함한다.

또한, 감성 기호인 이모티콘(emoticon, 그림말)을 이용하여 사용자들이 트윗 글에 자신의 감정을 많이 표시하고 있어, 트윗에서 사건어휘 자질 추출을 위해 긍정 감성기호 156개, 부정 감

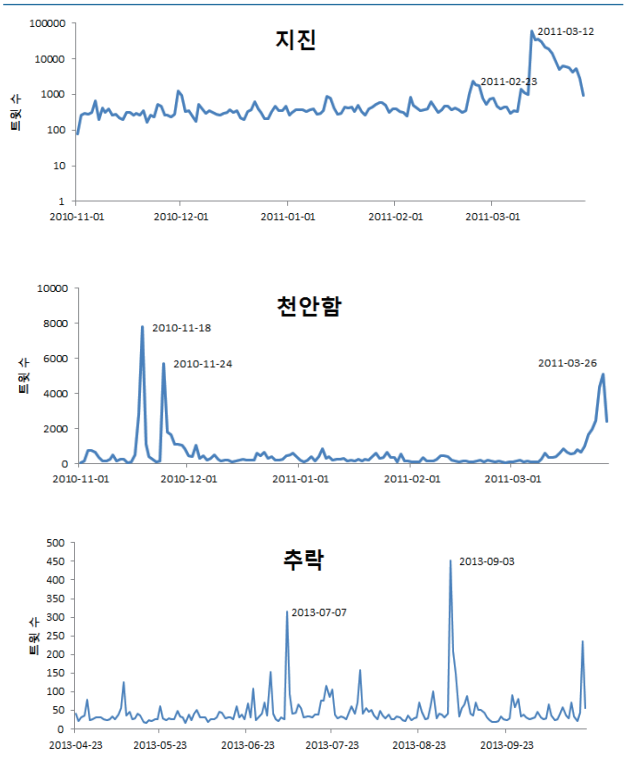


그림 5. 각 이슈에 대한 날짜별 트윗 수

성기호 96개를 사용하였다.

기본 사건 자질로 추출된 어휘들 중에서 상위 50개에 대해 해당 이슈의 전체 트위터 문서집합에서 감성 어휘 및 기호로 표현된 감성 자질과 함께 출현한 어휘 자질의 빈도수를 OpFreq(w, s)라 하겠다.

$$OpFreq(w,s,t_0) = \sum_{s \in D} \sum_{w \in D} tf(w,s,D) \quad (10)$$

여기서 tf(w, s, D)는 어휘 자질 w가 감성 자질 s과 함께 트윗 D에 시간 t_0에 같이 나타난 빈도수를 나타낸다.

시간 t_0에서 어휘 자질 w의 빈도수와 감성 자질 정보를 반영한 수식 OpScore은 다음과 같다.

$$OpScore(w,t_0) = Freq(w,t_0) + \alpha\, OpFreq(w,t_0) \quad (11)$$

여기서 감정표현과 같이 나타난 어휘의 가중치를 α로 조정하여 높게 반영하였다.

3. 감성과 시간별 자질을 반영한 사건 어휘 추출

시간별로 트윗 개수를 분석한 결과 어떠한 이슈에 대해 특정 사건이 발생했을 때 그 날짜의 트윗 개수는 전날에 비해 급격하게 증가하는 현상을 보였다. 만약 그 사건이 그 이전에는 발생하지 않은 새로운 사건 또는 발생한 적이 거의 없는 사건이라면 사건 어휘에 대한 Freq(w, t_0) 값이 그 이전 날들의 데이터보다 폭발적으로 증가함을 알 수 있었다. 이러한 특성을 반영하기 위해 시간 t_0에서 사건 어휘자질 w의 중요도를 계산하기 위해 카이제곱을 이용하여 계산하였다. 〈표 1〉은 카이제곱 값을 계산하기 위한 분할표이다.

표 1. 카이제곱 값을 계산하기 위한 분할표

	자질 w를 포함한 트윗 (w ∈ D)	자질 w를 포함하지 않은 트윗 (w ∉ D)
t_0 트윗 (D ∈ t_0)	a	b
t_0 이전 트윗 (D ∉ t_0, D∈t, t<t_0)	c	d

시간 t_0에서의 카이제곱은 다음과 같이 계산한다.

$$ChiSquare(w,t_0) = \frac{(a+b+c+d)(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \quad (12)$$

사건 어휘 자질의 순위화에 시간상에서의 특성을 반영한 ChiSquare(w, t_0)값, 리트윗 및 감정자질과 함께 출현한 정보를 반영한 OpScore(w,t_0)값을 이용한 최종 수식은 다음과 같다.

$$ChiOpSquare(w,t_0) = \lambda\, ChiSquare(w,t_0) + (1-\lambda)\, OpScore(w,t_0) \quad (13)$$

여기서 파라미터 λ값은 훈련 이슈('천안함')에 대한 학습을 통해서 0.3으로 설정했다.

본고에서의 새로운 이슈에 대한 추출 방법은 수식 (13)에서와 같이 날짜별로 추출한 사건 어휘에 대해서 시간상에서의 그 어휘를 포함한 트윗 개수의 변화 정도가 크고, 감정 표현과 함께 표현된 어휘들을 높게 반영한다.

Ⅳ. 시간 및 사용자 행동 분석 기반 토픽모델

소셜 데이터에서 사건 추출을 위한 시간 및 사용자 행동 분석 기반 LDA 모델을 살펴본다. 기계 학습 및 자연언어처리 분야에서 활발히 적용하고 있는 토픽 모델링은 문서 집합에서 발생하는 추상적인 "토픽"을 발견하기 위한 통계모델의 일종이

다. LDA 모델은 대표적인 토픽 모델로 2003년 기계학습 국제학회인 ICML에 처음으로 발표된 후 LDA 기반 다양한 연구가 이뤄지고 있다. LDA 모델은 지도학습과 비지도학습이 가능한 기계학습 모델로, 데이터의 집합에 대한 확률적 생성 모델(Generative Probabilistic Model)이다[5]. 생성 모델은 어떤 확률분포와 그 파라미터가 있다고 할 때, 랜덤 프로세스에 따라 데이터를 생성하는 모델이다. 문서의 토픽 분포와 각 토픽별로 특정 단어의 생성확률을 알고 있으면, 특정 문서가 만들어질 확률을 계산할 수 있다[9].

기존의 토픽 모델에 사용자 정보를 처음으로 추가한 논문인 [10]가 모든 문서가 그 문서를 작성한 저자를 통해 생성된다고 하였다. 같은 주제에 대해 서로 다른 사용자가 작성한 문서가 각 사용자에 특성에 따라 토픽 분포가 달라진다는 점을 LDA에 모델에 반영하여 실험을 하였다. 또한 시간상에서 갑자기 빈번해지는 어휘를 사건추출에 반영하고 정적인 사용자 정보를 반영한 LDA 모델[11] 연구도 있다.

본고에서는 이슈가 발생하였을 때 그 이슈에 따른 세부 토픽들을 추출하기 위해 앞서 설명한 시간 분석 및 사용자 행동 분석을 통한 LDA 모델을 이용하여 사건을 추출한다.

소셜 환경에서 일어나는 사건들의 특징과 문서를 작성한 사용자의 행동 분석 및 시간에 따른 어휘 빈도수를 LDA 토픽 모델에 반영한 시스템 구조는 다음과 같다.

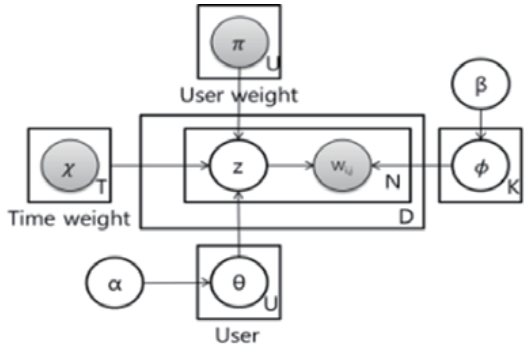


그림 6. 시간 및 사용자 분석 기반 LDA 모델

〈그림 6〉의 모델에 표시된 변수들은 다음과 같다.

T: 시간 $t \in \{1,\cdots,T\}$

U: 시간 t에 글을 쓴 사용자들. $u \in \{1,\cdots,U\}$

χ: 시간 t에서 단어 w의 ChiSquare값

π: User Authority. 각 사용자에 대한 신뢰도

θ: 토픽 k가 사용자 u에 나타날 확률

w: 문서 d에 나온 총 단어 $w \in \{1,\cdots,N\}$

φ: 단어 w가 토픽 k에 속할 확률

α: 각 문서가 토픽 K에 속할 초기 사전 확률

β: 각 단어가 토픽 K에 속할 초기 사전 확률

z: 문서 d에서 단어 w가진 토픽 비율

K: 토픽 수 $k \in \{1,\cdots,K\}$

LDA 모델에서의 파라미터 추정은 EM(Expectation Maximization), 깁스 샘플링(Gibbs Sampling) 등 다양한 추론 방법을 적용시킬 수 있다. 본고에서 LDA 모델 기반 사건 추출 실험에서의 파라미터 추정은 LDA 깁스 샘플링 오픈 소스[12]를 이용하였고, 토픽 개수 20, α값은 0.01, β값은 0.01, 반복횟수는 500으로 하였다.

〈그림 7〉은 이 모델에서 최종적으로 사건 추출하는데 필요한 확률테이블을 보여주고 있다. LDA 모델을 개선하여 날짜별 이슈가 된 토픽 그룹과 이슈와 관련된 사용자를 추출할 수 있다.

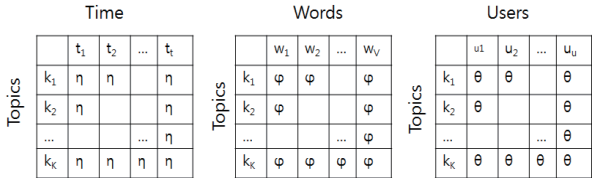


그림 7. 날짜별 토픽 단어와 사용자 확률 테이블

〈표 2〉는 사용자 행동분석과 시간 분석을 반영한 LDA 모델을 이용하여 추출한 사건 목록이다. 날짜별로 확률값이 높은 단어와 사건과 직접적으로 관련이 높은 사용자 그룹이 이슈 그룹으로 선택되었다.

표 2. LDA 토픽 모델의 사건 추출 예제

사건 (Topics)	상위 사건 어휘 (Top Words)	상위 사용자 (Top Users)
일본 지진	일본 지진 진도 8.9 지진해일 경보 일본 쓰나미	국가 기상서비스 주일본 한국영사관 일반사용자 일반사용자
방사능 유출	일본 대지진 방사능 유출 원전 방사능 지진 정보	일반사용자 일반사용자 일반사용자 일반사용자

실험을 통한 평가에서 48개 사건에서 상위 사건 어휘 10개에 대하여 정답포함률(accuracy)을 평가하기 위해 사건이 발생한 날짜의 모든 데이터에서 20개 토픽을 추출하였다. 비교분석을 위해 각 단계에서의 시간과 감성 자질을 사용한 사

건 추출(ChiOpScore) 방법과 기존의 LDA 모델을 이용한 사건 추출 (LDA모델), 시간과 사용자 행동 분석 기반 LDA 모델(TimeUserLDA)을 이용한 사건 추출 비교 실험에서 ChiOpScore 방법이 85.1%, LDA 모델이 84.3%, TimeUserLDA 모델이 94.8%의 성능을 보였다. 시간 변화 분석과 신뢰성있는 사용자 분석이 사건추출에 도움을 주고 있음을 확인할 수 있었다.

V. 결 론

본고에서는 활발한 사용자의 수가 아주 많고 매일 생성되는 트윗 글이 수억이 넘는 소셜 빅데이터 환경에서 사회적으로 큰 이슈가 되는 사건을 추출하기 위한 방법으로 시간분석과 사용자 분석을 반영한 토픽 모델링 기법을 소개하였다. 새로운 사회적 이슈가 발생했을 때 트윗의 개수가 크게 증가하고, 재난재해와 같이 사회적으로 이슈가 되는 글에서는 감성표현이 많다는 관찰을 통해 시간 분석과 감성 분석을 통해 사건 자질을 추출하였다. 또한 그 이슈와 관련해서 문서를 작성한 사용자들의 신뢰성 분석을 통하여 트위터 사용자를 신뢰성과 활동성이 높은 사용자, 일반 사용자, 활동력이 낮은 사용자로 분류함으로써 이슈에 대해 신뢰성이 높은 사용자를 분류하였다.

핵심사건을 추출하는 방법으로 기존의 LDA 토픽모델에 시간분석과 신뢰성 있는 사용자 분석을 반영하여 같은 날짜에 발생한 여러 사건들을 효율적으로 추출하는 토픽모델링 방법을 소개하였다.

참 고 문 헌

[1] Mendoza, M., Poblete, B. & Castillo, C. Twitter under crisis: Can we trust what we rt? In 1st Workshop on Social Media Analytics (SOMA '10). ACM Press, July 2010.

[2] Kanhabua, N. & Nejdl, W. Understanding the diversity of tweets in the time of outbreaks. In Proceedings of the 22nd international conference companion on World Wide Web, pp. 1335-1342. 2013.

[3] Benson, E., Haghighi, A., & Barzilay, R. Event discovery in social media feeds. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 389-398). Association for Computational Linguistics, 2011.

[4] Wikipedia, "Twitter", http://en.wikipedia.org/wiki/Twitter, 2017

[5] Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent Dirichlet Allocation. The Journal of Machine Learning research 3, pp. 993-1022

[6] Tsolmon, B. & Lee, K.-S. " A Graph-based Reliable User Classification ", Lecture Notes in Electrical Engineering 285, pp. 61-68, Springer Verlag. 2013.

[7] Kleinberg J. M. "Authoritative Sources in a Hyperlinked Environment", Journal of the ACM, 46(5), pp. 604-632, 1999

[8] Tsolmon, B. & Lee, K.-S. " Extracting Social Events based on Latent Dirichlet Allocation with Time and User Analysis ", Proceeding of the 37th Annual International ACM SIGIR Conference(SIGIR2014), pp. 1187-1190, 2014.

[9] Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. The author-topic model for authors and documents. In Proceedings of the 20th conference on Uncertainty in artificial intelligence (pp. 487-494). AUAI Press, 2004

[10] Griffiths, T. L., & Steyvers, M. Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235, 2004.

[11] Diao, Q., Jiang, J., Zhu, F. & Lim, E.P. Finding bursty topics from microblogs. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 536-544. 2012.

[12] GibbsLDA++: A C/C++ Implementation of Latent Dirichlet Allocation, http://gibbslda.sourceforge.net/

약 력

촐몽 바야르
(Tsolmon, Bayar)

2012년 전북대학교 컴퓨터공학부 학사
2014년 전북대학교 컴퓨터공학전공 석사
2014년~현재 벤처기업(Ametros Solutions) 소프트웨어 엔지니어
관심분야: 정보마이닝, 사물인터넷, 운영체제

이 경 순

1997년 한국과학기술원 전산학과 석사
2001년 한국과학기술원 전산학과 박사
2001년~2003년 일본 국립정보학연구소 연구원
2007년~2008년 미국 메사추세츠주립대학 방문교수
2004년~현재 전북대학교 컴퓨터공학부 교수
관심분야: 정보검색, 소셜데이터마이닝