Journal of Korean Institute of Intelligent Systems, Vol. 24, No. 6, December 2014, pp. 640-645
http://dx.doi.org/10.5391/JKIIS.2014.24.6.640

토픽 모델링을 이용한 댓글 그래프 기반 소셜 마이닝 기법

A Reply Graph-based Social Mining Method with Topic Modeling

이상연* · 이건명*†
Sang Yeon Lee, and Keon Myung Lee†

*충북대학교 컴퓨터과학과
*Department of Computer Science, Chungbuk National University

요 약

인터넷 상에서 많은 사람들은 사용자 간의 의사소통과 정보 공유, 사회적 관계를 생성하기 위한 방법으로 소셜 네트워크 서비스를 이용한다. 그 중 대표적인 트위터는 하루에 수백만 건의 소셜 데이터가 발생하기 때문에 수집되고 있는 데이터의 양이 엄청나다. 이 방대한 양의 데이터로부터 의미 있는 정보를 추출하는 소셜 마이닝이 집중적으로 연구되고 있다. 트위터는 일반적으로 유용한 정보 혹은 공유하고자 하는 내용을 팔로잉-팔로워 관계를 이용해 쉽게 전달하고 리트윗할 수 있다. 소셜 미디어에서 트윗 데이터에 대한 토픽 모델링은 이슈를 추적하기 위한 좋은 도구이다. 짧은 텍스트 기반인 트윗 데이터의 제한점을 극복하기 위해, 사용자를 노드로 사용자간 댓글과 리트윗 메시지의 여부를 간선으로 하는 그래프 구조를 갖는 댓글 그래프의 개념을 소개한다. 토픽 모델링의 대표적인 방법인 LDA 토픽 모델이 짧은 텍스트 데이터에 대해 비효율적인 것을 보완하기 위한 방법으로, 이 논문에서는 짧은 문서의 수를 줄이고 마이닝 결과의 질을 향상시키기 위한 댓글 그래프를 사용하는 토픽 모델링 방법을 소개한다. 제안한 모델은 토픽 모델링 방법으로 LDA 모델을 사용하였으며, 7일간 수집한 트윗 데이터에 대한 실험 결과를 보인다.

키워드 : 소셜 네트워크 서비스, 트위터, 소셜 마이닝, 댓글 그래프, 트렌드 추출, 토픽 모델링, LDA

Abstract

Many people use social network services as to communicate, to share an information and to build social relationships between others on the Internet. Twitter is such a representative service, where millions of tweets are posted a day and a huge amount of data collection has been being accumulated. Social mining that extracts the meaningful information from the massive data has been intensively studied. Typically, Twitter easily can deliver and retweet the contents using the following-follower relationships. Topic modeling in tweet data is a good tool for issue tracking in social media. To overcome the restrictions of short contents in tweets, we introduce a notion of reply graph which is constructed as a graph structure of which nodes correspond to users and of which edges correspond to existence of reply and retweet messages between the users. The LDA topic model, which is a typical method of topic modeling, is ineffective for short textual data. This paper introduces a topic modeling method that uses reply graph to reduce the number of short documents and to improve the quality of mining results. The proposed model uses the LDA model as the topic modeling framework for tweet issue tracking. Some experimental results of the proposed method are presented for a collection of Twitter data of 7 days.

Key Words : Social network services, Twitter, Social mining, Reply graph, Tread extraction, Topic modeling, LDA

1. 서 론

소셜 네트워크 서비스(Social Network Service, SNS)는 온라인상에서 인맥을 새롭게 쌓거나, 기존 인맥과의 관계를

접수일자: 2014년 9월 14일
심사(수정)일자: 2014년 9월 28일
게재확정일자 : 2014년 12월 12일
† Corresponding author
본 연구는 미래창조과학부 및 정보통신산업진흥원의 대학IT연구센터육성 지원사업/IT융합고급인력과정지원사업의 연구결과로 수행되었음(NIPA-2014-H0301-14- 1022).
This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA(National IT Industry Promotion Agency)

강화시킬 수 있는 서비스를 제공하고 개인의 관심, 취미, 사회적 위치 등 동질감이 있는 사람들 간에 커뮤니티를 만들어 준다. SNS 중 대표적인 서비스인 트위터는 트윗(tweet)과 리트윗(retweet)의 개념을 도입하여 공유하고자 하는 내용을 쉽게 전파할 수 있다.[1] 특히 리트윗은 팔로잉-팔로워(following-follower)를 통해 특별한 조건 없이 쉽게 관계를 맺을 수 있기 때문에 공감과 공유의 전파 속도를 빠르게 한다. 이러한 이유로 트위터는 사회적 이슈의 출현과 변화를 충분히 반영을 한다. 트윗 데이터를 분석하여 사회적 이슈 혹은 트렌드를 찾는 소셜 마이닝 분야가 활발히 진행되고 있고 트렌드를 이용하여 상업적으로 마케팅을 하는 수많은 웹사이트가 출현하였다.[2-4]

기존에는 트렌드를 추출하기 위한 방법으로써 토픽 모델을 이용한 트렌드 추출 기법으로 짧은 텍스트 데이터인 소셜 데이터에 비효율적인 토픽 모델링을 했었다. 이 논문에서는 댓글 그래프를 기반으로 유사 사용자간 군집을 연결요소로 추출하고 문서의 단위를 군집으로 표현하여 보다 토픽 모델링에 적합한 데이터 마이닝 기법을 제안한다.

2. 관련 연구

2.1 소셜 마이닝

SNS는 사회적 관계를 오프라인에서 온라인상으로 이동시킨 매개체로 기존 특정 주제를 중심으로 소통하던 방식과 달리 각각의 개인을 중심으로 공통의 관심사를 가지고 있는 사용자 사이에서 관계를 형성하고 이 관계를 통해 소통 및 공유가 일어난다. SNS는 사용하기 쉽고 전파 속도도 빠르고 매일 수억 건씩 방대한 양의 데이터를 생성하기 때문에 소셜 빅데이터라고 불린다. 따라서 이 방대한 양의 데이터로부터 의미 있는 정보를 찾아내려는 소셜 마이닝 분야가 등장하였다.

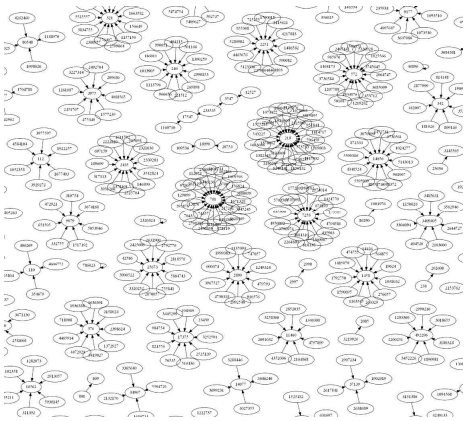

그림 1. 트위터에서 댓글과 리트윗을 이용한 소셜 그래프
Fig. 1. A Social Graph using Replys and Retweets from Twitter

소셜 마이닝은 SNS에 올라오는 글과 사용자를 분석해 사용자의 성향과 패턴 등을 분석하는 것으로 대표적인 그래프 알고리즘인 페이지랭크(PageRank)를 이용하여 가장 영향력 있는 사용자를 찾는 연구부터 소셜 그래프를 활용하여 커뮤니티(community)를 찾아 분석하는 연구까지 다양한 방면으로 진행되고 있다. 소셜 마이닝은 소비자를 분석함으로써 판매 및 홍보로 사용되고 여론변화나 사회적 흐름과 트렌드를 찾기 위해 사용되는 마이닝 기법이다.[5] 그 예로 18대 대통령선거에서는 이 기법을 이용하여 SNS를 통한 전략으로 내세워 홍보하는 역할을 하였다.

(그림 1)은 직접 수집한 트윗 데이터로부터 소셜 그래프를 만들기 위해 Graphviz를 이용한 단방향 그래프(Directed Graph)이다.

2.2 토픽 모델

토픽 모델(topic)은 텍스트 기반의 문서를 활용하기 위해 개발된 확률 모델이다.[6,7] 문서를 표현하기 위해 단어들에 대한 벡터 또는 용어 집합(bag-of-words)을 이용한다. 토픽 모델 중 잘 알려진 LDA(Latent Dirichlet Allocation)는 많은 문서들 안에서 잠재적으로 의미 있는 토픽을 발견하기 위한 확률적인 생성모델이다.[8] LDA는 Dirichlet 분포를 이용하여 텍스트 문서 내의 단어들이 어떤 특정 토픽에 포함될 확률을 계산한다. 여기서 각 문서들은 하나의 토픽이 아니라 여러 개의 토픽에 의해 확률적으로 표현되고, 각 토픽들은 단어들에 대한 특정 분포에 의해 표현되어진다. LDA중 MCMC(Markov Chain Monte Carlo) 알고리즘의 하나로 분포 혼합을 추정하는 깁스 샘플링(Gibbs sampling)은 주어진 모수에 의해 자료를 분류하고 그 분류에 의해 다시 모수를 추정한다. (그림 2)는 LDA에 대한 Plate Notation을 표현한 그림으로 α는 문서 당 토픽의 분포에 Dirichlet 사전확률(prior)에 대한 파라미터이고, β는 토픽 당 단어 분포에 Dirichlet 사후확률(posterior)에 대한 파라미터이다. 그리고 θ는 문서에 대한 토픽 분포, z는 특정 문서의 특정 단어에 대한 토픽, w는 특정 단어를 의미한다. LDA는 파라미터 α, β, N, M으로부터 θ와 z을 추정한다.

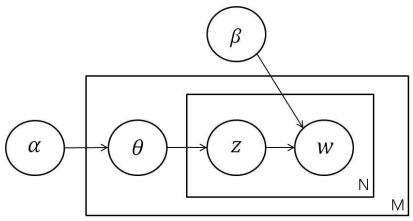

그림 2. LDA 토픽 모델
Fig. 2. LDA Topic model

2.3 그래프 클러스터링

클러스터링은 기계학습, 데이터마이닝 등 많은 분야에서 사용되고 있는 비모수적 분석 방법이다. 클러스터링이란 전체 데이터로부터 서로 간의 관계와 속성을 분류하여 비슷한 것끼리 군집을 이루도록 만드는 알고리즘이다. 그래프 클러스터링은 노드의 속성과 간선의 값을 이용하여 클러스터링을 할 수 있다. 각 클러스터에서는 내부적으로 간선이 많이 이어져 있고 서로 다른 클러스터 사이의 간선은 적게 군집하도록 한다. 클러스터링 방식의 대표적인 두 가지 알고리즘은 계층적 클러스터링(Hierarchical Clustering)과 파티션 클러스터링(Partition Clustering)이 존재한다. 계층적 클러스터링은 데이터간의 거리를 이용하여 트리를 만들고 그 트리로부터 가장 가까운 군집을 병합시키는 방법으로 상

향식 접근 방법을 이용한다. 파티션 클러스터링은 데이터를 k개의 파티션으로 나누는 방법으로 각각의 파티션은 하나의 군집으로 표현된다.

3. 댓글 그래프 기반 소셜 마이닝 기법

제안된 기법을 이용하기 위해서는 실시간으로 발생하고 있는 SNS 데이터를 스트림으로 수집해야 한다. 트위터와 같은 경우에는 기본적으로 오픈 소스로 제공되어 있으므로 라이브러리를 이용하여 수집할 수 있다. 이 제안된 기법은 사회적 이슈 혹은 트렌드에 대한 영향을 알아보기 때문에 하루 이상의 일정기간동안 수집해야 한다.

3.1 데이터 수집 및 전처리

데이터 수집 대상은 토픽 모델을 이용하기 때문에 텍스트 기반으로 이루어져 있고 댓글을 제공하는 SNS를 선택해야한다. 여기서 댓글은 댓글 그래프를 만드는데 있어 중요한 역할을 한다. 기본적으로 데이터는 다음과 같은 형태로 수집한다.

$$SD_i = (ID, date, lang, reply, txt)\,(i = 1, 2, \ldots, n)$$

ID는 데이터를 올린 사용자의 식별자이고 $date$는 이 글을 올린 날짜, $lang$은 SNS가 전 세계로 서비스하고 있을 경우 사용되는 언어 정보를 담고 있다. $reply$는 현재의 글이 누구로부터의 댓글을 한 글인지에 대한 사용자 정보가 들어가고 마지막으로 txt는 이 글을 올린 내용이 담긴다. 이 정보를 토대로 XML형식으로 저장도 가능하고 데이터베이스로도 저장할 수가 있다. 수집한 데이터의 크기는 수집한 기간과 SNS에 따라 다르지만 트위터를 사용하는 전 세계 사용자를 대상으로 한다면 하루에 수억 건의 데이터가 쌓이게 되므로 방대한 양의 데이터가 수집되고 이 데이터는 일반 컴퓨터에서 다룰 수 없는 크기가 된다.

이 데이터를 처리하기 위해서는 두 가지 방법 중 한 가지 방법을 선택해야한다. 첫 번째 방법은 빅데이터를 처리하기 위한 시스템을 사용하는 것이다. 그 중 대표적인 하둡(Hadoop)은 빅데이터 처리 프레임워크로 일반 컴퓨터 수십 대부터 수천 대의 클러스터로 구성되며, 빅데이터를 수용할 수 있는 HDFS(Hadoop Distributed File System)라는 분산 파일 시스템을 제공한다. 두 번째 방법은 데이터의 규모를 한정하는 것이다. 트위터는 전 세계로 서비스를 하지만 처리할 대상을 언어로 분류하여 영어권 혹은 한국어권으로 한정하게 되면 데이터의 규모를 줄일 수 있다.

수집한 데이터는 제일 먼저 자연어처리를 해야 한다. 이슈 혹은 트렌드를 찾는 것을 방해하는 요소인 특수 문자, 숫자, 대명사, 전치사, 정관사, URL 등의 불용어(stopword)를 지정하고 제거해야한다. 그리고 하나의 단어에 대해 파생된 접두파생어, 접미파생어, 접사, 접두사, 접미사를 제거하여 어간만 뽑아내는 어간 추출(stemming)을 해야 한다.

3.2 댓글 그래프 생성

댓글 그래프를 만들기 위해서는 사용자의 정보와 댓글 정보가 필요하다. 사용자는 노드로 표현하고 댓글의 여부를 그래프의 간선으로 표현하게 되면 댓글 그래프는 생성이 된다. 사용자 정보와 댓글 정보 또한 매우 크기 때문에 시각화를 하기 위한 메모리가 많이 사용되어 무작위 추출 방법을 사용해야 한다.

3.2.1 유사 사용자 클러스터링

이전 방법에서는 한사람으로부터 생성된 모든 트윗을 기준으로 하나의 문서를 표현하였다.[9,10] 하지만 제안된 방법은 댓글 그래프를 이용한 유사 사용자간 군집으로 문서를 표현한다. 댓글 그래프로부터 군집을 얻어내기 위한 방법으로 그래프 클러스터링 방법을 사용한다. 댓글 그래프는 사용자간의 네트워크를 댓글로써 표현하였기 때문에 네트워크 내의 커뮤니티를 찾는 방법을 이용한다. 커뮤니티를 찾는 방법에는 오버랩(overlap)을 고려하는 것과 하지 않는 방법이 있다. 여기서는 댓글의 특성을 고려하여 오버랩을 고려하는 방법을 선택하였다.

유사 사용자를 찾는 기준은 트위터에 이슈가 생성되고 이 이슈는 많은 사람들을 통해 전파와 공유가 되고, 전파와 공유를 할 수 있는 댓글로 그 내용에 대한 공감을 전달하기 때문에 특정 사용자의 글에 댓글을 남긴 사용자는 앞선 사용자와 유사하다고 판단한다. 댓글 그래프를 이용하면 기준 노드로부터 연결 요소(connected component)를 파악하고 이 연결 요소는 하나의 커뮤니티로 표현된다. 여기서 댓글의 수를 고려하여 가중치를 준다면 사용자간의 유사도를 측정할 수 있게 되고 임계값 이상의 간선만 표현한다면 각각의 커뮤니티는 보다 유사도가 높게 측정될 수 있다. 이렇게 생성된 커뮤니티는 하나의 군집으로 표현된다.

제안 방법은 하나의 노드를 기준으로 이 글이 누구의 댓글인지를 파악하고 댓글 그래프의 가장 상위에 있는 사용자를 찾게 된다. 가장 상위에 있는 사용자는 여러 사람이 될 수 있다. 찾게 된 상위 사용자로부터 댓글을 한 사람들을 찾게 되고 그 사람들에 댓글을 한 사람들을 찾는 반복적인 작업을 통해 커뮤니티를 생성하게 된다. 더 이상 커뮤니티에 속하는 사람이 추가가 되지 않는 다면 이 작업은 종료하게 된다. (그림 3)은 (그림 1)로부터 생성된 커뮤니티로 군집을 하였다.

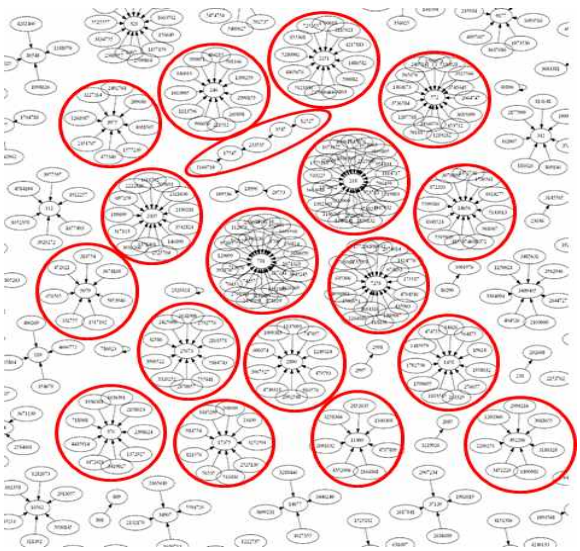

그림 3. 유사 사용자를 찾기 위한 댓글 그래프에서의 연결 요소

Fig. 3. Connected Components of Reply Graph to find similar user

3.2.2 클러스터 병합

군집을 마치고 모든 군집에 대해 두 군집간의 사용자를 비교한다. 이 과정은 두 군집 간의 사용자가 적정 기준 이상 겹치게 된다, 즉 두 군집이 유사하다면 하나의 군집으로 표현하게 된다.

두 군집을 병합하기 전 군집의 사용자 수가 10명 미만인 군집은 제거 대상이 된다. 너무 작은 군집이 하나의 문서로 표현되는데 있어서 너무 비중이 크기 때문에 제거를 해야 한다. 두 군집을 병합하기 위한 유사도 측정 방법으로 두 군집간의 거리 척도로써 자카드 계수(Jaccard coefficient)를 사용한다. 자카드 계수는 두 집합 A, B에 대해 다음과 같이 계산된다.

$$Jaccard\ coefficient = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

A와 B는 군집에 속하는 사용자 집합을 의미한다. 자카드 계수의 값이 0.8 이상이 된다면 두 군집은 충분히 같다는 것을 의미하고 두 군집은 병합을 하게 된다. 여기서 고려해야 할 사항은 $A \cap B$의 수가 B의 수와 같은 경우 $A \cap B$로 병합을 할 것인지 $A \cup B$로 할 것인지를 결정하는 것이다. 일반적인 경우는 병합 문제라면 $A \cup B$로 병합하는 것이 옳다고 할 수 있지만 군집 내의 유사도로 비교를 한다면 $A \cap B$로 병합하는 것이 더 높을 것이다. 그 이유는 군집이 이루어진 조건에 따라 군집 A와 B가 결정되었고 B는 A에 포함되므로 B라는 군집이 한 번 더 발생이 되었다는 것을 의미하기 때문에 $A \cap B$의 유사도가 높다. 이 논문에서는 두 군집이 병합할 때 유사 사용자를 중점으로 두기 때문에 $A \cap B$로 병합을 하는 것을 제안한다.

$$C = \begin{cases} A \cap B & if \quad Jaccard(A,B) > 0.8 \\ \varnothing & else \end{cases} \quad (2)$$

3.2.3 클러스터를 문서로 표현

군집의 사용자 수는 최소 10명부터 많게는 수백 수천명 이상이 되기 때문에 군집을 하나의 문서로 표현하기에 부당하다. 군집의 사용자 수에 따라 가중치를 주고 가중치에 따라 각 군집마다 표현할 문서의 수를 결정한다. 두 개 이상의 문서로 표현되는 군집의 경우에는 군집에 속한 사용자가 올린 글의 성격에 따라 분류한다. 이 방법을 사용하기 위해서는 유사 사용자간의 트윗 데이터로 하여금 그래프를 그리게 된다. 여기서의 그래프는 트윗 데이터를 노드로 표현하고 해당 글에 대한 리트윗 여부가 그래프의 간선으로 표현된다. 트윗 데이터를 가져올 시 어떤 글에 리트윗된 대상을 알 수는 있지만 어떤 글에 리트윗이 되었는지에 대한 정보는 담지 않는다. 다만 리트윗이 되었다면 글의 내용의 대부분이 같고 그 이외의 부분에 자신의 추가적인 내용을 담고 있어 텍스트의 유사도가 매우 높다는 것을 알 수 있다. 하지만 댓글의 경우에는 댓글 대상의 글과 무관하게 글을 올릴 수 있어 여기서는 제외가 된다. 이 방법은 트윗 데이터를 이용한 군집 내의 클러스터링이라고 볼 수 있다.

3.3 토픽 모델링

토픽 모델을 적용하기에 앞서 문서들은 용어 집합으로 표현해야 한다. 용어 집합은 단어 벡터와 달리 단어가 나열된 순서를 고려한다. 각 문서를 용어 집합으로 표현하고 깁스 샘플링을 적용하기 전에 파라미터 α와 β를 지정하고 토픽의 수와 반복 횟수를 지정해주어야 한다.[11] 여기서 토픽에 속한 단어들이 집합의 형태로 표현될 수 있도록 α를 1보다 큰 값으로 지정해 주어야 한다. 깁스 샘플링은 각 토픽에 해당하는 단어의 확률을 추정하는 작업을 계속적으로 반복하게 되므로 단어의 수와 문서의 수 그리고 반복 횟수에 따라 실행시간이 크게 차이난다. 이 제안된 방법은 유효한 단어의 수와 문서의 수를 줄이는 방법으로 기존 방법에 비해 실행 속도가 빠르다.

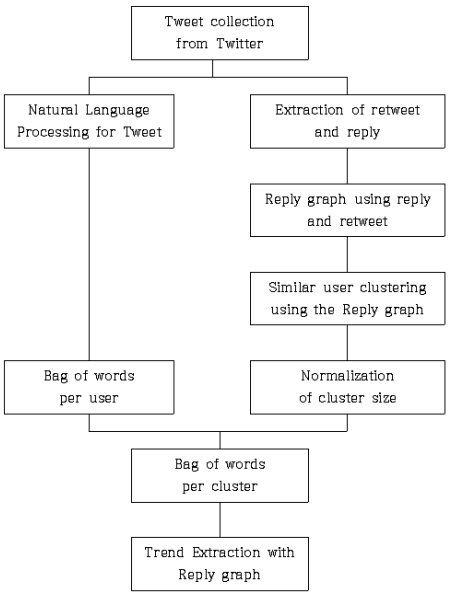

그림 4. 댓글 그래프 기반 트렌드 추출 시스템 아키텍처

Fig. 4. The System Architecture for Trend Extraction with Reply Graph

4.구현 및 실험환경

데이터는 SNS 중 트위터로 선정하였고 트윗 데이터를 수집하기 위하여 Java 기반 트위터 스트림 API인 Twitter4J(http://twitter4j.org/)를 통해 전 세계의 공개 트윗을 한국 시각 기준 2014년 9월 13일 00:00 부터 2014년 9월 19일 24:00 까지 총 7일간의 데이터를 수집하였다. 수집된 데이터의 크기는 약 6.0Gbyte이고 이 데이터는 총 24,695,089건의 트윗이며 트윗이 제공하는 언어는 67개로 되어 있다. 공개 트윗은 하루 평균 약 353만 건이 발생하고 그 중에 영어권은 약 127만 건(약 36.0%), 한국어권은 약 5.5만 건(약 1.6%)이 발생하였다. 한국어권은 공개 트윗의 수가 적기 때문에 제안된 기법을 사용하기 위해서 영어권을 기준으로 하였다. 영어권에서의 7일간 공개 트윗의 수는 8,906,567건이고 이 기간 동안 이용한 사용자 수는 5,018,532명이었다. 트윗 데이터 중 댓글의 수는 1,808,394건(20.3%)이었으며 리트윗의 수는 3,276,430건(36.8%)으로 확인되었다.

(그림 4)는 제안된 기법에 대한 시스템 아키텍처이다. 트위터를 수집하여 트렌드를 추출하기 위해선 크게

자연어처리와 군집화로 나눌 수 있다. 실험을 하기 위한 컴퓨터는 하나의 PC에서 성능 평가 하였다. PC환경은 Intel® Core™ i7-4770 3.40GHz CPU와 16GB RAM으로 Windows 7에서 실험하였다.

4.1 제안 기법 구현

총 7일간의 데이터로부터 데이터 수집은 XML파일 형식으로 수집하였으며 기존 데이터의 리트윗 정보를 담는 부분을 추가하여 저장하였다.

$$SD_i = (ID, date, lang, reply, retweet, txt)\,(i = 1, 2, \ldots, n)$$

데이터의 규모를 줄이기 위해서 대상을 공개 트윗으로 줄였기 때문에 하루에 약 1Gbyte 정도로 수집되었다는 것을 볼 수 있었다. 데이터의 규모가 아직 일반 컴퓨터에서 돌리기 부담스러울 정도의 크기이기 때문에 여기서 좀 더 데이터의 크기를 줄이기 위해 영어권 나라만 선택하여 보다 데이터의 크기를 줄였다.

수집한 데이터를 전처리하기 위해 3자 이하의 알파벳으로 구성된 단어를 제거하였고, 단어의 원형을 찾기 위한 방법으로 wordsmyth(http://www.wordsmyth.net/)에서 영단어 목록을 이용하여 검색 후 원형과 품사 정보를 가져왔다. 가져온 데이터를 통해 원형을 추출한 후 품사 중 명사와 동사를 제외한 불용어를 제거하였다. SNS는 개인적인 글을 올리는 특성이 있기 때문에 비속어와 은어가 많이 섞여있고 기존 사전으로부터 포함되지 않은 단어와 띄어쓰기를 하지 않은 형태를 찾을 수 있었다. 예를 들면 'stillnothingisimpossible'와 'whatttttttt', 'pettttty'와 같은 단어들이 포함되어 있기 때문에 이를 모아 비속어와 은어를 직접 추출해 사전을 만들었다. 수집한 트윗 데이터로부터 사전에 포함되지 않은 단어 중 100번 이상 출현한 단어가 4,861개였고 100번 이상 출현 단어들 중에서도 리트윗이 빠르게 전파하기 때문에 상당의 의미 없는 단어와 띄어쓰기가 되지 않은 단어도 많이 포함되어 있었다. 이들 중 비속어와 은어를 직접 찾아내어 제거 하였다. 마지막으로 어간 추출을 통해 파생어에 대해 처리하였고 이 처리된 단어를 토대로 용어 집합으로 표현하였다.

군집화를 하기 위해서 전처리된 트윗 데이터를 토대로 사용자와 댓글 혹은 리트윗 정보를 추출하였고 추출된 트윗 데이터 중 텍스트의 단어가 5개 미만인 경우는 제외하였다. 이렇게 모인 데이터의 댓글과 리트윗 수는 총 1,293,551개가 되었고 댓글과 리트윗을 같은 의미로 처리하여 댓글 그래프의 간선으로 표현하였다. 연결 요소로 커뮤니티를 얻기 위해서는 임의의 하나의 기준 노드로부터 상위 사용자를 찾아내고 상위 사용자로부터 댓글을 한 사용자를 찾아 추가하며 커뮤니티를 추출하였다. 이렇게 구성된 커뮤니티의 수는 약 5.0만개를 찾아내었고 이 커뮤니티가 각각 군집으로의 의미를 가지게 된다. 이 군집은 병합과정을 거치기 위해 제안한 방법인 자카드 계수를 이용하여 각각의 자카드 계수를 측정하였고 자카드 계수를 이용하여 병합된 군집의 수는 약 3.2만개로 1.8만개 정도의 군집이 제거 되었다.

4.2 토픽모델링 구현

토픽 모델에 적용시키기 위해 군집별 문서를 표현하고 각 문서를 용어 집합으로 표현하였다. 깁스 샘플링은 프로그래밍 언어 Java로 직접 구현하였고 토픽 모델을 적용하기 위해 깁스 샘플링에 대한 파라미터를 α는 25, β는 0.01로 주었고 토픽의 수는 100개 반복 횟수는 100번으로 지정하였다. α는 전체적으로 토픽에 포함된 단어의 확률을 균등하게 분배하기 위해 1보다 큰 25로 지정하였다. 이 실험의 결과로 토픽들이 단어에 대한 확률로 표현되어 있기 때문에 각 토픽에 대해 높은 확률 값을 갖는 상위 10개의 단어에 대해 추출하였다.

4.3 성능평가를 하기 위한 대조 기법 구현

평가를 하기 위한 대조 기법으로 불용어 제거, 비속어 및 은어 처리, 어간 추출을 한 데이터를 토대로 사용자 별 용어 집합으로 표현하는 전처리 과정만 하였다. 이 데이터 중 남은 데이터의 단어 수가 10개 미만인 트윗 데이터를 걸러 내었고 그 결과 7,086,947개의 영어권 공개 트윗 중 292,620개로 추출되었다. 이렇게 추출된 데이터는 하나의 사용자에 따른 용어 집합이 하나의 문서가 된다.

4.4 성능평가

성능 평가는 실행 시간, 메모리 사용, 마지막으로 정답률로 세 가지 측면에 대해 평가를 하였다. 실행시간은 깁스 샘플링에 대해서 얼마나 걸렸는지를 측정하였고, 메모리 사용은 깁스 샘플링을 돌리기 위한 메모리 사용 변화를 측정하였다. 정답률을 측정하기 위해서는 해당 기간에 나온 뉴스 기사로부터 단어가 5개 이상 출현하였을 경우를 정답으로 분류하였다.

이 실험 결과 (그림 5)와 같이 메모리 사용은 기존 기법은 약 18Gbyte로 상당한 메모리가 필요하지만 제안 기법은 약 8Gbyte로 절반 이하로 줄었다. 실행시간 또한 기존 기법은 4632초였고 제안기법은 1142초가 걸렸다. 마지막으로 정답률은 미세한 차이지만 기존 방법 대비 약 5% 향상된 것을 볼 수 있었다. 제안 기법에서는 댓글 그래프를 이용한 방법과 비속어, 은어에 대한 사전을 이용함으로써 이를 사용하지 않았던 방법에 비해 월등한 성능을 내었다.

	Proposal method	Existing method
Runtime	1142(s)	4632(s)
Memory usage	~ 8Gbyte	~ 18Gbyte
Accuracy	68%	63%

그림 5. 제안 기법과 대조 기법간의 성능 비교
Fig. 5. Comparison of Performance between Proposal Method and Existing Method

5. 결론

이 논문에서는 SNS기반 토픽 모델링을 보다 효율적으로 하기 위한 소셜 마이닝 기법을 소개하였다. 트렌드를 추출하는 방법은 이미 많이 소개되어 있지만 방대한 양의 데이터를 처리하기 위해서는 토픽 모델을 적용하기 전에 전처리 과정이 필요하다. 그 방법으로 SNS의 특성에 맞도록 짧은 텍스트 기반의 데이터를 토픽 모델에 적용하기 위해 기본적인 전처리 후 댓글 그래프를 통해 유사 사용자간 커뮤니티로 군집을 하였다. 기존 토픽 모델에서는 글 하나하나가 문서가 되었지만 제안 방법에서는 군집을 이용하여 문서의 단위를 축소시켜 학습에 필요한 반복 횟수를 줄였다. 그 결과 깁스 샘플링에서 속도와 메모리 측면에서 매우 좋은 결과를 얻을

수 있다는 것을 확인하였다.

각 토픽 별 높은 확률 값을 갖는 단어 10개를 추출하였고 추출한 기간의 뉴스기사로부터 해당하는 토픽이 기사에 출현하였는지에 따라 성능을 평가하였다. 기존 방법보다 정답 횟수가 좋았다는 것을 알 수 있었다. 향후 연구에서는 군집 내에서의 토픽과 전체 토픽과의 관계에 대해 실험할 예정이고 댓글 그래프로부터의 평판 분석에 대한 방법을 연구할 예정이다.

References

[1] H. Kwak, C. Lee, H. Park, S. Moon, "What is Twitter, a social network or a news media?," *Proceedings of the 19th international conference on World Wide Web*, pp. 591-600, 2010.

[2] M. Song and M. C. Kim, "RT2M: Real-Time Twitter Trend Mining System," *Proceedings of the IEEE 2013 International Conference on Social Intelligence and Technology*, pp. 64-71, 2013.

[3] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and Traditional Media using topic models," *Proceedings of the First Workshop on Social Media Analysis*, pp.338-349, 2011.

[4] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, Vol.3, pp.993-1022, 2003.

[5] T. L. Griffiths, M. Steyvers, "Finding scientific top-ics," *National academy of Sciences of the United States of America*, Vol.101, Suppl.1, pp.5228-5235, 2004.

[6] H. K. Peng, J. Zhu, D. Piao, R. Yan, "Retweet Modeling using Conditional Random Fields," *Proceedings of the 11th IEEE International Conference on Data Mining Workshops*, pp. 336-343, 2011.

[7] T. Hofman, "Probabilistic Latent Semantic Analysis," *Proceedings of UAI'99*, 1999.

[8] J. Weng, E. P. Lim, J. Jiang, "Twitterrank: Finding Topic-Sensitive Influential Twitterers," *Proceedings of the third ACM WSDM*, 2010.

[9] L. Hong, B. D. Davison, "Empirical Study of Topic Modeling in Twitter," *Proceedings of the SIGKDD Workshop on SMA*, 2010.

[10] D. M. Blei, "Introduction to Probabilistic Topic Models," *Communications of the ACM*, 2011.

[11] F. LU, B. Shen, J. Lin, H. Zhang, "A Method of SNS Topic Models Extraction Based on Self-Adaptively LDA Modeling," Proceedings of *2013 Third International Conference on Intelligent System Design and Engineering Applications, IEEE Computer Society*, pp.112-115, 2013.

저 자 소 개

이상연(Sang Yeon Lee)

2014년 : 충북대학교 컴퓨터공학부 학사

2014년~현재 : 충북대학교 컴퓨터과학과 석사과정

관심분야 : AI, Big Data, Machine Learning

E-mail : jaeimveilion@gmail.com

이건명(Keon Myung Lee)

1990, 1992, 1995 년: KAIST 전산학과 학사, 석사, 박사

1995년~1996년 : 프랑스 INSA de Lyon, Post-Doc.

1996년 : 미국 PSI 사, Staff Scientist

1996년 ~현재 : 충북대학교 교수

관심분야 : AI, Machine learning, Data Mining, Big data applications

E-mail : kmlee@cbnu.ac.kr