creative commons

COMMONS DEED

저작자표시-비영리-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

군집화 기법과 문서 순위를 이용한 한국어 트윗 상의 토픽 추출

Topic extraction on Korean tweets using clustering and document ranking

2013 년 8 월

서울대학교 대학원

산업공학과 데이터마이닝 전공

강 성 진

군집화 기법과 문서 순위를 이용한 한국어 트윗 상의 토픽 추출

Topic extraction on Korean tweets using clustering and document ranking

지도 교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함
2013 년 7 월

서울대학교 대학원
산업공학과 데이터마이닝 전공
강 성 진

강성진의 공학석사 학위논문을 인준함
2013 년 7 월

위 원 장 _____ 윤 명 환 _____ (인)

부위원장 _____ 조 성 준 _____ (인)

위 원 _____ 박 종 헌 _____ (인)

초 록

트위터(Twitter)는 전세계적으로 가장 많이 사용되는 마이크로블로깅 서비스(microblogging service)로 실시간 정보 공유가 가능한 특성을 지니고 있기 때문에 많은 대중들에게 빠르게 영향을 준다. 하지만 트위터 상에서 의미있는 정보를 찾는 과정에는 어려움이 많다. 하루에 작성되는 트윗(tweet)의 양은 2억 건이 넘었고, 이 중 대부분은 의미 없는 잡담과 중복되는 내용이다. 그렇기 때문에 트위터 상에서 정보를 얻으려면 같은 토픽에 대해서는 하나의 묶음으로 정보를 제공해야 할 필요성이 존재한다. 이와 관련해 활발한 연구가 이루어지고 있는 영어 트위터와는 달리 한국어 트위터에 대한 연구는 전무한 실정이다.

따라서 본 연구에서는 한국어 트윗 상에서 토픽을 추출하는 전 과정을 제안한다. 연구에서 소개하는 방법은 네 단계로 이루어진다. 첫 단계에서는 트위터에서 제공하는 오픈 API를 이용하여 한국어로 작성된 트윗들을 수집한다. 두 번째 단계에서는 수집한 트윗들을 분석이 가능한 벡터 스페이스 모델 형태로 변환한다. 각 트윗에서 명사만을 추출하여 분석에 사용한다. 기존 사전에 '네이버 실시간 검색어'와 '네이버 핫토픽 키워드'들을 추가한 사용자 정의 사전을 사용하여 신조어, 약어 등 인터넷 공간에서 실제로 사용되는 언어를 보강한다. 세 번째 단계에서는 군집화 기법을 통해 트윗들을 주제 별로 하나의 묶음으로 나눈다. 네 번재 단계에서는 세 번째 단계에서 생성된 군집 내에서 각 군집을 잘 설명할 수 있는 트윗을 선정하기 위해 트윗들의 순위를 매기고, 상위 순위 트윗들을

이용하여 토픽을 추출한다.

군집화 결과 총 평균이 99.4%의 순수도(purity)를 보임에 따라 각 주제별로 트윗의 군집화가 잘 이루어졌음을 알 수 있었다. 또한 정성적인 내용 분석을 통하여 군집 별로 상위 순위 트윗들과 하위 순위 트윗들을 비교함으로써 상위 순위 트윗들로 토픽을 추출하는 것이 각 군집을 대표하는 내용을 잘 반영하는 것으로 나타났다.

주요어 : 트위터, 한국어 트윗, 군집화, 문서 순위, 토픽 추출

학　번 : 2011-23454

목 차

제 1 장 서 론 1
제 2 장 기존 연구 6
2.1. 텍스트 군집화 관련 연구 7
2.2. 토픽 추출 관련 연구 8
제 3 장 연구 방법 9
3.1. 트윗 수집 10
3.2. 데이터 전처리 12
3.3. 트윗 군집화 19
3.4. 트윗 순위 선정 23
제 4 장 실험 설정 및 결과 25
4.1. 데이터 설명 25
4.2. 실험 설정 28
4.3. 실험 결과 31
제 5장 결 론 46
5.1. 결론 46
5.2. 향후 연구 방향 47
참고 문헌 49

표 목차

[표 1] 사용자 정의 사전에 추가될 단어의 예 13
[표 2] 데이터 클렌징 규칙의 예 15
[표 3] k-평균 군집화 알고리즘 20
[표 4] 계층적 군집화 알고리즘 21
[표 5] 트윗 순위 선정 알고리즘 23
[표 6] 분석에 사용한 트윗 데이터 주제 25
[표 7] 분석에 사용된 트윗과 명사 수 27
[표 8] 군집 수에 따른 평균 silhouette 값 (k-평균 군집화) 31
[표 9] k-평균 군집화 순수도 결과 32
[표 10] 군집 수에 따른 평균 silhouette 값 (계층적 군집화) 33
[표 11] 계층적 군집화 순수도 결과 34
[표 12] 군집 1의 상위 순위 트윗 결과 35
[표 13] 군집 1의 하위 순위 트윗 결과 36
[표 14] 군집 2의 상위 순위 트윗 결과 37
[표 15] 군집 2의 하위 순위 트윗 결과 38
[표 16] 군집 3의 상위 순위 트윗 결과 39
[표 17] 군집 3의 하위 순위 트윗 결과 39
[표 18] 군집 4의 상위 순위 트윗 결과 40
[표 19] 군집 4의 하위 순위 트윗 결과 41
[표 20] 군집 5의 상위 순위 트윗 결과 42
[표 21] 군집 5의 하위 순위 트윗 결과 42
[표 22] 군집 6의 상위 순위 트윗 결과 43
[표 23] 군집 6의 하위 순위 트윗 결과 44

그림 목차

[그림 1] 트위터 사용자 인터페이스 2
[그림 2] 트위터 ‘실시간 트렌드’ 서비스 3
[그림 3] 트윗 토픽 추출 단계별 과정 5
[그림 4] 한국어 트윗 상의 토픽 추출 전체 모형 9
[그림 5] 사용자 정보 수집 방식 10
[그림 6] 수집된 트윗 저장 방식 10
[그림 7] 트위터 데이터베이스 테이블 11
[그림 8] 한나눔 형태소 분석기 작업 흐름도 14
[그림 9] TF-IDF를 이용한 벡터 스페이스 모델 예시 18
[그림 10] 군집화 과정 예시 19
[그림 11] 계층적 군집화 예시 21
[그림 12] 분석에 사용한 트윗 예시 26
[그림 13] 계층적 군집화 덴드로그램 34

제 1 장 서 론

정보의 홍수 시대에 사는 우리는 다양한 채널을 통해서 수많은 정보를 접하며 살아간다. 전통적인 뉴스 채널인 신문, TV를 통해 정보를 접하기도 하고, 인터넷의 보급으로 대중화된 온라인 매체들을 통해 정보를 획득할 수도 있다. 이러한 채널들을 통한 정보 획득 이외에 최근 부상하고 있는 뉴스 채널로 마이크로 블로깅 서비스(Microblogging service)를 들 수 있다. TV, 라디오와 같은 기존 매체들과는 다르게 양방향 정보 교류가 가능하다는 점에서 마이크로 블로깅 서비스에 관한 관심이 증대되고 있다(Wu et al. 2011).

많은 사람들이 사용하는 대표적인 마이크로 블로깅 서비스로는 트위터(Twitter), 페이스북(Facebook), 미투데이 등이 있다. 이 중, 트위터는 전세계적으로 가장 많이 사용되는 마이크로 블로깅 서비스이다(Suh et al. 2010). 트위터의 사용자 인터페이스는 [그림 1]과 같다. 사람들은 트위터를 통해 트윗(tweet)이라 불리는 140자 이내의 단문 메시지를 자유롭게 올릴 수 있고 자신이 원하는 사람이 올린 트윗들을 팔로잉(following)을 통해 받아 볼 수 있다. 그리고 리트윗(retweet)을 통해 다른 사람이 올린 트윗을 재생산해 낼 수 있고 멘션(mention)과 답장(reply)를 통해 특정 사람에게 메시지를 전달할 수 있다. 이와 같은 자유롭고 다양한 사용자 환경을 바탕으로 수많은 이야기들이 트위터 상에서 빠르게 퍼져나간다.

[그림 1] 트위터 사용자 인터페이스

트위터는 실시간으로 정보 전달이 가능하여 상대적으로 많은 대중들에게 영향을 줄 수 있다(Java et al. 2007). 특히, 영향력 있는 몇몇 사용자들을 통한 정보 전파 특성과 다수의 일반 사용자들을 통한 입소문(word-of-mouth)의 특성은 트위터를 일종의 뉴스 채널이라고 볼 수 있다(Kwak et al. 2010).

하지만 정보를 획득하기 위한 뉴스 채널로써 트위터는 몇 가지 한계점을 지니고 있다. 트위터 상에는 의미 없는 잡담과 중복되는 내용이 많아 올바른 정보를 찾아내는 일은 어렵다(Becker et al. 2011). 실제로 트위터에서 제공하는 검색 서비스를 이용하여 특정 키워드에 대한 검색 결과에는 키워드와 관련된 핵심 내용이 아닌

잡담이나 스팸(spam) 트윗이 상당수 포함되어 있다. 트위터에 하루에 올라오는 트윗 양은 2011년에 이미 2억 건을 넘을 정도로 방대한 양의 정보가 생성된다[1]. 이와 같이 사용자가 선택하기 어려울 정도로 많은 양의 트윗이 생성되고 있다는 점도 정보 획득을 힘들게 하는 하나의 장애물이라고 볼 수 있다.

[그림 2] 트위터 '실시간 트렌드' 서비스

트위터는 자체적으로 특정 시점에서 화제가 되고 있는 키워드를 제시해주는 '실시간 트렌드' 서비스를 제공하고 있다. [그림 2]와 같이 매 순간 10개의 실시간 트렌드를 제공해주고 각각을 클릭하면 이와 관련된 트윗들을 화면 오른편에 제공해 준다. 이 때 리트윗이 많이 된 트윗이 '인기 트윗'이라는 이름으로 가장 상단에 나오고, 그 밑으로는 시간 순으로 관련 트윗들이 나온다. 하지만 이와 같은 방식으로는 키워드와 관련된 정확한 이야기를 찾기 어렵다. 그 이유로 첫 째, 리트윗이 많이 된 인기

[1] 트위터 공식 블로그 (blog.twitter.com)

트윗이 대표성을 갖기에 적절하지 않을 수 있다. 한 예로 정확한 정보를 제공하지 않는 유명인의 트윗이 자주 리트윗 되어 인기 트윗에 등장하는 경우가 종종 있다. 둘 째, 나머지 트윗들은 시간 흐름으로 나열되어 정보 파악을 위해서는 많은 양의 트윗을 살펴보는 방법 밖에 없다. 마지막으로 실시간 트렌드 자체에 문제가 있는 경우도 있다. 실시간 트렌드는 종종 약어, 기호, 온라인 용어를 포함하고 있다. 이러한 단어들은 일반 사용자들이 쉽게 파악하기 어려운 부분이라 오히려 혼란을 가져다 줄 수 있다.

트위터는 현재 뉴스 매체로의 가치를 가지고 있지만 이를 제대로 활용할 수 있는 시스템이 갖춰지지 않았다. 최근 이러한 문제점의 대안으로서 트윗 상에서의 토픽 추출 연구가 활발하게 진행되고 있다(Sankaranarayanan et al. 2009; Becker et al. 2011). 토픽이란 화제 혹은 이야깃거리를 의미한다[2]. 일반적으로 트위터 분석에서 토픽의 의미는 '화제가 되는 키워드' 혹은 '화제가 되는 키워드와 관련된 이야깃거리'라고 정의 된다.

하지만 지금까지 트윗 상에서 토픽을 추출하는 연구는 대부분 영어 트윗을 기반으로 한 분석에 국한되어 왔다. 그 이유로는 (1) 자유롭게 사용이 가능한 공개된 영어 트윗 데이터의 존재, (2) 다수의 공개된 전처리 모듈을 이용한 전처리 과정의 수월함을 들 수 있다. 이와는 반대로 한국어 트윗의 경우 공개된 데이터가 없고, 전처리를 제대로 할 수 있는 모듈이 존재하지 않는다. 특히 한국어와 영어의 언어적 차이 때문에 영어 트윗 상에서 토픽을 추출하는 연구 방법들을 한국어 트윗에 직접 적용이 어려운 경우가 많다. 이러한 문제를 해결하기 위해 한국어 트윗 상에서 토픽을

[2] 네이버 국어 사전 (krdic.naver.com)

추출하는 전 과정에 대한 연구의 필요성이 대두되고 있다.

[그림 3] 트윗 토픽 추출 단계별 과정

본 연구에서는 한국어 트윗 상에서 토픽을 추출하는 전체 체계(framework)를 제안한다. 트윗 상에서 토픽을 추출하는 과정은 [그림 3]과 같이 단계별로 나누어진다. 우선 트위터에서 한국어 트윗을 수집하고 수집된 데이터를 분석이 가능한 형태로 전처리를 수행한다. 전처리 과정이 끝난 데이터에 대해 군집화를 수행하고, 마지막으로 군집 별로 설명력를 기준으로 트윗 순위를 선정한다. 이 결과 나오는 상위 순위 트윗들을 통해 토픽을 추출해낸다.

제 2 장 기존 연구

다양한 마이크로 블로깅 서비스 중 특히 트위터에 대한 분석은 매우 빠르게 성장하는 연구 분야이다(Ellen, J. 2011). 하지만 트위터와 마이크로 블로깅 서비스에 관련된 상당수의 초기 연구들은 (1) 서비스 자체의 사용법을 논하고 커뮤니티를 찾는데 집중하거나(Java et al. 2007), (2) 마이크로 블로깅의 역할에 대해 논하거나(Zhao & Rosson 2009), (3) 트위터 생태계 전반에 대해 연구하였다(Kwak et al. 2010). 초기에는 마이크로 블로깅 연구 분야에 새롭게 등장하는 문제들의 정의가 제대로 되어있지 않았기 때문에 내용적인 분석보다는 역할, 사용법 등에 대한 연구가 선행되어 왔다.

최근 들어 트위터를 비롯한 마이크로 블로깅 서비스들의 내용을 분석하는 연구들이 활발히 진행되고 있다. 특히 군집화와 토픽 추출을 이용한 마이크로 블로깅 서비스 분석은 가장 대표적인 연구 분야라고 할 수 있다.

2.1. 텍스트 군집화 관련 연구

문서들의 군집화에 대한 연구는 트위터와 같은 마이크로 블로깅 서비스가 시작되기 이전부터 활발하게 진행되어 왔다. Kim & Lee(2004)는 온라인 문서에 대해 군집화와 범주화(categorization)를 이용하여 정보를 분류하는 시스템을 구축하였다. Dittenbach et al.(2002)는 문서들을 계층적으로 군집화하기 위해 기존의 SOM(Self-organizing Map)에서 변형된 GHSOM(Growing hierarchical Self-organizing Map)를 제안하였다. 그러나 위와 같은 전통적인 문서 군집화 기법들은 대체로 웹 페이지와 같은 긴 문장에 적용되는 경우가 많다.

짧은 문장에 군집화를 적용시킨 연구 역시 다수 존재한다. Benerjee et al.(2007)은 위키피디아(Wikipedia)에서 추가적인 특성을 정의하고 이를 이용하여 군집화 성능을 향상시키는 방법을 제안하였다. 하지만 이 방법은 문법이 잘 지켜지지 않는 트위터 데이터에 적합하지 않다. Zeng et al.(2004)는 군집화 문제를 순위 문제로 변환하여 생각하였다. 검색 엔진에 보낸 쿼리와 그에 따른 결과로 나온 문서들의 중요도 순위 레이블(label)을 수작업으로 분류하였다. 계산된 순위를 이용하여 회귀 모형을 학습하고 이 결과를 이용하여 선택된 중요 구문들을 바탕으로 군집화를 수행하였다. 그러나 이 방법은 사람이 직접 중요도를 판단하여 수작업으로 순위를 입력하는 과정이 포함되어 시간이 오래 걸리는 단점이 있다.

2.2. 토픽 추출 관련 연구

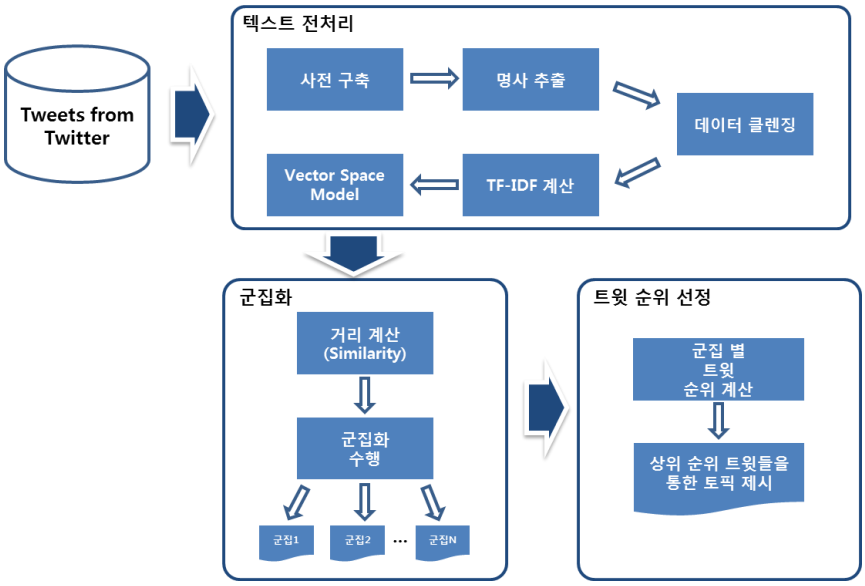
Bossard et al.(2009)는 사용자 쿼리 데이터에 대한 분석을 수행하였다. 모든 사용자 쿼리 간의 유사도(similarity)를 'Jaccard' 척도로 계산하였고 이 결과를 가지고 k-평균 군집화를 수행하였다. 각 군집 별로 군집 중심에 가장 가까운 하나의 문장을 추출하여 요약 정보로 제공하는 방법을 제안하였다.

Becker et al.(2011)은 트위터를 이용하여 어떠한 사건을 탐지(event detection)하는 모델을 제안하였다. 우선 4가지 종류의 특성들을 정의하였다. 이 4가지는 각각 시간과 관련된 'temporal features', 트위터에서 사람들 간의 소통을 나타내는 리트윗, 답장 등과 관련된 'social features', 특정 토픽과 관련된 'topical features', 마지막으로 트위터에서만 쓰이는 해쉬 태그(hash-tag, #)와 같은 특성인 'Twitter-centric features'이다. 그리고 이를 이용하여 트윗들을 군집화한다. 이 결과 나온 군집 단위로 앞서 정의한 4가지 종류의 특성들을 다시 구하고 Support vector machine(SVM)을 이용하여 새롭게 들어오는 트윗을 자동으로 분류하는 모델을 제안하였다.

Kim et al.(2012)는 핵심 토픽(core topic)을 기반으로 트윗을 군집화하는 Core-topic based clustering(CTC)를 제안하였다. 핵심 토픽은 따옴표 ("")사이에 있거나 대문자로 이루어진 단어나 구문들을 뽑아 선정한다. 군집화 결과를 평가하여 상위 k 군집을 선정하고 여기에 포함된 핵심 토픽들을 통해 핵심 정보를 파악하는 것이 가능하다.

제 3 장 연구 방법

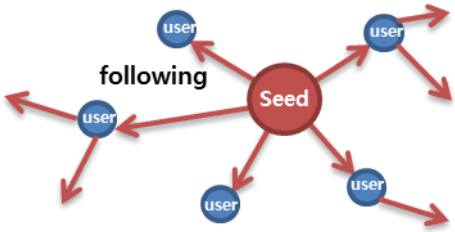
본 장에서는 이번 연구에서 제안된 한국어 트윗 상의 토픽 추출 전 과정을 기술한다. 전체적인 과정은 [그림 4]와 같다. 각 과정은 크게 트윗 수집, 텍스트 전처리, 군집화, 트윗 순위 선정으로 나누어 생각해 볼 수 있다. 각각의 과정에 대한 세부 설명은 이어지는 절에서 상세히 기술한다. 3.1 절에서는 트윗을 수집하는 방법과 이를 저장한 데이터베이스에 대해 기술한다. 3.2 절에서는 사전 구축, 명사 추출, 데이터 클렌징, TF-IDF 계산, 벡터 스페이스 모델 구축을 포함하는 텍스트 전처리 전 과정을 기술한다. 3.3 절에서는 텍스트를 군집화하는 기법에 대하여 기술하고, 3.4 절에서는 문서 간의 유사도를 이용한 순위 선정 방법에 대하여 기술한다.



[그림 4] 한국어 트윗 상의 토픽 추출 전체 모형

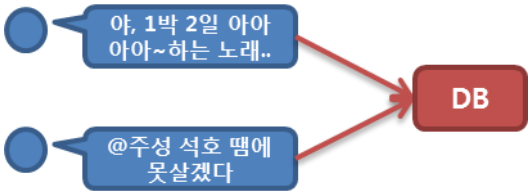
3.1. 트윗 수집

트위터에서 제공하는 API(Application Programming Interface)를 이용하여 한국어 트윗을 수집한다. 이 때 트윗을 수집하는 방식은 크게 두 단계로 나누어 진행된다. 첫 번째 단계에서는 사용자의 프로필(profile)과 관계(relation)를 수집한다. [그림 5]와 같이 특정 사용자를 시드(seed)로 이용하여 해당 사용자를 팔로잉하는 사용자와 해당 사용자가 팔로잉하는 사용자들의 계정 정보를 순서대로 수집한다.



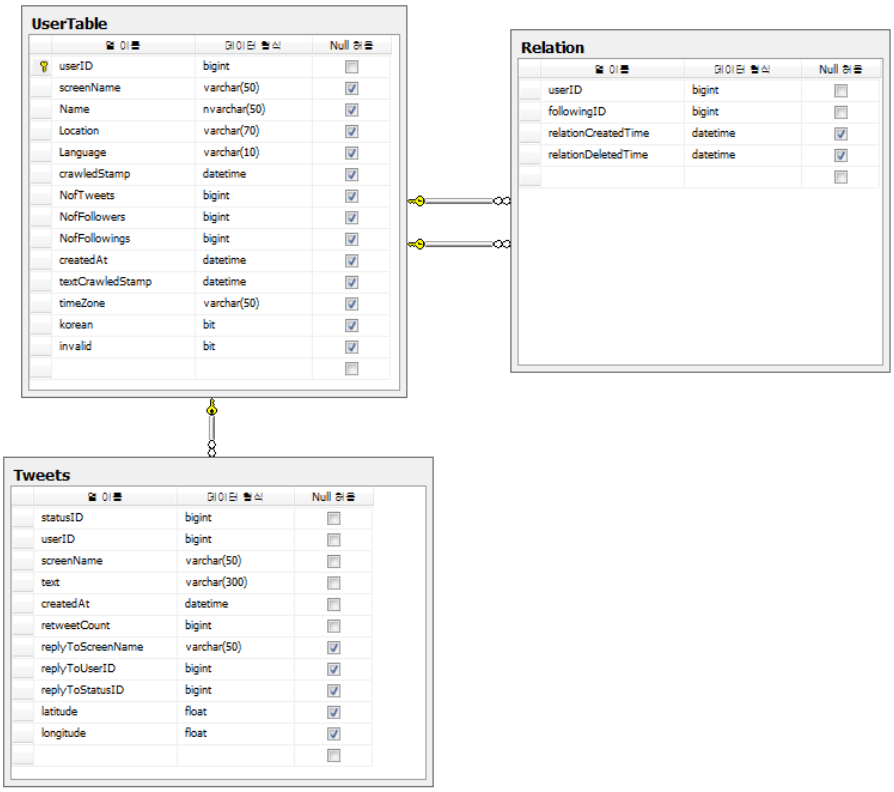
[그림 5] 사용자 정보 수집 방식

두 번째 단계에서는 첫 번째 단계에서 수집된 사용자들이 작성한 트윗을 수집한다. 수집한 트윗들은 [그림 6]과 같이 데이터베이스에 저장된다.



[그림 6] 수집된 트윗 저장 방식

수집된 사용자 정보와 트윗 정보들은 데이터베이스에 [그림 7]과 같이 3개의 테이블로 저장된다. ‘UserTable’ 테이블은 ID, 대화명, 지역, 언어, 작성한 트윗 수, 자신이 팔로잉한 사람 수, 자신의 팔로워(follower) 수 등을 변수로 갖는다. ‘Relation’ 테이블은 ID와 자신이 팔로잉하는 사람의 ID 등을 주요 변수로 갖는다. ‘Tweets’ 테이블의 경우 각 글의 고유한 값인 statusID, 작성한 사람을 나타내는 UserID, 내용을 나타내는 text 등을 주요 변수로 갖는다.



[그림 7] 트위터 데이터베이스 테이블

3.2. 데이터 전처리

가. 사용자 정의 사전 구축

트위터는 누구나 자유롭게 작성할 수 있는 온라인 마이크로 블로깅 서비스이다. 온라인에서 문서를 작성하는 정해진 규칙이 없기 때문에 사용자들이 트윗을 작성함에 있어 기존 사전에는 포함되지 않은 단어를 사용하는 경우가 많다. 온라인의 특성상 비속어, 신조어, 축약어 등 일반적인 사전에는 포함되지 않은 말들이 자주 사용된다. 특히 사용자들이 트윗을 작성할 때 140자라는 글자 수 제약이 있기 때문에 문장 구조나 단어 선택이 정상적인 문법의 범주를 벗어나는 경우가 많다. 또한 연예인, 정치인에 관한 이야기와 영화, TV 프로그램을 언급하는 경우가 많기 때문에 고유명사가 빈번히 사용된다. 그렇기 때문에 트윗들을 제대로 분석하려면 위 언급된 단어들을 포함하도록 추가적인 사전 구축 작업이 필요하다.

트윗들에서 토픽을 추출하기 위해서 자주 등장하는 신조어와 인물, 영화, 공연, 지역 이름, 예능 프로 이름 등과 같이 의미가 있는 고유명사를 보강해야 한다. 이러한 부분들을 보완할 수 있는 사전 구축을 위해 ‘네이버 실시간 검색어’와 ‘네이버 핫토픽 키워드’를 사용한다. ‘네이버 실시간 검색어’는 단위 시간 동안 네이버 검색창으로 입력되는 검색어 중 입력 횟수의 증가 비율이 가장 큰 검색어를 순서대로 보여주는 서비스이다. ‘네이버 핫토픽 키워드’는 뉴스, 카페, 블로그, 미투데이 등에서 생성된 문서를 기반으로 많이 언급되고 있는 키워드를 집계해서 보여주는

서비스이다. 이 두 가지를 이용하여 사전을 구축할 경우 온라인 공간에서 실제로 사람들이 사용하는 신조어와 고유명사를 선정할 수 있다. 사용자 정의 사전에 총 1083개 단어를 추가하였다. [표 1]은 사전에 추가될 단어들의 예시이다.

[표 1] 사용자 정의 사전에 추가될 단어의 예

각시탈	은교	멘탈붕괴	오메가 3	디카프리오
강예빈	인기가요	강기윤	빅토리아	오바마
곽진영	조여정	강석훈	호날두	개그콘서트
무한도전	지현우	문재인	박찬호	개콘
빅뱅	달샤벳	이정희	유지애	신보라
원더걸스	에프엑스	황영철	묵향	아이폰
유인나	김국진	강은희	레오나르도	애플

나. 명사 추출

텍스트 데이터를 분석에 사용하려면 각각의 문장를 형태소 단위로 분석하여 나온 품사들을 벡터 형태로 변환시켜야 한다. 현재 한국어 형태소 분석기 중 대표적인 것은 '한나눔 한국어 형태소 분석기[3]' 와 '꼬꼬마 형태소 분석기[4]' 가 있다. 본 연구에서는 '한나눔 한국어 형태소 분석기' 를 사용하여 분석을 수행하였다. [그림 8](시맨틱 웹 첨단연구센터, 2011)은 '한나눔 형태소 분석기' 의 전체적인 작업 흐름을 보여준다. 각 분석 단계 별

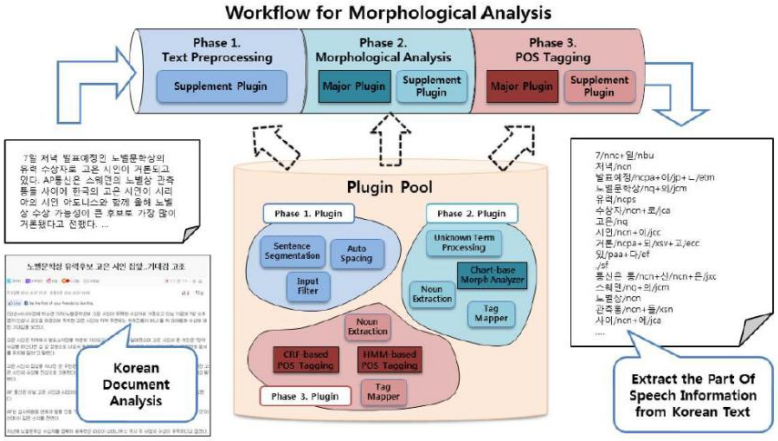
[3] 한나눔 한국어 형태소 분석기 (kldp.net/projects/hannanum)
[4] 꼬꼬마 형태소 분석기 (kkma.snu.ac.kr)

역할은 다음과 같다(시맨틱 웹 첨단연구센터, 2011).

Phase1. Text preprocessing: 문장, 경계 인식, 필터링, 자동 띄어쓰기 등 형태소 분석 이전에 필요한 전처리 작업을 수행

Phase2. Morphological Analysis: 입력 문장에 대해서 어절 단위로 발생 가능한 모든 형태소 분석 결과를 생성

Phase3. Pos Tagging: 가장 유망한 형태소 분석 결과들을 선택하여 입력 문장에 대한 최종 품사 태깅 결과를 반환



[그림 8] 한나눔 형태소 분석기 작업 흐름도

일반적으로 텍스트 데이터를 분석하는데 있어 선택적으로 품사를 사용한다. 본 연구에서는 형태소 분석 결과 나온 품사 중 명사만을 추출하여 분석에 사용한다. 한국어의 경우 문장에서 명사만을 뽑아 살펴보아도 전체적인 내용이 파악 가능한 경우가 많다. 명사 추출의 예를 살펴보면 다음과 같다.

예시 문장)“롯데마트가 판매하고 있는 흑마늘 양념 치킨이

논란이 되고 있다”

→”롯데마트”, “판매”, ”흑마늘”, “양념”, “치킨”, “논란”

위의 결과를 보면 추출된 명사들을 살펴보는 것만으로도 주요 내용을 파악하는 것이 가능함을 알 수 있다.

다. 데이터 클렌징(data cleansing)

트윗에서 명사들을 추출한 후 추가적으로 데이터 클렌징 과정을 수행한다. 트위터에 자주 사용되지만 실제 분석에는 필요 없는 형태의 단어들을 일정한 규칙을 이용하여 제거한다. 분석 목적에 따라서 데이터 클렌징 규칙은 가변적이다. 본 연구에서는 트위터에서 다른 사용자에게 메시지를 보내는 개념인 @ID 형태의 데이터와 사진, 동영상 혹은 특정 페이지를 연결 시켜주는 URL과 같은 경우 토픽 추출에 불필요한 노이즈(noise)라고 볼 수 있다. [표 2]는 이번 분석에 사용한 데이터 클렌징 규칙 목록을 나타낸다.

[표 2] 데이터 클렌징 규칙의 예

@ID 제거
URL 제거
숫자 제거
한자, 영어 제거
특수문자, 문장부호 제거
‘ㅋㅋㅋ’, ‘ㅏㅏㅏ’와 같은 자음과 모음 제거
사용자 정의 사전에 없는 한 글자 단어 제거
빈도수가 적은 단어 제거

텍스트 데이터 분석 시 분석 목적에 맞는 데이터 클렌징은 크게 두 가지 장점을 갖는다. 첫 번째로, 분석의 정확도를 향상시킨다. 분석에 불필요하다고 판단되는 내용을 제거 함으로써 노이즈가 줄게 되고 향후 군집화나 기타 방법론을 적용 시킬 때 결과 값의 정확도가 향상될 수 있다. 두 번째로, 분석의 효율성을 증가시킨다. 데이터 클렌징을 통해 불필요한 단어들이 제거되고 이는 텍스트를 벡터화함에 있어 차원을 축소시키는 결과를 가져온다. 고차원의 벡터로 표현되는 텍스트 데이터 차원 축소는 분석의 효율성 향상에 큰 도움을 줄 수 있다.

라. 벡터 스페이스 모델(vector space model)

데이터 클렌징을 마치고 텍스트 데이터를 벡터 스페이스 모델로 변환시키는 과정을 수행한다. 텍스트 데이터를 분석할 때 벡터 스페이스 모델과 같이 벡터 형태로 데이터를 변환하는 것은 매우 유용하다. 이때 각각의 벡터에 들어갈 성분 값으로 TF-IDF(term frequency – inverse document frequency)를 사용한다. TF-IDF는 하나의 문서 단위에서 단어들이 차지하는 중요도를 평가하는 일반적인 방법으로, 각 문서에서 출현 빈도가 높고, 전체 문서들에서 등장하는 빈도가 낮은 특징을 갖은 일수록 높은 값을 가진다. 각각의 트윗들에 대해 TF-IDF 벡터를 구하는 식은 다음과 같다.

$$V_d = (tf(t_1) \cdot idf(t_1), \ldots, tf(t_i) \cdot idf(t_i), \ldots, tf(t_M) \cdot idf(t_M))$$

$tf(t_i)$: 트윗 d에서 단어 i의 term frequency(TF) $df(t_1)$: 단어 i를 가지고 있는 트윗의 수 $idf(t_i) = \log(\frac{n}{df(t_i)})$: 단어 i의 inverse frequency(IDF) n: 전체 트윗 수 M: 전체 단어 수

단어 별로 각 트윗 내에서의 출현 빈도를 계산하고 전체 트윗을 대상으로 그 단어가 포함된 트윗의 수를 계산한다. 이 값들이 각각 tf와 df값이 된다. 각 문서에서는 빈도가 높고, 전체 문서들에게서는 빈도가 낮은 특징을 찾기 위한 TF-IDF의 목적에 맞게 tf값과 idf값 (전체 트윗 수를 df로 나눈 값에 로그를 취한 값)을 곱한다. 이렇게 계산된 TF-IDF 값이 큰 단어의 경우 해당 트윗에서 중요한 단어라고 할 수 있다.

위의 식을 이용하여 여러 트윗들을 벡터화 한 뒤, 각 문서를 하나의 열로 모은 매트릭스 형태로 변환하여 벡터 스페이스 모델로 표현한다. [그림 9]는 트윗 별 단어 빈도 수로 정리된 매트릭스를 TF-IDF값으로 이루어진 벡터 스페이스 모델로 변환하는 예시이다.

	Term1	Term2	Term3	Term4	...	TermN
Document1	1	1	1	0		0
Document2	0	0	1	1		0
Document3	0	0	0	1		1
Document4	0	1	0	1		1

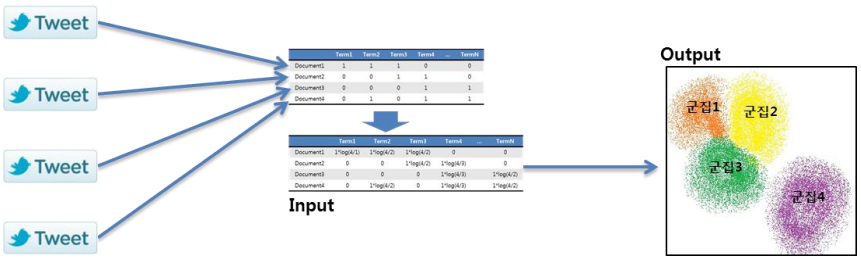
	Term1	Term2	Term3	Term4	...	TermN
Document1	1*log(4/1)	1*log(4/2)	1*log(4/2)	0		0
Document2	0	0	1*log(4/2)	1*log(4/3)		0
Document3	0	0	0	1*log(4/3)		1*log(4/2)
Document4	0	1*log(4/2)	0	1*log(4/3)		1*log(4/2)

[그림 9] TF-IDF를 이용한 벡터 스페이스 모델 예시

위의 예시는 총 4개의 문서(document)에 총 N개의 단어가 있는 경우이다. 문서 1의 단어 2에 해당하는 TF-IDF 계산을 살펴보자. 변환 전의 매트릭스를 통해 문서 1에서 발생하는 단어 2의 빈도는 1이 되는 것을 알 수 있다. 이 값은 문서 1의 단어 2에 tf값을 가리킨다. 단어 2의 경우 전체 4개의 문장 중 총 2개의 문장에서 등장한다. 따라서 단어 2의 df값은 2가 된다. 이렇게 구한 df값과 전체 문장 수를 이용하여 idf값을 구하면 log(4/2)가 된다. 앞서 구한 tf값과 idf값을 곱하면 문서 1의 단어 2에 해당하는 TF-IDF값이 나온다.

3.3. 트윗 군집화

동일한 주제의 트윗들을 하나의 묶음으로 분류하기 위해 군집화를 수행한다. 군집화가 이상적으로 이루어지면 각 군집이 하나의 토픽을 나타내게 된다. 군집화 제대로 되야 다음 단계에서 군집 별로 하나의 토픽을 추출할 수 있기 때문에 이 과정은 매우 중요하다. [그림 10]은 군집화 과정의 예시를 간략하게 보여준다. 우선 수집한 트윗들을 벡터화하고, 각 벡터 별로 계산된 TF-IDF 값을 입력 값으로 군집화를 수행한다. 그 결과 해당 트윗이 속한 군집 번호를 출력 값으로 얻는다.



[그림 10] 군집화 과정 예시

군집화는 학습 데이터나 레이블이 필요 없는 비교사 학습(unsupervised learning) 방법으로, 문서를 유사한 주제끼리 분류하는데 매우 효과적이다(Treeratpituk & Callan, 2006). 본 연구에서는 트윗 군집화를 위해 k-평균 군집화와 계층적 군집화 방법 두 가지를 사용한다.

가. k-평균 군집화 (Han et al., 2006)

k-평균 군집화는 분할 군집화(partitioning clustering) 기법 중 대표적으로 많이 사용되는 방법이다. 미리 군집의 수, k를 정하고 중심값(centroid)을 기준으로 k개의 군집으로 객체들을 나눈다. 기법이 구현되는 방식은 다음 [표 3]와 같다.

[표 3] k-평균 군집화 알고리즘

1. 주어진 k에 대해서, 임의의 k개의 객체를 중심값으로 군집 설정
2. 나머지 객체와 중심값들 사이의 거리를 계산
3. 거리 계산 결과 가장 가까운 군집에 객체들을 할당
4. 모든 객체들이 각 군집에 할당 된 후, 다시 군집 별 중심값 계산
5. 더 이상 변화가 없을 때까지 2~ 4단계 반복

k-평균 군집화는 계산이 복잡하지 않고 쉽게 사용이 가능하다는 장점이 있다. 하지만 예외 상황이나 노이즈가 많은 텍스트에는 적용이 어려울 수 있다.

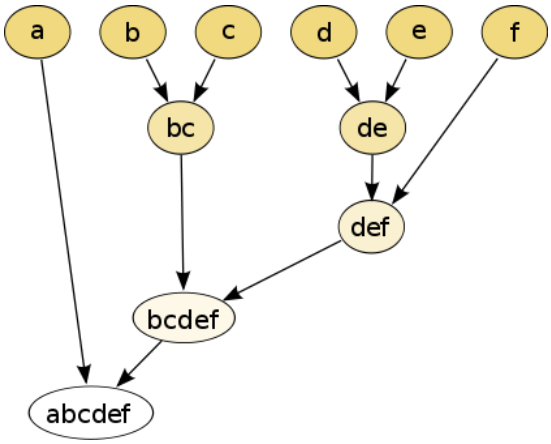
나. 계층적 군집화 (Han et al., 2006)

계층적 군집화(hierarchical clustering)는 계층적으로 군집을 생성하는 기법으로 분할 군집화 기법과 비교하여 상대적으로 높은 성능과 낮은 효율성을 가진다. 계층적 군집화 방법은 크게 응집 방식의 계층 군집화(agglomerative hierarchical clustering)와 분할 방식의 계층 군집화(divisive hierarchical clustering)로

나뉜다. 분할 방식의 계층 군집화는 실제 분석에서 상대적으로 드물게 사용된다(Kaufman et al. 2009). 본 연구에서는 응집 방식의 계층 군집화를 분석에 사용한다. 기법이 구현되는 방식은 다음 [표 4]와 같다.

[표 4] 계층적 군집화 알고리즘

1. 각 객체를 하나의 독립적인 군집으로 두고 시작
2. 모든 군집 간의 상호 거리 계산
3. 가장 거리가 짧은 두 군집을 하나의 군집으로 병합
4. 모든 군집이 하나의 군집으로 병합될 때까지 2, 3단계 반복



[그림 11] 계층적 군집화 예시

[그림 11]은 응집 방식의 계층적 군집화의 예시이다. {a}, {b}, {c}, {d}, {e}, {f} 6개의 각기 다른 객체가 있다고 가정하자. 처음에는 하나의 객체가 각기 다른 군집을 이루고 있다. 각 군집별로 거리를 계산한 결과 {b}와 {c} 그리고 {d}와 {e} 사이의

거리가 가장 가깝기 때문에 이들을 병합한다. 그 결과 {a}, {b, c}, {d, e}, {f}와 같이 4개의 군집이 생성된다. 위와 같은 방식으로 남은 4개의 군집 간의 거리를 계산하고 가장 가까운 군집을 병합하면 {d, e}와 {f}가 하나의 군집을 이뤄 {d, e, f}가 된다. 이러한 방식으로 전체 객체가 하나의 군집인 {a, b, c, d, e, f}를 이룰 때까지 반복적으로 병합을 수행한다.

3.4. 트윗 순위 선정

군집화가 끝난 트윗들은 각 군집에 주제 별로 분류가 된다. 각각의 군집들이 하나의 토픽을 갖도록 잘 나뉘었다면 마지막으로 토픽이 어떠한 내용인지를 정확히 파악하는 과정이 필요하다. 이는 각 군집 별로 트윗의 순위를 계산하여 나오는 상위 순위 트윗들을 통해 파악할 수 있다. 트윗들의 순위를 계산하는 방법은 다음 [표 5]와 같다.

[표 5] 트윗 순위 선정 알고리즘

1. 전체 객체(트윗)의 평균을 계산하여 각 군집의 중심값을 구함
2. 각 군집의 중심값과 군집 내 객체(트윗)의 코사인 유사도 계산
3. 유사도가 가장 큰 값부터 순서대로 객체(트윗)들의 순위를 선정

코사인 유사도(cosine similarity)를 계산하는 방법은 다음과 같다.

$$sim\left(d_i, d_j\right) = \cos(\theta) = \frac{V_{d_i} \cdot V_{d_j}}{\left\|V_{d_i}\right\| \left\|V_{d_j}\right\|}$$

d_k : 문서 k
V_{d_k} : 문서 k의 TF-IDF 벡터

문서 i와 문서 j의 코사인 유사도는 각 문서들의 TF-IDF 벡터의 내적 값을 각 벡터의 스칼라 값의 곱으로 나누어 구할 수

있다. 이 값은 문서 i의 벡터와 문서 j의 벡터가 이루는 각도 (θ)에 코사인(cosine)을 취한 값과 같다. 코사인 유사도 값은 두 문서 벡터가 얼마나 유사한지를 판단하는 척도이다.

실제로 군집화에 사용하는 입력 값은 트윗 자체가 아닌 트윗에서 명사를 추출하여 만든 벡터들이다. 이 벡터들의 평균 값을 구한다면 토픽을 설명하는 명사들의 경우 대다수의 트윗에 존재하므로 값이 어느 정도 유지가 되겠지만 그렇지 않은 명사들은 평균을 구하면서 거의 0에 가까운 결과가 나올 것이다. 고로 중심 값(centroid)은 핵심 명사들을 제외한 나머지 명사들의 값이 0에 가까운 벡터가 된다. 이러한 결과는 결국 핵심 단어를 많이 포함한 트윗에게 높은 유사도 값을 주어 상위 순위가 나오는 결과를 가져오게 한다.

이와 같은 순위 결정 방법을 통해 사용자들은 잡담이나 스팸성 트윗이 아닌 진짜 정보를 담고 있는 트윗들을 빠르고 정확하게 얻을 수 있다.

제 4 장 실험 설정 및 결과

4.1. 데이터 설명

본 연구에서 사용한 데이터는 대표적인 마이크로 블로깅 서비스인 트위터이다. 자체적으로 구축한 트윗 데이터베이스에서 2013년 1월 1일에 작성된 트윗 중 사전에 정의한 5개의 주제와 관련된 트윗을 무작위로 200개씩 총 1,000개의 트윗을 분석에 사용하였다.

[표 6] 분석에 사용한 트윗 데이터 주제

번호	주제
1	김태희-비 열애
2	싸이 타임스퀘어 공연
3	손현주 SBS 연기대상 수상
4	택시법 통과
5	지동원 임대

각각의 주제는 [표 6]과 같다. 특정 주제와 관련된 트윗들을 분석에 사용한 이유는 향 후 분석에 있어 비교사 학습 방법인 군집화 기법의 순수도(purity) 측정과 결과 해석에 있어서의 타당성을 확인하기 위해서이다.

[그림 12]은 분석에 사용한 트윗들 중 일부 예시를 나타낸다. 예시 트윗들을 살펴보면 '@ID', URL 사용, 'RT'와 같은 트위터에서 자주 사용되는 특별한 특성이 나타나는 것을 알 수 있다. 이 같은 특성들을 3.2 절에서 설명한 전처리 방법을 통해 실제 분석에서는 제거되었다.

Tweet
우와...비와 김태희라니.....디스패치가 터트린다는게이거였구나...
디스패치가 한건 제데로 했네,비랑 김태희라니.상상도 못함ㅋㅋ
디스패치 새해부터 ㅋㅋㅋㅋㅋㅋ 비랑 김태희?ㅋㅋㅋㅋㅋ
새해맞이 열애설이... 모든 이의 예상을 깨고 김태희-비(정지훈)이라니... 디스패치 홈페이지가 안 열린다 ㄷㄷㄷ
와 디스패치 ㄷㄷㄷㄷㄷㄷㄷㄷㄷㄷㄷㄷㄷ김태희였다니 .. - w : http://t.co/OWWkeHkc - m : http://t.co/KjpbDBWJ
쩐다 [택시법 이준기 햄버거] 나의 운명은 어디에?... http://t.co/i5KiC9hQ
정부·버스업계 " 택시법 통과 유감 " http://t.co/xoaKiKFB(국민투표 찬반합시다. 국회의원이 아닌국민에게 ... ?
냠.... 2013년 [떡국맛있게끓이는법 택시법] 나의 운명은 누구?... http://t.co/HkkvRsBD
냥.. 확밀아 RT @wan1210:[공무원봉급표2013 택시법] 내가 평생 모을 수 있는 돈은?... http://t.co/xPXK0POs
http://t.co/10kEOs5Y 택시법 국회 본회의 통과 ㄱ. 진짜 버스 파업 대비해야 하나;;
유재석 노홍철도 싸이와 헤머와 함께 타임스퀘어 생방 공연..ㅎㄷㄷ
우옷... 한밀아 [유영진 싸이 타임스퀘어] 새해 내게 일어날 일 일아보기!!!!... http://t.co/AN6K55Cy
싸이, 유재석·노홍철과 美 타임스퀘어서 말춤··폭소 http://t.co/m3QC2tf3 무한도전! \(^0^*)/
'ABC' 미국 타임스퀘어 2013 년 new year 프로그램에 싸이 방금 공연 했음. 근데 유느님, 노홍철, 하하 찬조 출연 했음. 유느님 춤 끝내줬음.
2012년 12월 31일, 싸이 타임스퀘어 공연 http://t.co/C1ZHAg3X
역시 손현주 형님이 탔다!! 대상!!!!
아쌕ㅋㅋㅋ손현주 대상ㅋㅋㅋㅋ요싴ㅋㅋㅋㅋ
손현주 대상 축하! 인정!
손현주 대상~!!!!!
역시 추적자 아저씨가 대상이군!!!!!손현주 짱b
지동원 아욱국 임대~!!!!!!드디어 갔구나!!! 큐피알 아님 아욱국 갔음 좋겠다 했는데 이번시즌 지쿠 동반상승 기대!!!
RT @naver_sports: [해외축구] 지동원, 아우크스부르크 임대 확정...구자철과 한솥밥 http://t.co/Tf7np6Kg
RT @cdk0649: 선덜랜드 감독, "지동원 임대 보낼 것" http://t.co/RQOAsAcf
RT @cdk0649: 선덜랜드 감독, "지동원 임대 보낼 것" http://t.co/RQOAsAcf
선덜랜드, 지동원 임대 이적 검토 http://t.co/md1AFiI0

[그림 12] 분석에 사용한 트윗 예시

트윗 데이터를 전처리 한 결과 (987 * 533) 크기의 매트릭스 형태로 TF-IDF 벡터 스페이스 모델이 구축되었다. 전처리 과정에서 총 1,000개의 트윗 중 13개의 트윗이 제거 되었고, 남은 987개의 트윗에서 총 533개의 명사가 추출되었다. 분석에 사용된 트윗과 명사 수는 [표 7]와 같다.

[표 7] 분석에 사용된 트윗과 명사 수

분석에 사용된 총 트윗 수	분석에 사용된 총 명사 수
987	533

4.2. 실험 설정

가. 군집화 시 객체 간 거리 계산

수집된 1,000개의 트윗 중 전처리 과정을 거치고 남은 987개의 트윗을 사용하여 군집화를 수행한다. k-평균 군집화와 계층적 군집화 두 가지 방법을 사용한다.

k-평균 군집화와 계층적 군집화는 각 객체 간의 거리를 계산하여 군집화를 수행한다. 본 연구에서는 텍스트 분석에서 많이 사용되는 코사인 유사도를 이용하여 거리를 계산한다. 문서 간 코사인 유사도의 계산 방법은 3.4절에서 서술한 내용과 동일하다.

유사도의 값을 직접 거리로 사용할 수 없기 때문에 다음 식과 같이, 유사도를 이용한 간단한 계산을 통해 거리를 구할 수 있다.

$$distance = 1 - similarity$$

위와 같이 간단한 계산식을 통해 거리를 구할 수 있는 이유는 코사인 유사도 값이 [0, 1]의 범위에 포함되기 때문이다. 1에서 유사도를 뺀 값이 클수록 두 문장이 덜 비슷하고, 작을수록 더 비슷하다. 이는 우리가 생각하는 거리와 같은 의미를 갖는다.

나. 군집 수 결정 방법

k-평균 군집화와 계층적 군집화를 수행할 때 최적의 군집 수를 정하기 위해 군집을 평가하는 지표인 Silhouette 값 (Han et al., 2006)을 이용한다. Silhouette 값의 수식은 다음과 같다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

$a(i)$: 군집 내에서 객체 i와 모든 다른 개체들의 평균 거리
$b(i)$: 객체 i가 속하지 않는 군집의 모든 개체와 i의 평균 거리

Silhouette 값은 군집 내 각 객체들 사이의 거리가 작을수록, 각 군집 사이의 거리가 클수록 큰 값을 갖는다.

Silhouette 값은 [-1, 1]의 범위의 값을 갖고, 이 값이 클수록 군집 내 각 객체들 사의 거리가 작고 각 군집 사이의 거리가 크다는 것을 의미한다. 달리 말하면 Silhouette 값이 큰 군집은 잘 나뉜 군집이라고 할 수 있다.

최적의 군집 수를 찾기 위해 k-평균 군집화와 계층적 군집화에 각각 실험을 수행한다. 실험 방법은 군집 수를 2부터 9까지 늘려가면서 모든 군집의 silhouette 값의 평균이 가장 큰 군집 수를 찾아 그 때의 군집 수를 최적 값으로 정한다.

다. 군집화 결과 평가 방법

군집화의 결과를 평가하기 위해 군집 별로 순수도를 구한다. 순수도의 계산 방법은 다음과 같다.

$$Purity = \frac{\text{군집 내 가장 많이 나온 주제의 트윗 개수}}{\text{군집 내 전체 트윗 개수}}$$

군집 내에서 다수인 주제의 트윗 개수를 군집 내 전체 트윗 개수로 나누면 그 군집이 얼마나 순수한지를 평가할 수 있다. 군집화는 실제 트윗이 속한 주제에 대한 정보를 사용하지 않는 비교사 학습이므로 이와 같은 평가 방법을 적용할 수 있다.

라. 토픽 추출 결과 평가 방법

3.4절에서 설명한 문서 순위 계산 방법을 통해 순위를 산출하고 이를 바탕으로 토픽을 파악한다. 정성적으로 결과를 평가하는 위해 문서 순위 상의 상위 트윗들과 하위 순위 트윗들을 비교하여 상위 순위 트윗들이 토픽을 잘 반영하고 있는지를 살펴본다.

4.3. 실험 결과

가. k-평균 군집화 결과

초기 군집 수를 결정하기 위해 군집 수를 2에서 9까지 조정하면서 silhouette 값을 비교해본 결과 군집 수가 6일 때가 최적인 것으로 나왔다. [표 8]는 군집 수에 따른 평균 silhouette 값을 나타낸 표이다.

[표 8] 군집 수에 따른 평균 silhouette 값 (k-평균 군집화)

K	2	3	4	5	6	7	8	9
Avg. silhouette	0.1170	0.1606	0.2292	0.2645	**0.2822**	0.2615	0.2551	0.2383

미리 정한 5개의 주제에서 뽑은 트윗 데이터를 이용하여 분석을 수행하였지만 silhouette 결과는 군집 수가 6일 때가 최적으로 나왔다.

군집 수를 6으로 하였을 때 k-평균 군집화 결과는 [표 9]와 같다.

[표 9] k-평균 군집화 순수도 결과

Cluster No.	Cluster Name	데이터 수	순수도
1	김태희-비 열애	204	0.9706
2	싸이 타임스퀘어 공연	198	1
3	손현주 SBS 연기대상 수상	195	1
4	택시법 통과	194	1
5	지동원 임대	57	1
6	지동원 임대 2	139	1
Total		**987**	**0.9939**

k-평균 군집화 결과를 바탕으로 순수도를 계산한 결과 전체 순수도는 0.9939라는 높은 수치가 나왔다.

각 군집 별로 살펴본 결과 '지동원 임대'와 관련된 트윗들이 2개의 군집으로 나뉘어 총 6개의 군집이 나왔다. 두 군집에 대해 내용을 살펴보면 실제로 같은 이야기를 다루지만 5번 군집의 경우 '소문'을 이야기하고 있고, 6번 군집은 '사실'을 이야기 하고 있다. 이와 관련된 자세한 내용과 논의 사항은 뒤에 올 (라. 결과 논의) 부분에 수록하였다.

나. 계층적 군집화 결과

k-평균 군집화와 마찬가지로 초기 군집 수를 결정하기 위해 군집 수를 2에서 9까지 조정하면서 silhouette 값을 비교해본 결과 군집 수가 5일 때가 최적인 것으로 나왔다. [표 10]는 군집 수에 따른 평균 silhouette 값을 나타낸 표이다.

[표 10] 군집 수에 따른 평균 silhouette 값 (계층적 군집화)

K	2	3	4	5	6	7	8	9
Avg. silhouette	0.0889	0.1379	0.2007	**0.2629**	0.2592	0.2595	0.2479	0.2484

미리 정한 5개의 주제에서 뽑은 트윗 데이터를 이용하여 분석을 수행하였는데 silhouette 결과가 이와 동일하게 5개를 최적으로 나왔다.

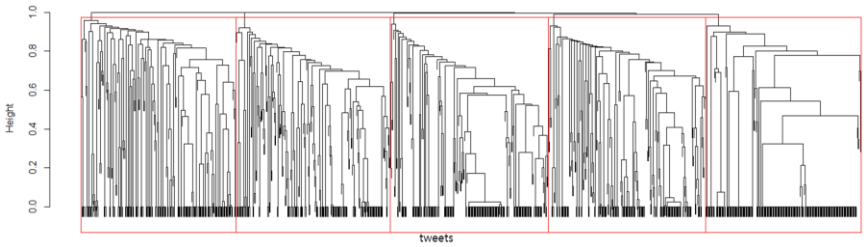
군집 수를 5으로 하였을 때 계층적 군집화 결과는 [표 11]과 같다.

[표 11] 계층적 군집화 순수도 결과

Cluster No.	Cluster Name	데이터 수	순수도
1	김태희-비 열애	199	0.9849
2	싸이 타임스퀘어공연	200	0.99
3	손현주 SBS연기대상 수상	195	1
4	택시법 통과	196	1
5	지동원 임대	197	0.9949
Total		**987**	**0.9939**

계층적 군집화 결과를 바탕으로 순수도(purity)를 계산한 결과 전체 순수도는 0.9939로 k-평균 군집화와 마찬가지로 높은 수치가 나왔다.

계층적 군집화를 시각적으로 보여주는 덴드로그램 (dendro-gram)의 결과를 살펴보면 5개의 군집으로 나뉘는 것이 어느 정도 타당성을 가진다는 것을 알 수 있다. 계층적 군집화의 덴드로그램은 [그림 13]와 같다.



[그림 13] 계층적 군집화 덴드로그램

다. 트윗 순위 선정 및 토픽 추출 결과

(가.)에서 수행한 k-평균 군집화 결과대로 나뉜 군집 별로 트윗들의 순위를 정한다. 순위는 각 군집의 중심값과 객체 간의 코사인 유사도를 통해 계산된다. 각 군집 별로 상위 순위 트윗들과 하위 순위 트윗들을 비교하고, 상위 트윗들에서 토픽을 추출해 내는 과정을 거친다.

ㄱ. 군집 1 결과

군집 1은 '김태희-비 열애' 관련 주제로 이루어진 군집이다. 군집에 포함된 객체 수는 204개이고 순수도는 0.9706이 나왔다. 군집 1에서 트윗 순위를 정한 뒤 상위 순위 트윗들을 살펴보면 다음 [표 12]와 같다.

[표 12] 군집 1의 상위 순위 트윗 결과

Similarity	Tweet
0.7457	헐ㅋㅋㅋ

상위 순위부터 5개의 트윗을 살펴보면 ‘디스패치’, ‘김태희’, ‘비’, ‘열애’, ‘데이트’와 같은 주요 내용들이 포함된 것을 알 수 있다. 이를 토대로 군집 1의 토픽은 ‘디스패치에서 김태희와 비의 열애설 발표’라는 걸 파악할 수 있다.

[표 13] 군집 1의 하위 순위 트윗 결과

Similarity	Tweet
0.0285	[뜨거운감자]택시법 http://t.co/2laqfnoo
0.0619	[사설] 세금 2 조원 ‘택시법’은 탈법적 포퓰리즘이다 – http://t.co/orKEubqu
0.1488	@BDVIOLET 디스패치 맞긔!! 진짜 태희언니 그얼굴로 그런옷 입는다는게 쇼킄ㅋㅋㅋㅋㅋㅋ 너도 작년에 너가 먹은 술과 아름다운 양식많큼 복 많이 받아!!!! 올해도 잘 부탁해!!ㅋㅋㅋㅋ
0.1524	태희언니....♥ .....잘가여....새해첫날부터 흡... 디스패치대박이 다진짝ㅋㅋㅋㅋㅋㅋㅋㅋ 엄마는 비팬일뿐이고...엄마는 지금 멘붕일뿐이고 ㅋㅋㅋㅋㅋㅋㅋ 엄마그렇다고 아침도 안하면 어떠케...
0.1690	엥? 모두들 [디스패치 김태희 CF 계약서] 나의 솔로부대 계급 알아보기... http://t.co/9aBJa7oo

[표 12]의 상위 순위 트윗 결과와는 다르게 [표 13]의 하위 순위 트윗 결과는 의미를 파악하는 것이 어렵다. 하위 순위 트윗들의 경우 군집 1의 내용과 관계 없는 ‘택시법’과 관련된 내용이 나오거나 광고 링크를 포함한 스팸인 경우가 있다. 김태희와 비를 언급한 트윗의 경우에는 개인적인 감상과 같은 잡담이 많이 포함되어 정확한 내용 파악이 어렵다.

ㄴ. 군집 2 결과

군집 2는 ‘싸이 타임스퀘어 공연’관련 주제로 이루어진 군집이다. 군집에 포함된 객체 수는 198개이고 순수도는 1이 나왔다. 군집 2에서 트윗 순위를 정한 뒤 상위 순위 트윗들을 살펴보면 다음 [표 14]와 같다.

[표 14] 군집 2의 상위 순위 트윗 결과

Similarity	Tweet
0.8513	싸이 타임스퀘어 공연 저거 뭐야.. 노홍철 유재석도 나오고 말이지.. 영등포 타임스퀘어는 아닌거 같고! 진짜 출세했구나
0.8449	유재석 노홍철도 싸이와 헤머와 함께 타임스퀘어 생방 공연.. ㅎㄷㄷ
0.7892	싸이 11 억뷰,노홍철 유재석 타임스퀘어에서 싸이와 함께~ http://t.co/ehhT9c8m
0.7781	RT @batangso: 싸이 흥하니 같이 흥하는구려. RT @geniusoh: 우왕!!!!!!!"@progfrog: 싸이, 유재석, 노홍철, MC 해머 at 타임스퀘어~~ 유재석, 노홍철 언제 간거지? 대단하다~~ http://t.co/SROv ...
0.7746	@wnawna00 싸이,유재석,노홍철 in 뉴욕타임스퀘어 공연 12.12.31 http://t.co/0ryJoKqO @youtube 에서

상위 순위부터 5개의 트윗을 살펴보면 ‘싸이’, ‘타임스퀘어’, ‘공연’, ‘유재석’, ‘노홍철’, ‘MC 해머’와 같은 주요 내용들이 포함된 것을 알 수 있다. 이를 토대로 군집 1의 토픽은 ‘싸이, 노홍철, 유재석, MC해머가 함께 뉴욕 타임스퀘어에서 연말 공연’라는 걸 파악할 수 있다.

[표 15] 군집 2의 하위 순위 트윗 결과

Similarity	Tweet
0.1588	헐랭.. 가요대전 RT @korean_pop_news:[비 싸이 타임스퀘어] 페북 방문자 추적기 사실?... http://t.co/FUpNtGaH
0.1829	냠... MBC [노준규 김기현 싸이 타임스퀘어] 나의 페북 가치는 얼마일까?... http://t.co/ETFqQIgj
0.1888	ㅜㅜ 해피뉴이어 RT @wanitools:[싸이 싸이 타임스퀘어] 페북 방문자 추적기 진짜?... http://t.co/Ovdp6gXc
0.1900	싸이 타임스퀘어 직캠 동영상 http://t.co/wiUDDbIm
0.1917	움... 마지막 [비 싸이 타임스퀘어] 나는 솔로다... http://t.co/tjbmBoQ5

[표 14]의 상위 순위 트윗 결과와는 다르게 [표 15]의 하위 순위 트윗 결과는 의미를 파악하는 것이 어렵다. 하위 순위 트윗들의 경우 '싸이', '타임스퀘어'라는 단어는 포함하지만 광고 링크를 포함한 스팸임을 알 수 있다. 사람들이 광고를 많이 보도록 유도하기 위해 화제가 되고 있는 단어들을 포함한 트윗들이 하위 순위에는 나타났다.

ㄷ. 군집 3 결과

군집 3은 '손현주 SBS연기대상 수상' 관련 주제로 이루어진 군집이다. 군집에 포함된 객체 수는 198개이고 순수도는 1이 나왔다. 군집 3에서 트윗 순위를 정한 뒤 상위 순위 트윗들을 살펴보면 다음 [표 16]와 같다.

[표 16] 군집 3의 상위 순위 트윗 결과

Similarity	Tweet
0.8319	손현주 받을만한데요?RT @gift_all: sbs 연기대상 손현주 대상 말세네요. 채시라나 장동건이 타야 할것을 쯧쯧, 염정아도 얼마나 잘했는데~
0.8132	[대상 - 손현주] 2012 SBS 연기대상 http://t.co/rH9DCJ8W @youtube 추카합니다
0.7996	sbs 연기대상 손현주. 정말 제대로 된 시상이다!!!!!!!!!!!! 캬하.
0.7461	손현주 연기대상 받을만해. 추적자 넘 재밌었지
0.7354	뚜아 손현주씨 대상타셨음?? 추카추카드립니다!!

상위 순위부터 5개의 트윗을 살펴보면 '손현주', 'SBS 연기대상', '대상', '시상'과 같은 주요 내용들이 포함된 것을 알 수 있다. 이를 토대로 군집 3의 토픽은 '손현주 SBS 연기대상 대상 수상'이라는 걸 파악할 수 있다.

[표 17] 군집 3의 하위 순위 트윗 결과

Similarity	Tweet
0.2081	10 년전 우리집에서 단막극 촬영을 할때 소박하시던.. 불과 3 년전 소주한잔 하시면서 태현아 나중에 찐한거 함하자..늘 꿈을 꾸시던 정말 몇달전 촬영중 피곤한 상태에서 반갑게 인사해주시던... SBS 연기대상 손현주 형님 축하드립니다 진심으로!!
0.2143	손현주씨 대상은 충분히 공감이 간다 그 배우 장미빛인생부터 알아봤다 장미빛인생 같은 슬픈드라마는 별루 안좋아하는데 아직도 기억에 남는다
0.2228	RT @minokkim: 2013 년은 밝았지만. 손현주가 대상받아 너무 기쁘다~한때 내 이상형이었는데ㅋ이제야 제대로 평가받아 넘 좋넹~!!ㅎ

[표 16]의 상위 순위 트윗 결과와는 다르게 [표 17]의 하위 순위 트윗 결과는 의미를 파악하는 것이 어렵다. 하위 순위 트윗들의 경우 손현주의 대상 이야기를 포함하지만 개인적인

감상이 내용의 주를 이루고 있어 정확한 내용 파악이 어렵다.

ㄹ. 군집 4 결과

군집 4는 ‘택시법 통과’관련 주제로 이루어진 군집이다. 군집에 포함된 객체 수는 194개이고 순수도는 1이 나왔다. 군집 4에서 트윗 순위를 정한 뒤 상위 순위 트윗들을 살펴보면 다음 [표 18]와 같다.

[표 18] 군집 4의 상위 순위 트윗 결과

Similarity	Tweet
0.7190	@KevinGu10 우리의 누구당 같다... 어쩜 하나같이 똑같냐... 미국이라고 다를 게 없어... 우리는 말도 안되는 법이 하나 유야무야 통과됨. 이른바 택시법이라고. 택시가 대중교통이란다...
0.7079	@mofun_arcade @Umbrella_KR 우리에게는 통과된 택시법이 있어서 대중교통 택시!! 어?
0.6887	와... 택시법 통과됬네.. 이건아니잖아 =_=
0.6692	헝 택시가 대중교통으로 인정받는 택시법이 통과됐군요.헐...지금 택시 돌아가는 꼴을 보면 절대 그렇게 될 수 없는데..
0.6470	정부, '택시법' 통과 "허탈하고 안타깝다": 국토해양부는 1 일 국회 본회의에서 택시를 대중교통 수단으로 인정하는 이른바 '택시법'(대중교통 육성 및 이용 촉진법 개정안)이...기사보기

상위 순위부터 5개의 트윗을 살펴보면 ‘택시법’, ‘통과’, ‘대상’, ‘대중교통’과 같은 주요 내용들이 포함된 것을 알 수 있다. 이를 토대로 군집 4의 토픽은 ‘택시법 통과, 택시 대중교통 인정’이라는 걸 파악할 수 있다.

[표 19] 군집 4의 하위 순위 트윗 결과

Similarity	Tweet
0.1451	RT @koreaspiritnana: 중국은 계속 연 10%씩 국방비를 늘리고 아베 역시 일본의 국방비를 크게 늘린다는데 한국은 택시법 포퓰리즘 같은 복지를 위해 국방비와 r&d 예산이 대폭 삭감됬다 어려울수록 허리띠를 졸라 매야하는데 ...
0.1698	@hopesumi 택시법에 노동자 처우개선은 명시 됬나요..?아니면 역시나 사업주 배만 채워주는 건가요.?
0.1719	RT @ksh1429: 복지예산 확대, 국채발행 백지화(!), 강정기지 예산통과, 부자증세, 택시법-마트규제법. 여야가 통과시킨 예산안의 주요내용입니다. 안보를 내세우던 새누리가 '국방' 예산을 줄이는데 합의했다는게 의외군요. http://t. ...

[표 18]의 상위 순위 트윗 결과와는 다르게 [표 19]의 하위 순위 트윗 결과는 의미를 파악하는 것이 어렵다. 하위 순위 트윗들의 경우 '택시법' 이라는 단어는 포함지만 토픽과 관련 없는 정치적인 견해가 내용의 주를 이루고 있어 정확한 내용 파악이 어렵다.

ㅁ. 군집 5 결과

군집 5는 '지동원 임대' 관련 주제로 이루어진 군집이다. 군집에 포함된 객체 수는 57개이고 순수도는 1이 나왔다. 군집 5에서 트윗 순위를 정한 뒤 상위 순위 트윗들을 살펴보면 다음 [표 20]과 같다.

[표 20] 군집 5의 상위 순위 트윗 결과

Similarity	Tweet
0.7985	유럽 축구 소식에 정통한 관계자는 1 일(한국시간) "선덜랜드가 지동원의 독일 분데스리가 아우크스부르크로 임대를 허락했다. 지난달 30 일(한국시간) 마틴 오닐 감독이 지동원을 직접 만나 이야기를 나눈 뒤 임대 쪽으로 가닥을 잡았다"고 전했다
0.5811	RT @Kampfflieger: 지동원의 아우크스부르크 이적 임대 소식이 있었나본데 아직 독일에서 가장 최신 소식은 아마도 키커지(오프라인 잡지 12 월 31 자)에 언급된 로이터가 전임 단장때 영입대상이었던 선수들과의 협상을 계속 진행한다는게 아 ...

상위 순위부터 2개의 트윗을 살펴보면 ‘선덜랜드’, ‘독일’, ‘분데스리가’, ‘지동원’, ‘아우크스부르크’, ‘임대’와 같은 주요 내용들이 포함된 것을 알 수 있다. 이를 토대로 군집 5의 토픽은 ‘선덜랜드의 지동원이 독일 분데스리가 아우크스부르크 임대 소식’이라는 걸 파악할 수 있다.

[표 21] 군집 5의 하위 순위 트윗 결과

Similarity	Tweet
0.2751	RT @Kampfflieger: 개인적으로 지동원 선수를 좋아하고 분데스리가로 진출했으면 하는 선수로 그를 1 번으로 꼽아왔기에 온다면 매우 기쁘겠지만 EPL 에서 부진해서 BL 로 임대되어 온다는 것이 조금 아쉽다. 바이에른으로 당당하게 입성해주길 ...
0.2923	선덜랜드, 지동원 임대 이적 검토 http://t.co/md1AFiI0

[표 20]의 상위 순위 트윗 결과와는 다르게 [표 21]의 하위 순위 트윗 결과는 부족한 정보를 제공한다. 하위 순위 트윗들을 살펴보면 지동원이 임대를 할 것이라는 내용은 포함하고 있으나

어느 팀으로 갈지에 대한 언급이 없다.

ㅂ. 군집 6 결과

군집 6은 군집 5와 마찬가지로 ‘지동원 임대’ 관련 주제로 이루어진 군집이다. 군집에 포함된 객체 수는 139개이고 순수도는 1이 나왔다. 군집 6에서 트윗 순위를 정한 뒤 상위 순위 트윗들을 살펴보면 다음 [표 22]과 같다.

[표 22] 군집 6의 상위 순위 트윗 결과

Similarity	Tweet
0.9300	RT @cdk0649: [단독] 지동원, 아우크스부르크 임대 확정…구자철과 한솥밥 http://t.co/cAqrzUjc"
0.9093	지동원 아우크스부르크 임대 확정,지동원 구자철 한솥밥 http://t.co/EnmZSFCK
0.8887	RT @naver_sports: [해외축구] 지동원, 아우크스부르크 임대 확정…구자철과 한솥밥 http://t.co/Tf7np6Kg

상위 순위 3개의 트윗을 살펴보면 ‘지동원’, ‘아우크스부르크’, ‘임대’, ‘확정’, ‘구자철’, ‘한솥밥’과 같은 주요 내용들이 포함된 것을 알 수 있다. 이를 토대로 군집 6의 토픽은 ‘지동원이 구자철이 소속 된 아우크스부르크로 임대 확정’이라는 걸 파악할 수 있다.

[표 23] 군집 6의 하위 순위 트윗 결과

Similarity	Tweet
0.2751	RT @Kampfflieger: 개인적으로 지동원 선수를 좋아하고 분데스리가로 진출했으면 하는 선수로 그를 1 번으로 꼽아왔기에 온다면 매우 기쁘겠지만 EPL 에서 부진해서 BL 로 임대되어 온다는 것이 조금 아쉽다. 바이에른으로 당당하게 입성해주길 ...
0.2923	선덜랜드, 지동원 임대 이적 검토 http://t.co/md1AFiI0

[표 22]의 상위 순위 트윗 결과와는 다르게 [표 23]의 하위 순위 트윗 결과는 부족한 정보를 제공한다. 하위 순위 트윗들을 살펴보면 지동원이 임대를 할 것이라는 내용은 포함하고 있으나 '구자철' 과 '아우크스부르크'와 같은 주요 내용을 언급하지 않았다.

라. 결과 논의

군집화 결과 k-평균 군집화, 계층적 군집화 할 것 없이 매우 높은 순수도가 나왔다. 이는 미리 정한 5개의 주제가 매우 상이했기 때문에 라고 생각할 수 있다. 특히, 분석에 명사만을 사용하였기에 실제로 분석에 사용된 단어들은 주제 별로 많은 차이를 보였고, 그 결과 이와 같은 높은 순수도 값을 얻을 수 있었다.

k-평균 군집화 결과, 군집 5와 군집 6의 경우 같은 주제에 대한 트윗들이 둘로 나뉘었음을 보여준다. 토픽을 추출하여 내용을 비교해보면 군집 5의 경우 '지동원 임대'와 관련된 소문을 이야기하고 있고, 군집 6의 경우 '지동원 임대'가 확정되어 구자철과 한솥밥을 먹는다라는 내용을 다루고 있다. 군집 5와 군집

6 간의 내용의 차이가 있기 때문에 ‘지동원 임대’와 관련된 트윗들을 두 개의 군집으로 나뉜 것은 잘된 분석 결과라고 할 수 있다.

군집 별로 상위 순위 트윗들과 하위 순위 트윗들을 비교해보면 상위 순위 트윗들은 군집의 토픽을 충분히 이해할 수 있는 내용들이고 하위 순위 트윗들의 경우 토픽을 제대로 반영하지 못함을 알 수 있다. 이를 통해 상위 순위를 트윗을 선정하는 방법이 타당함을 확인할 수 있었다.

제 5장 결 론

5.1. 결론

이번 연구에서는 한국어 트위터에서 토픽을 추출하는 전 과정을 제시하였다. 트위터를 수집하는 것부터 전처리, 군집화, 토픽 추출에 이르는 전체 흐름을 하나의 체계(framework)로 제안하였다. 각 과정 별로 연구의 성과를 요약하면 다음과 같다.

첫째, 트위터 데이터를 전처리하는 과정에서 트위터 분석에 적합한 새로운 사전을 구축하여 분석의 정확도를 높였다. 신조어와 고유명사를 포함한 분석을 수행하기 위해 '네이버 실시간 검색어'와 '네이버 핫토픽 키워드'에서 가져온 단어들로 새로운 사용자 정의 사전을 구축하였다.

둘째, 평균 Silhouette 값을 이용하여 최적의 군집의 수를 정한 다음 k-평균 군집화와 계층적 군집화를 수행하였고 그 결과 매우 높은 순수도 값이 나오는 것을 확인하였다. 이는 전처리 과정에서 명사만을 이용한 점이 높은 순수도에 영향을 미쳤다고 판단할 수 있다.

셋째, 군집화 과정 이후 군집 별 문서 순위를 계산하였다. 문서의 중심값과 각 트윗 간의 코사인 유사도를 이용하여 구한 결과 타당성이 있음을 정성적 평가를 통하여 확인하였다. 각 군집의 상위 순위 트윗들은 군집 내의 핵심 토픽을 기술하는 반면, 하위 순위 트윗들은 중심단어와 관련이 적은 잡담, 개인적 이야기, 광고

등을 포함하고 있었다.

5.2. 향후 연구 방향

트윗 상에서 토픽을 추출하는 연구는 최근 몇 년간 크게 발전하였으나, 한국어 트윗 상에서 토픽을 추출하는 연구는 전무한 실정이다. 본 연구에서는 한국어 트윗 상에서 토픽을 추출하는 전 과정을 제시하였다. k-평균, 코사인 유사도 등 간단한 기법들과 제한적인 데이터를 사용하였다. 따라서 한국어 트윗 상에서 토픽을 추출하는 기법 향상에 대한 향후 연구 방향은 아래와 같다.

첫째, 제한적인 데이터가 아닌 대규모의 데이터를 실험에 사용해 본다. 본 연구에서는 미리 정한 5가지 주제의 트윗을 선정하여 분석에 사용하여 좋은 결과를 얻었으나 실제 다양한 주제의 대규모 데이터에 적용이 가능한지에 대한 의문이 남는다. 하루 전체 데이터 혹은 그 이상의 데이터에 대한 실험을 통해 결과를 검증해 보아야 한다.

둘째, 본 연구에서 사용하였던 k-평균 군집화와 계층적 군집화 이외에도 텍스트 군집화에 적합하다고 알려진 LDA(latent Dirichlet Allocation), Affinity propagation과 같은 방법론들에 대한 적용을 통해 군집화 성능 향상을 얻을 수 있는지 확인한다.

마지막으로, 문서 순위를 결정하기 위하여 언어적인 특성 이외에도 트위터가 가지고 있는 특성인 리트윗, 답장, 해쉬태그와

같은 정보들을 활용하여 문서 별로 순위 계산 시 가중을 줄 수 있다. 이와 같은 정보들을 트윗의 중요도를 판단하는 지표(Boyd, Golder & Lotan, 2010)로 사용함으로써 분석의 정확도 향상을 가져올 수 있다.

새로운 정보의 채널인 트위터는 활발한 연구 주제이자 다양한 서비스의 원천으로 활용되고 있다. 본 연구에서 수행한 한국어 트윗 상에서의 토픽 추출 연구가 추가 보완된다면 트윗 상에서 다루어지는 토픽들을 활용한 다양한 서비스에 적용 가능할 것이라고 기대해본다.

참고 문헌

시맨틱 웹 첨단연구센터, “한나눔 한국어 형태소 분석기 사용자 매뉴얼”, 한국과학기술원, 2011.06.05

Banerjee, Somnath, Krishnan Ramanathan, and Ajay Gupta. "Clustering short texts using wikipedia." *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007.

Becker, Hila, Mor Naaman, and Luis Gravano. "Beyond trending topics: Real-world event identification on twitter." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM' 11)*. 2011.

Bossard, Aurélien, Michel Généreux, and Thierry Poibeau. "CBSEAS, a summarization system integration of opinion mining techniques to summarize blogs." *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*. Association for Computational Linguistics, 2009.

Boyd, Danah, Scott Golder, and Gilad Lotan. "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter." *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 2010.

Dittenbach, Michael, Dieter Merkl, and Andreas Rauber. "Organizing and exploring high-dimensional data with the Growing Hierarchical Self-Organizing Map." *Proceedings of the 1st international conference on fuzzy systems and knowledge discovery (FSKD 2002)*. Vol. 2. 2002.

Ellen, J. "All about microtext: A working definition and a survey of current microtext research within artificial intelligence and natural language processing." *Proceedings of the Third International Conference on Agents and Artificial Intelligence*. 2011.

Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

Java, Akshay, et al. "Why we twitter: understanding microblogging usage and communities." *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007.

Kaufman, Leonard, and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. Wiley. com, 2009.

Kim, Han-joon, and Sang-goo Lee. "An intelligent information system for organizing online text documents." *Knowledge and information systems* 6.2 (2004): 125-149.

Kim, Sungchul, et al. "Finding Core Topics: Topic Extraction with Clustering on Tweet." *Cloud and Green Computing (CGC), 2012 Second International Conference on*. IEEE, 2012.

Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?."*Proceedings of the 19th international conference on World wide web*. ACM, 2010.

Sankaranarayanan, Jagan, et al. "Twitterstand: news in tweets." *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2009.

Suh, Bongwon, et al. "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network." *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, 2010.

Treeratpituk, Pucktada, and Jamie Callan. "Automatically labeling hierarchical clusters." *Proceedings of the 2006 international conference on Digital government research*. Digital Government Society of North America, 2006.

Wu, Shaomei, et al. "Who says what to whom on twitter." *Proceedings of the 20th international conference on World wide web*. ACM, 2011..

Zeng, Hua-Jun, et al. "Learning to cluster web search results." *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004.

Zhao, Dejin, and Mary Beth Rosson. "How and why people Twitter: the role that micro-blogging plays in informal communication at work." *Proceedings of the ACM 2009 international conference on Supporting group work*. ACM, 2009.

Abstract

Topic extraction on Korean tweets using clustering and document ranking

Sungjin Kang

Department of Industrial Engineering

The Graduate School

Seoul National University

Twitter is the most used microblogging service in the world which spreads effects to the mass quickly because of real-time sharing of information. But it is not easy to find useful information on twitter. More than 200 million tweets are updated daily and most of them are meaningless or repetitive. That is why there is a need for combining information within the same topic in order to acquire information on twitter. Active research is in process in the case of English Twitter related to this topic, whereas there are none in the Korean Twitter.

So this research offers the entire process of extracting topics

from the Korean Tweets. The method introduced in this research consisted of four steps. In the first step, tweets are collected. Using the open API provided by Twitter, one collects Korean tweets. In the second step, one transforms these tweets into a vector space model so that it could be analyzed. Only nouns extracted from each tweets are analyzed. Simultaneously, we use a custom-designed dictionary which starts from the original dictionary, then adds the 'Naver real-time search word' and 'Naver hot topic keyword.' In the third step, tweets are divided into groups according to topic through clustering. In the fourth step, tweets that have been grouped from step three are ranked, and the high-ranked tweets are used to extract important topic.

As a result of calculating the purity in order to evaluate the clustering result, the total average was verified to be clustered according to topic at the rate of 99.4%. By comparing high-ranked and low-ranked tweets by groups, it was shown that extracting topics from high-ranked tweets well represented each group's important topic.

Keywords : Twitter, Korean tweets, clustering, document ranking, topic extraction

Student Number : 2011-23454