지역별 감성 분석을 위한 트위터 데이터 수집 시스템 설계

최기원* · 김희철*

*인제대학교 디지털 항노화 헬스케어학과

Design of Twitter data collection system for regional sentiment analysis

Kiwon Choi* · Hee-Cheol Kim*

*Inje-University

Institute of Digital Anti-aging Healthcare

E-mail : kiwon2819@naver.com* · heeki@inje.ac.kr*

요 약

오피니언 마이닝은 텍스트 속의 감성을 분석해 낼 수 있는 방법으로 작성자의 정서 상태 파악이나 대중의 의견을 알아내기 위해 사용된다. 이를 통해서 개인의 감성을 분석할 수 있듯이 텍스트를 지역별로 수집하여 분석한다면 지역별로 가지고 있는 감정 상태에 대해서 알아 낼 수 있다. 지역별 감성분석은 개인 감성분석에서 얻어 낼 수 없었던 정보를 얻어낼 수 있으며 해당 지역이 어떠한 감정을 가지고 있을 때, 그 원인에 대해서도 파악할 수 있다. 지역별 감성 분석을 위해서는 각 지역별로 작성된 텍스트 데이터들이 필요하므로 트위터 크롤링을 통해서 데이터를 수집해야 한다. 따라서 본 논문에서는 지역별 감성분석을 위한 트위터 데이터 수집 시스템을 설계한다. 클라이언트에서는 특정 지역 및 시간대의 트윗 데이터를 요청하며, 서버에서는 클라이언트로부터 요청받은 트윗 데이터를 수집 및 전송한다. 지역이 가지는 위도, 경도 값을 통해 해당 지역의 트윗 데이터를 수집하며, 수집한 데이터들을 통해 텍스트를 지역 및 시간별로 관리할 수 있다. 본 시스템 설계를 통해 감성분석을 위한 효율적인 데이터 수집 및 관리를 기대한다.

ABSTRACT

Opinion mining is a way to analyze the emotions in the text and is used to identify the emotional state of the author and to find out the opinions of the public. As you can analyze individual emotions through opinion mining, if you analyze the text by region, you can find out the emotional state you have in each region. The regional sentiment analysis can obtain information that could not be obtained from personal sentiment analysis, and if a certain area has emotions, it can understand the cause. For regional sentiment analysis, we need text data created by region, so we need to collect data through Twitter crawling. Therefore, this paper designs a Twitter data collection system for regional sentiment analysis. The client requests the tweet data of the specific region and time, and the server collects and transmits the requested tweet data from the client. Through the latitude and longitude values of the region, it collects the tweet data of the area, and it can manage the text by region and time through collected data. We expect efficient data collection and management for emotional analysis through the design of this system.

키워드

감성 분석, SNS 크롤링, 오피니언 마이닝, 데이터 수집

Ⅰ. 서 론

오피니언 마이닝은 사람들의 의견, 감성, 태도 및 감정을 분석하는 연구 분야로서 자연어 처리 분야에서 가장 활발한 연구분야 중 하나이며 웹 마이닝 및 텍스트 마이닝 분야에서도 널리 연구되고 있다. 오피니언 마이닝은 텍스트 속에 담긴 작성자의 감성을 분석하여 정서적 상태를 파악하거나 특정 주제에 대한 의견을 얻어 내기 위하여 사용된다.[1][2]

이러한 감성 분석을 위해서는 사람에 의해 작성된 텍스트 데이터가 필요하며 최근 오피니언 마이닝에서는 이를 위해 주로 Social Network Service(SNS)를 이용하고 있다. 인터넷 상에서 SNS의 이용률이 급격하게 증가함에 따라 SNS는 개인의 감정과 의견을 표현하는 대표적인 공간이 되었다. 사람들이 SNS를 통하여 자신이 작성한 글을 공개하고, 더 나아가 자신의 감정과 의견을 타인과 공유하기 때문이다. 이처럼 개인의 감정 및 의견을 담고 있는 수많은 텍스트가 존재하는 SNS는 데이터 수집을 위한 훌륭한 도구로 사용될 수 있다.[3][4]

SNS의 데이터는 다양한 SNS 크롤링 API를 통해 수집할 수 있다. API는 해당 SNS의 데이터들을 수집하기 위한 다양한 기능을 제공한다. 예를 들어, 특정 유저의 데이터만 수집할 수도 있고 특정 범위 내에서 작성된 데이터를 수집할 수도 있다.[5]

이처럼 API에서 지원하는 기능을 통해 다양하게 데이터를 수집할 수 있는데, 본 논문에서는 특정 범위를 지정하여 지역별로 텍스트 데이터를 수집하는 데에 초점을 맞추고 있다. 개인의 텍스트를 수집하여 지역별로 텍스트를 분류한다면 오피니언 마이닝을 통해 지역이 가지고 있는 감정 상태에 대하여 알아 낼 수 있기 때문이다. 지역별 감성 분석은 개인의 텍스트 감성 분석에서 알아낼 수 없었던 여러 가지 정보를 얻어 낼 수 있다. 예를 들어, 지역의 감정을 분석해내어 해당 지역의 감정 상태를 날짜별로 분석해낼 수 있으며, 왜 그러한 감정을 가지고 있는지의 원인에 대해서도 파악할 수 있다.

따라서 본 논문에서는 지역별 감성 분석의 기초가 되는 지역별 트위터 데이터 수집 시스템을 설계한다. 2장에서는 트위터 데이터의 크롤링 방법에 대하여 설명하고 3장에서는 지역별 데이터 수집을 위해 지역 코드를 얻는 방법에 대하여 설명한다. 4장에서는 시스템의 구성에 대해 소개하고 5장에서는 결론 및 향후 연구과제에 대하여 논의한다.

II. 트위터 수집

트위터의 데이터를 수집하기 위해선 데이터를 크롤링할 수 있는 API를 사용해야 한다. 대표적으로 Java용 API인 “Twitter4j”와 R용 API인 “TwitterR”이 존재하며 본 시스템에서는 Java로 데이터를 수집하여 위하여 Twitter4J를 사용한다. Twitter4J를 사용하기 위해서는 twitter4j-core jar 파일이 필요하며 API의 권한을 얻어야 한다. 인증을 통해 권한을 얻어야 만이 API를 사용할 수 있으며 본 논문에서는 OAuth 인증을 사용하였다. OAuth 인증을 위해선 apps.twitter.com에서 consumer key, consumer secret, access token, access token secret의 4개의 코드를 얻어야 한다.

그림 1. Twitter4J 권한 획득을 위한 4개의 Key

파일과 권한의 기본 설정이 끝났다면 트위터에 접근할 수 있는 Twitter 객체를 생성할 수 있으며 Query 객체와 Status 객체를 사용하여 특정 조건을 통한 트위터 검색과 데이터 수집을 수행할 수 있다.

Query는 트위터의 조건 검색을 위해 필요한 객체이다. 특정 키워드를 포함한 검색, 특정 언어로 작성된 데이터의 검색, 특정 날짜 또는 특정 범위 내의 데이터 검색 등 다양한 기능을 제공한다. 지역별 데이터 수집을 위해 가장 필요한 조건은 특정 범위 내의 데이터 검색이다. Query 객체 내의 메서드를 통해 범위를 설정할 수 있는데 이때 필요한 것이 해당 지역의 GeoCode 이다. GeoCode는 특정 지역의 위도와 경도를 가지는 값이다. 본 논문에서는 원하는 지역의 GeoCode 값을 얻기 위해 Google Map의 정보를 사용했으며, 이는 3장에서 설명한다.

Status는 트위터 데이터의 정보를 가지는 객체이다. Query에 의해 설정된 조건과 부합하는 데이터들은 Status형의 객체로 반환되는데 Status 안에는 텍스트, 작성일자, 작성한 유저의 정보 등 다양한 정보를 포함한다.[6][7]

III. 지역 코드 수집

Twitter4J를 통해 지역별로 데이터를 수집하기 위해서는 각 지역의 위도와 경도값을 가지는 GeoCode가 필요하다. GeoCode는 수집하길 원하는 지역에 따라 값이 달라지며 같은 지역의 GeoCode라고 해도 값을 제공해주는 곳에 따라서 GeoCode가 달라 질 수 있다. 따라서 본 시스템에서는 GeoCode의 정확성 및 유연성을 위하여 Google Map의 정보를 통해 위치 값을 제공받을 수 있는 Google Maps Geocoding API를 사용한다. Google Map API 역시 Twitter4J와 마찬가지로 API의 권한을 얻기 위한 코드를 얻어야 한다. 그 후, 데이터를 요청할 URL을 생성한다. URL에는 geocode api 주소, output format, 요청 지역명, 권한 인증키 등이 포함되어있다. output format은 json 또는 xml로 설정할 수 있으며 설정한 형식으로 데이터를 제공받을 수 있다. 요청 지역명은 자신이 GeoCode를 알기 원하는 지역의 이름이다. 이렇게 생성된 URL을 통해 원하는 지역의 위치 값을 원하는 포맷으로 제공받을 수 있다. Java의 DomParser 또는 JsonParser를 통해 제공된 데이터 포맷에서 location 값을 추출할 수 있으며 location 내에 포함된 경도 값인 lng과 위도 값인 lat을 사용하여 Query의 GeoCode를 설

정하여 원하는 지역의 트위터 데이터를 수집할 수 있다.[8][9]

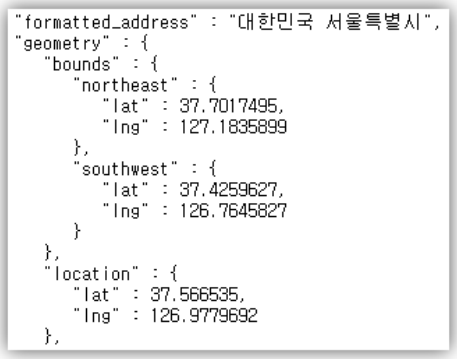

그림 2. Json 형식으로 반환된 서울 위치 정보

Ⅳ. 시스템 구성

시스템은 지역 코드 수집, 트위터 수집, 데이터 저장으로 구성된다.

지역 코드 수집에서는 사용자가 원하는 지역의 GeoCode를 수집한다. 본 시스템에서는 서울과 6개의 광역시를 기준으로 GeoCode를 수집한다. 어플리케이션을 통해 감성 분석을 원하는 지역의 목록을 선택하면 선택된 지역의 목록을 SOAP 기반 웹서비스를 통해 Hadoop 서버로 전송한다.

트위터 수집에서는 전송받은 지역의 GeoCode를 Query의 위치정보로 설정한다. 본 시스템에서는 지역을 서울과 6개 광역시로 설정하고, 전송받은 GeoCode 위치로 부터 반경 15KM를 기준으로 트위터 데이터를 수집한다. 또한 작성된 트윗의 언어를 한글로 제한, 수집할 트윗의 개수 설정 등 기본적인 조건을 설정한다. 트위터 수집은 일정한 시간당 수집할 수 있는 데이터의 양이 제한되어 있다. 따라서 1시간 간격으로 데이터를 수집하는 과정을 배치 프로세싱 한다.

데이터 저장에서는 수집된 트위터 데이터에서 필요한 정보를 추출하고 데이터 베이스에 저장한다. 수집을 요청한 데이터는 Status 객체 형태로 반환되며, Status 객체 안에는 트위터 데이터 자체에 대한 정보와 작성한 유저에 대한 정보들이 포함되어 있으므로 감성 분석에 필요한 요소들을 추출해야 한다. 대표적으로 감성 분석에 필요한 트위터 텍스트를 추출해야 하며, 텍스트들을 구분할 수 있는 ID 등이 필요하다. 이렇게 수집된 데이터는 Hadoop의 비관계형 데이터베이스인 Hbase를 이용하여 저장 및 관리한다. 테이블의 Row Key는 트윗의 ID로 설정한다. 트윗의 ID는 작성된 글의 고유 ID이며 중복되지 않는 값이다. Column Family는 User의 정보를 담는 User Family와 Tweet의 정보를 담는 Tweet Family로 구성한다.

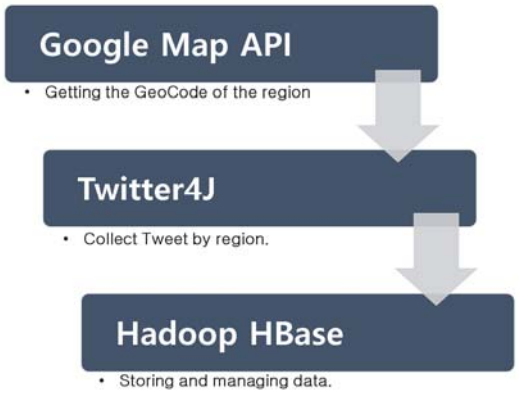

그림 3. Tweet 데이터 수집 프로세스

Ⅴ. 결 론

본 연구에서는 트위터를 이용하여 지역별로 텍스트를 수집할 수 있는 시스템을 설계하였다. 이 시스템은 지역별 감성 분석을 위해 원하는 지역의 트위터 데이터를 수집하며 트위터 데이터에서 감성 분석에 필요한 정보를 추출하여 데이터베이스에 저장 및 관리한다.

지역별 감성 분석을 위해선 각 지역에서 작성된 텍스트가 필요하며, 대량의 텍스트 분석을 위해선 사전에 데이터를 많이 수집해야 한다. 따라서 본 시스템을 통해 각 지역의 데이터를 수월하게 수집할 수 있으며, 1시간 간격으로 트위터 크롤링을 반복 수행하기 때문에 대량의 텍스트 수집에 용이하다.

후속연구에서는 본 시스템을 통해 수집된 데이터들을 이용하여 지역별로 감성 분석을 하여 결과를 제공하는 시스템을 연구한다. 지역별로 수집된 데이터를 분석하여 각 지역의 감정 상태를 알아내고, 감정 상태와 연관된 원인들을 분석하면 유의미한 결과를 도출해 낼 수 있을 것으로 사료된다.

감사의 글

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2017R1D1A3B04032905)과 2017년도 산업통상자원부의 '창의산업융합 특성화 인제 양성사업'의 지원을 받아 연구되었음(과제번호 N0000717)

참고문헌

[1] Liu, Bing., Sentiment analysis and opinion mining, Synthesis lectures on human language technologies, 5.1, 1-167, 2012.

[2] Pang, Bo., Lillian Lee., Opinion mining and sentiment analysis, Foundations and Trends® in Information Retrieval, 2.1‐2, 1-135, 2008.

[3] Gokulakrishnan, Balakrishnan, et al., Opinion mining and sentiment analysis on a twitter data stream, Advances in ICT for emerging regions (ICTer), 2012 International Conference on. IEEE, 182-188, 2012.

[4] Pak, Alexander, and Patrick Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, LREc, Vol. 10, 1320-1326, 2010.

[5] Dopson, Brian, Cardavian Lowery, and Deepti Joshi. Collection and analysis of social media datasets, Journal of Computing Sciences in Colleges, 30.2, 254-261, 2014.

[6] 소주현, 최병현, 송지영, 김용혁, 안드로이드 스마트폰에 기반을 둔 사용자 정보 트윗 시스템, 한국정보과학회 학술발표논문집, 38(2D), 93-96, 2011.

[7] 박종은, 권오진, 이홍창, 이명준, 트위터 사용자를 위한 스마트폰 그룹 채팅 시스템, 한국정보과학회 학술발표논문집, 38(1D), 190-193, 2011.

[8] 정창훈, 김철진, 위치 기반 서비스를 위한 동적 위치 인지 기법, 한국산학기술학회 논문지, 15(7), 4562-4572, 2014.

[9] Middleton, Stuart E., Lee Middleton, and Stefano Modafferi, Real-time crisis mapping of natural disasters using social media, IEEE Intelligent Systems, 29.2, 9-17, 2014.