http://dx.doi.org/10.14400/JDC.2015.13.2.43

효율적인 트윗 분석 시스템 설계 및 구현 방법

최민석
삼육대학교 경영정보학과

An Efficient Method for Design and Implementation of Tweet Analysis System

Minseok Choi
Dept. of Management Information Systems, Sahmyook University

요 약 다양한 소셜 네트워크 서비스의 등장과 사용자의 급증으로 소셜 네트워크 상에서 생산되는 데이터가 급증하고 있다. 전파 속도가 빠르고 개인적 성향의 의견들을 많이 포함하고 있는 소셜 네트워크 데이터의 특성으로 이를 분석하여 다양한 방면으로 활용하려는 요구도 증가하고 있다. 이러한 요구에 부응하여 실시간으로 대용량 데이터를 분석 처리하기 위한 다양한 기술 및 서비스들이 등장하고 있지만, 단기간에 적은 비용으로 그것들을 적용하기에는 어려움이 따른다. 본 논문에서는 새로운 기술이나 서비스의 도입 없이 효과적으로 트윗을 분석하기 위한 시스템 설계 및 구현 방법을 제안한다. 리눅스 기반의 호스팅 서버에 MySQL 데이터베이스와 PHP 스크립트를 이용하여 트윗 데이터를 수집하고 분석하는 모니터링 시스템을 구축하여 제안된 방법을 검증하였다.

주제어 : 트위터, 트윗, 트윗 분석, 소셜 데이터 분석, 소셜 네트워크

Abstract Since the popularity of social network services (SNS) rise, the data produced from them is rapidly increased. The SNS data includes personal propensity or interest and propagates rapidly so there are many requests on analyzing the data for applying the analytic results to various fields. New technologies and services for processing and analyzing big data in the real-time are introduced but it is hard to apply them in a short time and low coast. In this paper, an efficient method to build a tweet analysis system without inducing new technologies or service platforms for handling big data is proposed. The proposed method was verified through building a prototype monitoring system to collect and analyze tweets using the MySQL database and the PHP scripts.

Key Words : twitter, tweet, tweet analysis, social data analysis, social network

1. 서론

스마트 폰, 테블릿 등의 휴대 통신 단말의 보급과 트위터[1], 페이스북[2] 등의 소셜 네트워크 서비스(Social Network Service, SNS)의 등장은 사회적 소통 방식의 패러다임을 바꾸어 놓았다. 사람들은 누구나 쉽게 자신

Received 6 November 2014, Revised 26 December 2014
Accepted 20 February 2015
Corresponding Author: Minseok Choi
Email: mschoi@syu.ac.kr

ISSN: 1738-1916

© The Society of Digital Policy & Management. All rights reserved. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

의 생각이나 의견을 소셜 네트워크상에 올릴 수 있으며, 이러한 데이터들은 네트워크상에서 빠르게 전파된다. 이러한 소통 방식의 변화는 대량의 사회적 데이터의 생산을 유발하게 된다. 대표적 SNS의 하나인 트위터의 2010년 통계자료[3]를 보면 하루에 약 5천만건의 트윗(tweet, 트위터에 등록되는 메시지)이 등록되던 것이 2013년 8월 통계자료[4]를 보면 하루에 약 5억 건의 트윗이 등록되었으며 이는 평균적으로 1초에 5,700건의 트윗이 생성된 것이다. 이러한 사회적 데이터의 폭발적 증가는 데이터의 수집, 저장/관리, 분석 및 활용의 전반에 걸친 다양한 기술에 대한 연구를 촉발시켰다.

데이터를 다루는 기술은 크게 데이터의 획득, 데이터의 저장 및 관리, 데이터의 분석 분야로 나눌 수 있다. 데이터의 획득에서 저장, 관리에 이르는 부분은 데이터 자체의 특성에 관심을 가진 분야로 이전의 정형화된 데이터를 다루던 SQL 기반의 관계형 데이터베이스 기술이 아닌 빠르게 증가하는 비정형 데이터를 다루기 위한 분산 파일 시스템(DFS), 맵리듀스(MapReducde), NoSQL 기반의 기술들이 활발하게 연구되고 있다. 한편 데이터 분석 분야는 대량의 데이터를 기반으로 현실의 다양한 문제들을 해결하기 위한 노력에 초점을 두고 있으며, 해결하려는 문제에 따라 다양한 학문 분야의 지식 및 기술의 활용이 요구된다. 데이터 분석을 통한 인사이트(insight) 제공기술은 차세대 구글(google)을 실현할 수 있는 핵심 기술로도 언급되고 있다[5].

빅데이터(big data)라고 불리는 이러한 대용량 데이터의 활용이 사회 전반에서 다양하게 요구되고 있지만, 앞서 언급한 것 같이 빅데이터 활용을 위한 도구, 플랫폼, 분석 방법 등이 기존의 전통적 기술과 다름에서 오는 생소함이 일종의 진입 장벽으로 작용 될 수 있을 것이다. 따라서 개인이나 조직의 소셜 데이터 활용 폭을 높이기 위해서는 보다 쉽고 효율적인 접근방법이 요구된다.

본 논문에서는 대표적 소셜 미디어인 트위터 데이터를 수집하고 분석하기 위한 시스템 구현 방법을 제안한다. 트위터의 경우 많은 사용자와 빠른 전파속도로 여론 형성에 기여도가 높으며 상품 개발, 마케팅, 고객 대응, 위기관리 등 여러 방면으로 활용 가능한 데이터들의 획득이 용이하다. 데이터 분석을 위하여 기존의 SQL 기반의 관계형 데이터베이스를 활용하여 데이터를 수집하고 관리하며, 웹 표준 기술들을 이용하여 분석 결과를 표현하는 시스템을 구현함으로써 적은 비용으로 소셜 데이터를 다양하게 활용할 수 있는 방법을 제안하고자 한다.

본 논문의 구성은 2장에서 관련 동향을 소개하고, 3장에서는 제안된 시스템의 세부적인 구현 방법 및 기술을 설명한다. 4 장에서 고찰과 결론 그리고 향후 과제로 마무리한다.

2. 관련 동향

2.1 기존 연구

빅데이터 분석을 위한 대표적인 프레임워크는 하둡(Hadoop)[6]이 있다. 하둡은 분산파일 시스템(HDFS)과 맵리듀스를 포함하며 낮은 시스템 요구사항에 높은 데이터 신뢰성을 가지며, 대용량 데이터 처리에 최적화된 시스템이다[7]. 송지훈[7] 등은 하둡을 이용한 트윗 메시지 분석 시스템을 제안하였으며, 한국어 형태소 분석과정을 맵리듀스에 포함시켜 시간을 단축시키는 방법을 사용하였다. 이혁주[8] 등도 SNS 데이터의 의미 분석을 위한 하둡의 맵리듀스 기반 분석 모듈의 구조를 제안하였다.

사용자 간의 관계 정보는 소셜 네트워크 서비스가 가지는 차별화된 특성 중에 하나이며 이러한 관계 정보의 분석은 SNS 분석의 주요 대상 중에 하나이다. 손종수[9] 등은 SNS 데이터가 가지는 관계 정보를 분석하기 위하여 그래프 이론을 적용하여 사용자간 최단거리를 빠르게 찾기 위한 최단 경로 탐색 방법을 제안하였다. 구윤모[10] 등은 마이크로블로그 서비스인 미투데이를 대상으로 사용자들의 관계를 직관적으로 파악하기 위한 소셜 네트워킹 패턴 분석 및 가시화 시스템을 제안하였다.

의사 결정을 위한 데이터 분석도구로서 기존의 정형화된 데이터에 사용되던 데이터 마이닝 등의 분석 기법은 대용량의 비정형 데이터에 대하여 새로운 분석 패러다임과 분석기법이 요구되고 있다. 이병엽[11] 등은 기존의 데이터 마이닝 분석 기법을 소셜 미디어 데이터 분석에 활용하기 위한 방법을 제안하였다. 허정[12] 등은 빅데이터 기반 의사결정 지원 시스템 '소셜위즈덤'에 포함된 이슈 이벤트 검색 시스템을 제안하였다. 유병국[13] 등은 트위터 검색 네트워크 분석 활용을 통한 마케팅 전략 수립에 관하여 연구하였다. 진설아[14] 등은 트위터 데이터를 분석하여 토픽의 변화 시점 및 패턴을 파악하

여 기업에 대한 부정적 여론을 빠르게 파악하고 즉각적인 대응이 가능하도록 하기 위한 방법을 제안하였다.

2.2 소셜 이슈 탐지 및 분석 기술 동향 분석

트윗 분석 시스템의 기능 설계를 위하여 국내의 대표적 소셜 분석 서비스인 소셜메트릭스와 펄스K의 서비스 구조 및 내용을 분석하였다. 소셜메트릭스[15]는 자연어 처리 기술과 텍스트 마이닝 기술을 기반으로 블로그, 트위터 문서를 분석한 정보를 제공하는 서비스이다. [Fig 1]과 같이 사용자가 입력한 키워드에 대하여 소셜 미디어에 노출 빈도 추이와 연관어 정보를 제공한다.

[Fig. 1] Service page of social metrics [15]

펄스K[16]는 트위터, 페이스북, 미투데이 및 블로그 등의 소셜 미디어를 모니터링하고 분석할 수 있는 서비스이다. 소셜 인지도와 호감도를 기반으로 [Fig. 2]와 같이 평판을 수치화하고 모니터링 하는 기능을 제공한다.

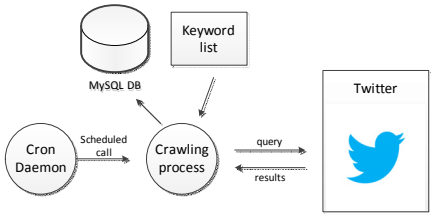
[Fig. 2] Service page of pulse K [16]

3. 트윗 분석 시스템

제안된 트윗 분석 시스템은 소셜메트릭스와 같이 관심 키워드를 기반으로 발생 추이 및 연관어 정보를 모니터링 할 수 있는 기능에 추가적으로 다중 키워드의 비교 모니터링 및 분석, 실시간 위기 검출 그리고 트위터의 특성을 반영한 전파 경로 분석 기능 등을 구현하였다.

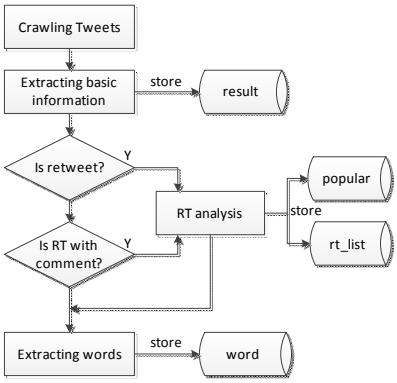
3.1 데이터 수집

트위터는 외부 연동을 위한 다양한 공개 API를 제공하고 있다. 그 중 관심 키워드가 포함된 트윗을 수집하기 위해서 트위터 검색 API[17]를 이용한다. 검색 API는 검색 키워드를 입력하면 최근의 트윗 중에서 해당 키워드가 포함된 트윗을 한 번에 100건까지 제공해 준다. 일정 시간 간격으로 검색을 수행하여 트윗을 수집할 수 있도록 데몬(daemon) 프로그램을 작성할 수도 있지만, 보다 쉬운 구현을 위하여 [Fig. 3]과 같이 리눅스의 크론(cron) 데몬을 이용하여 PHP로 작성된 수집 프로그램을 주기적으로 실행하도록 하였다. 트위터 검색 API는 15분당 180회만 호출을 허용하는 제한이 적용되기 때문에 등록된 키워드의 수에 따라 적절한 실행 주기를 조절해 줄 필요가 있다.



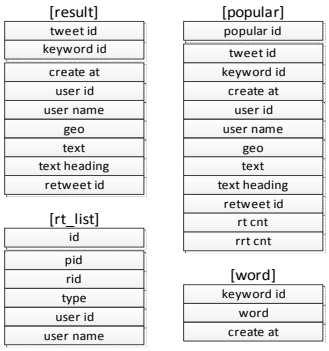
[Fig. 3] Process diagram of data crawling

수집 프로세스에서 수집된 트윗 정보는 [Fig. 4]와 같은 과정을 거쳐 데이터베이스에 저장된다.



[Fig. 4] Flow chart of crawling and storing process

각 과정에서 추출된 데이터를 저장하기 위한 테이블 구조는 [Fig. 5]와 같다.



[Fig. 5] Structure of main tables in database

result 테이블은 트윗의 기본 정보들이 저장되는 테이블이다. 검색된 모든 트윗은 일단 result 테이블에 저장이 된다. 트윗의 확산 정보를 확인하기 위해서는 리트윗(retweet) 정보를 확인하여야 하지만, 시스템의 구축 초기의 검색 API에서는 리트윗에 관한 정보를 제공해 주지 않았다. 일반적으로 리트윗된 글의 처음에는 'RT @name'가 붙게 되기 때문에 해당 패턴으로 시작되는 트윗에 대하여 원본 트윗을 찾아 popular 테이블에 등록하고 리트윗 관계 정보를 rt_list에 저장하였다. 이때 원본 검색의 효율을 높이기 위하여 각 트윗의 앞쪽 일부를 추출하여 따로 저장하여 검색에 이용하였다.

일부 사람들은 리트윗을 하면서 자신의 의견을 첨언할 수 있기 때문에 트위터의 공식 리트윗이 아닌 이전 방식의 리트윗을 이용한다. 일반적으로 인용RT라고 불리는 방식은 새롭게 트윗을 생성하면서 리트윗 하려는 글의 내용을 복사해 붙여 넣는 방식이다. 그러한 인용RT를 찾기 위하여 트윗에서 '@name ' 패턴을 찾아서 검색하는 과정을 추가 하였다. [Fig. 6]은 리트윗과 인용RT의 차이를 보여준다.

[Fig. 6] Example of retweet and retweet with comment

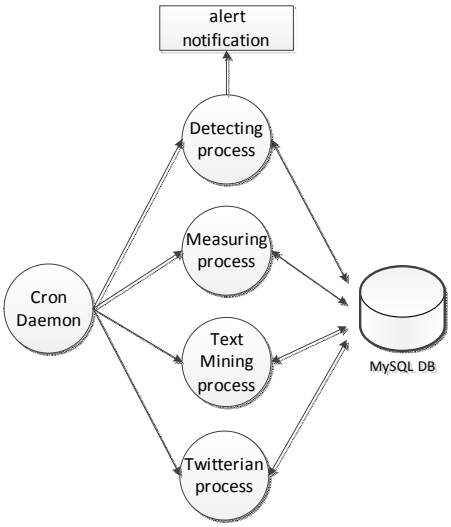
현재의 검색 API에서는 리트윗 정보를 제공해 주고 있기 때문에 리트윗에 대하여는 검색 과정을 거치지 않고 원본의 확인이 가능하지만 인용RT의 경우는 여전히 검색을 통하여 확인하여야 한다.

수집 프로세스의 마지막 단계는 트윗 내용에서 사용된 단어들을 추출하여 저장하는 것이다. 단어 추출은 형태소 분석의 복잡한 과정을 거치지 않고 조사 및 어미 분석을 통하여 간단히 추출하였다.

3.2 데이터 분석 및 요약

수집 프로세스에서 수집된 트윗 데이터를 MySQL DB에 계속 축적하면서 분석 시스템을 운영할 경우 데이터 포화에 따른 성능 문제가 발생하게 된다. 시스템 구축 초기 단계에서 유명 정치인들을 위주로 키워드를 설정하고 테스트 운영한 결과 2달을 채우지 못하고 급격한 DB 성능 저하 및 저장 용량 초과 문제를 맞게 되었다. 이를 해결하기 위하여 일정 시간 단위로 데이터 분석 작업을 수행한 후 그 결과를 저장하여 결과 분석에 이용하고, result 테이블에 수집된 트윗 원본은 일정 기간 경과 후 삭제하는 방법을 적용하였다. 일반적으로 트윗의 수명주기(lifetime)는 매우 짧은 편이기 때문에 하루 정도의 데이터만으로도 실시간 분석 및 트윗 확인 용도로 사용하기에 지장이 없으며 나머지는 분석된 데이터만을 보관하여도 데이터 활용에 크게 문제가 없을 것이다. 필요하다면 용도에 맞게 보관 기관을 적절히 조정하여서 운영할 수 있다. 리트윗된 트윗은 따로 popular 테이블에 보관되기 때문에 이슈가 된 트윗은 삭제되지 않고 보관된다. [Fig. 7]은 데이터 분석 및 모니터링 프로세스 운영을 도식화 한 것이다. 역시 크론 데몬을 활용하여 프로세스 별로 설정된 주기마다 PHP로 작성된 스크립트가 실행되도록 처리하였다.

Detecting process는 해당 키워드에 대한 급격한 트윗 증가를 감지하여 위험 상황을 알려주기 위한 프로세스이다. 구현 시스템에서는 15분 주기로 실행하였으며 시간당 트윗 빈도를 계산하여 관리자가 설정해 놓은 임계 값 이상의 변이가 발생할 경우 DB에 기록하고 필요할 경우에는 휴대폰 문자 서비스(sms) 등을 이용하여 관리자에게 통보할 수 있도록 구현하였다.



[Fig. 7] Process diagram for data analysis

Measuring process는 30분 주기로 실행하였으며 30분 단위의 트윗 관련 각종 통계 데이터를 계산하여 DB에 기록하고 설정된 기간이 지난 트윗들을 result 테이블에서 정리하는 작업을 수행한다. 리트윗이 되어 popular 테이블에 저장된 트윗의 경우도 리트윗 횟수에 임계값을 설정하여 임계값보다 적은 트윗들은 정리하였다.

Text mining process는 추출된 단어들을 정리하는 프로세스이다. 각 트윗에서 추출된 단어의 수가 많기 때문에 모든 단어를 가지고 실시간 히스토그램 분석을 할 경우 시간이 지체되기 때문에 발생 빈도가 높은 단어들만을 추려서 따로 저장하고 그 단어들에 대해서만 단위 시간 히스토그램 데이터를 계산하여 축적하였다. 정리가 끝난 나머지 단어들은 삭제하여 DB 부하를 줄이도록 하였다.

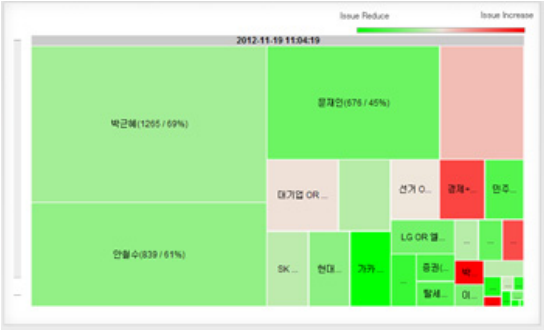
Twitterian process는 해당 키워드에 대한 트윗을 많이 남긴 사용자를 확인하기 위한 프로세스이다. 일정 시간 후 원본 트윗이 삭제되기 때문에 사용자 별로 트윗 작성 통계를 구하여 남겨두어야 한다. 1시간 단위로 실행하여 해당 시간 동안 사용자별 트윗의 개수를 계산 후 임계치 보다 상위에 속하는 사용자들의 정보를 기록하여 축적하였다.

3.3 사용자 인터페이스

사용자 인터페이스 구현은 웹을 기본 플랫폼으로 하여 구현하였다. 그래프 등의 시각화 표현을 사용하기 위하여 HTML5를 기반으로 javascript로 구현된 chart 라이브러리를 이용하였다. 분석된 결과를 사용자에게 제공하기 위한 대쉬보드는 6가지 메뉴로 구성하였다.

[Fig. 8] Main menu of user dashboard

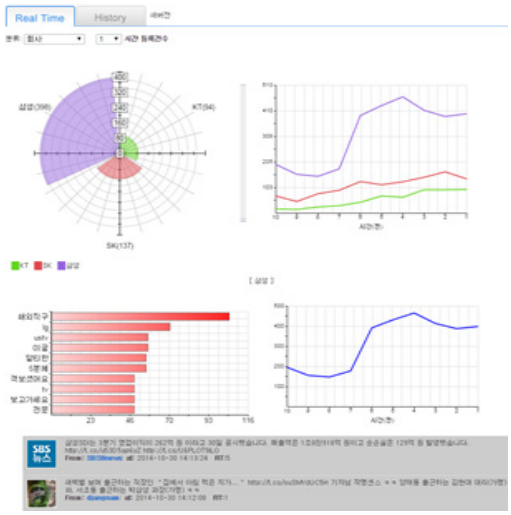
Issue tree map은 등록된 키워드들의 상태를 한 눈에 파악할 수 있도록 하기 위하여 트리맵(tree map)을 이용하여 표현하였다. [Fig. 9]와 같이 트리맵 상에 키워드 별로 트윗 발생 빈도를 면적으로 표현하였으며, 색상을 이용하여 현재 증가 추세(붉은색)인지 감소 추세(녹색)인지를 표현하였다. 매 분 단위로 데이터를 자동 갱신하여 지속적인 모니터링이 가능하도록 구현하였다.



[Fig. 9] Tree map of registered keyword

Issue measuring은 [Fig. 10]과 같이 키워드 별로 현재 단위 시간당 트윗 발생 건수와 과거 발생량 추이, 추출된 연관 단어들과 실제 수집된 트윗 내용을 확인할 수 있도

록 구현하였다. 이 페이지도 자동 갱신 기능을 구현하여 특정 키워드에 대한 모니터링이 가능하도록 하였다.



[Fig. 10] Sample page of issue measuring

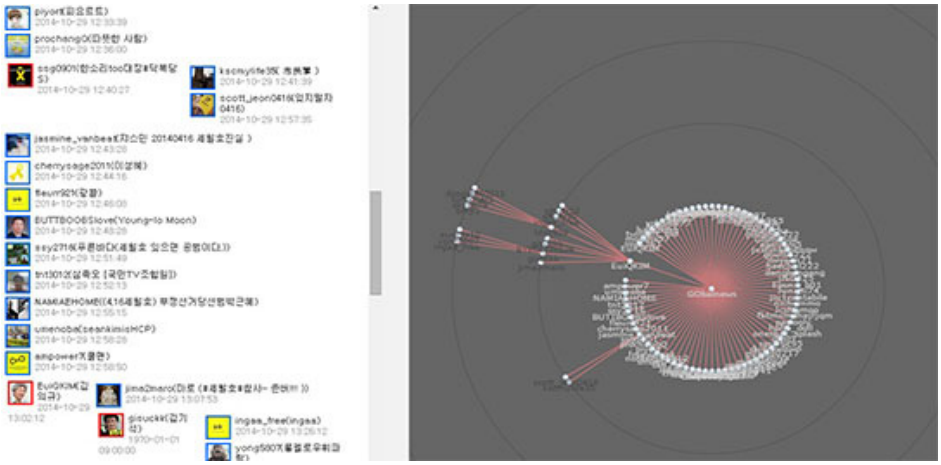
Risk 검출 메뉴는 앞서 설명한 detecting process에서 감지된 위험 상황에 대한 로그를 확인하고 상황 분석을 할 수 있도록 기능을 구현하였다.

Issue relationship 메뉴는 키워드 별로 리트윗이 많이 된 트윗들을 확인하고 [Fig. 11]과 같이 전파 트리와 전파 그래프를 이용하여 해당 트윗의 전파 상황을 시각화 하여 표현하였다. 리트윗이 아닌 인용RT의 경우 인용RT 트윗에 다시 리트윗이나 인용RT가 가능하기 때문에 전파 경로를 3단계까지 표현하였다.

Text mining 메뉴는 [Fig. 12]와 같이 키워드 별로 트윗에서 추출된 단어들의 히스토그램을 분석하여 그 결과를 확인할 수 있도록 구현하였다.

[Fig. 12] Sample page of text mining

Twitterian 메뉴는 키워드 별로 설정된 기간별로 트윗을 많이 작성한 사용자를 확인할 수 있도록 구현하였다. 사용자의 트윗이 리트윗이 많이 되어 보관되지 않았다면 작성한 원본 트윗은 삭제되기 때문에 해당 사용자가 작성한 트윗을 직접 보여주는 기능은 제공되지 못하였고 사용자 트위터로의 연결 기능만을 구현하였다.



[Fig. 11] Visualization of tweet propagation

3.4 시스템 운영 평가

제안된 시스템의 평가를 위하여 리눅스 기반의 웹 호스팅 서비스와 기본 제공되는 MySQL 데이터베이스를 이용하여 트윗 모니터링 시스템을 구축하였다. 정치인들과 대기업, IT 관련 제품 등의 키워드 17개를 등록하고 6개월 간 운영한 결과 약 2.5백만 건의 트윗이 수집되어 분석되었으며 약 1.4백만 개의 주요 연관어가 추출되었고, 약 66만 건의 인기 트윗과 약 19.5백만 건의 리트윗 전파 경로가 수집되었다. 제안된 방법에 따라 일정 시간 단위로 데이터를 분석하여 결과를 누적시키는 방식을 적용하여 장기 운영에도 데이터 누적에 따른 데이터베이스의 성능 저하나 데이터 분석의 시간 지체 현상은 발생하지 않았다.

4. 고찰 및 결론

본 논문에서는 대표적인 소셜 네트워크 서비스인 트위터의 트윗을 분석하기 위한 시스템 구현 방법을 제안하였다. 소셜 네트워크가 가지는 여러 특성들에 의해 트윗 데이터를 분석하여 활용하고자 하는 요구가 증가하고 있지만 막연히 빅데이터라는 장벽에 방향을 잡지 못하고 있다면 제안된 방법을 통하여 기존의 웹 서버와 관계형 데이터베이스를 기반으로 익숙한 분석 방법들을 이용하여 원하는 정보를 획득하고 활용하는 것이 가능할 것이다.

제안된 시스템의 평가를 위하여 저렴한 비용의 리눅스 기반의 웹 호스팅 서비스와 기본 제공되는MySQL 데이터베이스를 이용하여 시스템을 구현하여 운영한 결과 데이터베이스 및 서비스 운영에 무리가 없음을 확인하였다. 향후 단어 추출을 위한 형태 분석 및 품사 분석 방법에 대한 보완과 트윗 내용의 긍/부정 자동 분석 방법에 대한 연구가 추가되면 시스템의 완성도가 높아질 것으로 생각된다.

REFERENCES

[1] http://www.twitter.com
[2] http://www.facebook.com
[3] https://blog.twitter.com/2010/measuring-tweets
[4] https://blog.twitter.com/2013/new-tweets-per-second-record-and-how
[5] Special Report : The next Google, Nature, Vol.455, Sep. 2008.
[6] http://hadoop.apache.org
[7] Song Ji Hoon, Lee Si Jin, Park Hyo Dong, Twitter message analysis system design using Hadoop, Proceedings of Korean society for internet information 2012, Vol. 13, No. 1, pp.169-171, 2012.
[8] Hyeokju Lee, Myoungjin Kim, Hanku Lee, Hyogun Yoon, Design and Implementation of an Analysis module based on MapReduce for Large-scalable Social Data, Proceedings of Korea Computer Congress 2011, Vol.38, No.1(B), pp.357-360, 2011.
[9] Jong-Soo Sohn, Soo-Whan Cho, Kyung-Lag Kwon, In-Jeong Chung, Improved Social Network Analysis Method in SNS, Journal of Intelligent Information System, Vol. 18, No. 4, pp.117-127, 2012.
[10] Yun-Mo Koo, Jeongjin Lee, Jinwook Seo, A Visual Analytics for Analyzing Social Networking Patterns among Microbloggers, Journal of Korea Game Society, Vol. 12, No. 3, pp.77-86, 2012.
[11] Byoung-Yup Lee, Jong-Tae Lim, Jaesoo Yoo, Utilization of Social Media Analysis using Big Data, Journal of the Korea Contents Association, Vol. 13 No. 2, pp 211-219, 2013.
[12] Jeong Heo, Pum-Mo Ryu, Yoon Jae Choi, Hyun Ki Kim, Cheol Young Ock, An Issue Event Search System based on Big Data for Decision Supporting: SocialWisdom, Journal of KIISE, Vol. 40, No. 7, pp. 381-394, 2013.
[13] Byong-Kook Yoo, Soon-Hong Kim, Marketing Strategies using Social Network Analysis: Twitter's Search Network, Journal of the Korea Contents Association, Vol. 13 No. 5, pp. 396-407, 2013.
[14] Seol A Jin, Go Eun Heo, Yoo Kyung Jeong, Min Song, Topic-Network based Topic Shift Detection on Twitter, Journal of Korea Society for Information Management, Vol. 30, No. 1, pp. 285-302, 2013.
[15] http://www.socialmetrics.co.kr
[16] http://www.pulsek.com

[17] https://dev.twitter.com/rest/public/search

최 민 석(Choi, Min seok)

- 1996년 2월 : 한양대학교 전자공학과(공학사)
- 1998년 8월 : 한양대학교 전자공학과(공학석사)
- 2004년 8월 : 한양대학교 전자공학과(공학박사)
- 2012년 3월 ~ 현재 :삼육대학교 경영정보학과 조교수
- 관심분야 : 정보처리, 정보시스템
- E-Mail : mschoi@syu.ac.kr