트위터 사용자 정보 기반의 유사성 순위 시스템

양새동 · 김재윤 · 사잔 쿠말 · 김창수 · 정회경

배재대학교 컴퓨터공학과

Twitter User Information based Users Similarity Ranking System

Xi-tong Yang · Jae-Yoon Kim · Sajan Kumar · Chang-Su Kim · Hoe-Kyung Jung

Department of Computer Engineering, PaiChai University

E-mail : withchyang1@gmail.com, foxboss@hanmail.net, sajank@hotmail.com, MIE-ddoja@pcu.ac.kr, hkjung@pcu.ac.kr

트위터는 140자를 한 번에 올릴 수 있는 트윗을 사용하여 전 세계적으로 다양한 사람들과 소통할 수 있다. 또한, 트위터는 팔로우 기능을 제공하여 메신저와 같은 신속성도 제공한다. 이로 인해 트위터를 사용하는 사용자의 수가 급증하였고, 스마트 폰의 대중화로 인해 생활의 일부분이 되었다. 하지만 트위터의 많은 데이터로 인해 사용자의 정보와 유사한 사용자나 정보가 추천되지 않는 단점을 가지고 있다. 이러한 문제점을 보완하기 위하여 본 논문에서는 사용자의 정보 기반으로 유사성을 필터링하여 순위를 정하고 사용자에게 유사한 사용자나 정보를 추천하는 시스템을 제안한다.

본 논문에서 제안하는 시스템은 사용자의 트위터 계정을 사용하여 데이터를 수집하는 모듈과 수집된 데이터를 필터링 및 추천하는 모듈로 구성되어 있다. 이러한 모듈들은 Open API와 Mahout을 사용하여 설계 및 구현하였다.

ABSTRACT

Twitter is using Tweets to post 140 characters at a time to interact with different people around the world. In addition, Twitter will also provide speed, such as instant messaging by providing the follow feature. This was used for increasing the number of users because of the tweeter, a portion of the life was due to the popularity of smart phones. However, because of the large amount of data of the tweeter has a disadvantage similar to the user information or user information is not recommended. In this paper, in order to compensate for this problem to establish a ranking filter the similarity information based on a user's system, we propose that the user or the like similar to the user information.

The system proposed in this paper consists of the collected data and modules to collect data using a user account in the filtering and the like to the tweeter module. These modules use the Open API and Mahout designed and implemented.

키워드

Mahout, Open API, Social Network, Twitter

Ⅰ. 서론

휴대용 스마트 기기와 광대역 인터넷망이 형성되어 트위터를 사용하는 사용자들이 증가하고 있으니까 트윗 건수도 증가하고 있다[1]. 그리고 트위터의 성장과 트위터를 사용하여 효과적으로 홍보를 하는 해외의 성공 사례에 자극을 받아 국내에서도 트위터를 도입하는 기관이나 기업들이 증가하고 있다. 트워터는 사용자가 실시간으로 정보를 생산하고 소비하며 이러한 데이터들은 정형 데이터 보다 비정형 데이터가 많아 기존의 시스템에서는 데이터 처리가 비효율적이다[2]. 또한, 이러한 데이터들은 대용량으로 생산되고 있어 이를 처리하기 위한 기술을 요구한다.

본 논문에서는 트워터를 사용하는 사용자들의 계정을 기반으로 트위터의 키워드를 수집하여 사용자 성향과 유사한 사용자나 그룹, 정보를 추천하는 시스템 구현을 제안한다. 제안하는 시스템은 빅데이터가 요구하는 기능을 충족할 수 있고, 간편하게 구현이 가능하다. 또한, 모듈별로 투명성

을 제공하여 다양한 시스템으로 변경하여 사용할 수 있다.

Ⅱ. 관련 연구

본 장에서는 제안하는 시스템에서 사용하는 Open API와 Mongo-Hadoop 커넥터, Mahout에 대해 기술한다.

2.1 Open API

트위터에서는 REST(REpresentational State Transfer) 기반의 Open API를 제공하며 OAuth라는 인증을 통하여 사용자를 인증하거나 트위터가 제공하는 기능을 API 기반으로 사용할 수 있다[3]. 그리고 REST API는 기존의 기술보다 간단하게 구현할 수 있고 확장성을 제공한다.

2.2 Mongo-Hadoop 커넥터

Mongo-Hadoop 커넥터는 NoSQL인 MongoDB와 분산 처리 기반에 사용되는 Hadoop을 연결하여 사용한다[4]. 그리고 오픈 소스 기반의 라이브러리이며 MongoDB에서 Hadoop의 MapReduce를 사용하여 데이터를 입출력할 수 있도록 제공한다. 또한, 다양한 Hadoop의 에코시스템에도 적용하여 사용할 수 있다.

2.3 Mahout

Mahout은 Hadoop 기반으로 분산처리가 가능하고 확장성을 제공하는 기계학습용 라이브러리이다[5]. 그리고 다양한 수학 라이브러리와 효과적인 패키지들을 제공하며 데이터를 분류하고 군집, 패턴 마이닝, 벡터유사도 등 다양한 알고리즘을 제공한다.

Ⅲ. 시스템 설계 및 구현

제안하는 유사성 순위 시스템은 3단계로 구성되어 있고 구조는 그림 1과 같다. Open API를 이용하여 사용자의 인증을 얻고 데이터를 수집 및 저장하는 단계와 MongoDB와 Hadoop, Mahout을 통해서 데이터를 필터링 및 분석하는 단계, 그리고 최종 결과를 사용자에게 제공하는 단계로 구성되어 있다.

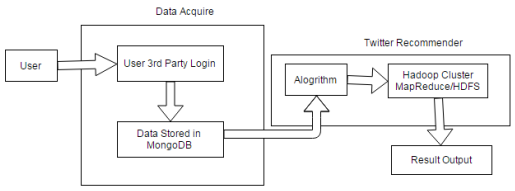

그림 1. 유사성 순위 시스템의 구조도

3.1 사용자의 데이터 수집과 저장

본 시스템에서 데이터를 수집하는 방법은 두 가지로 분류된다. 첫 번째는 트위터에서 제공하는 Open API를 사용하여 사용자의 데이터를 수집한다. 두 번째로는 Twitter4j를 사용하여 사용자의 데이터를 수집한다. 그리고 수집한 데이터는 MongoDB에 저장하며 구조는 그림 2와 같다.

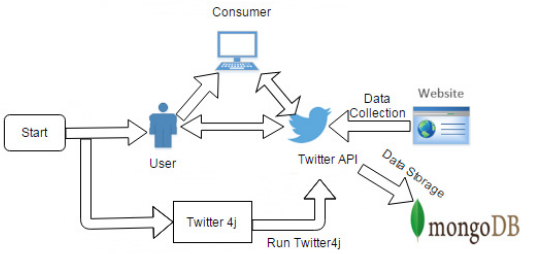

그림 2. 데이터 획득

3.2 키워드 필터링과 분석

트위터 사용자 올린 트윗의 키워드 분석에는 Mongo-Hadoop 커넥터와 Mahout이 제공하는 피어슨 상관관계(Pearson Correlation Similarity)와 유클리드 거리(Euclidean Distance Similarity) 알고리즘을 사용하여 처리한다. 본 시스템에서 키워드 분석의 처리 순서는 그림 3과 같다.

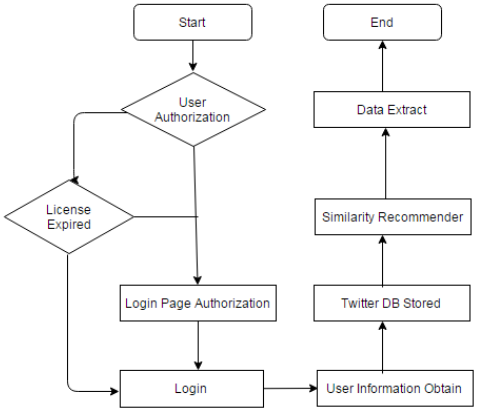

그림 3. 키워드 분석의 처리 순서도

3.3 유사성 순위 시스템

유사성 순위 시스템은 사용자의 트윗 내용의 유사성과 친구, 관심이 있는 친구의 상관성을 분석하고 유사성을 계산하여 사용자의 친구 목록에 있는 사용자의 유사성에 따라 순위를 정하고 제공한다. 제안하는 시스템의 운영체제는 Ubuntu 14.04 LTS를 사용하고, Intel i5 CPU와 8 GB 메모리를 갖춘 PC 환경에서 구현하였다. 그리고 본 시스템의 처리 결과는 웹브라우저를 통해 사용자에게 제공된다.

Ⅳ. 실 험

실험에는 유사한 사용자와 정보를 추천하는 방식으로 진행하였다. 첫 번째로 본 시스템을 이용

하여 트위터 사용자의 트윗 내용을 필터링하고 사용자의 친구 목록을 얻다. 두 번째는 이런 정보의 상관성을 계산하고 유사한 사용자 순위를 사용자에게 추천한다. 유사한 사용자 추천은 5명으로 범위를 설정하고 죄종 수출 화면은 그림 4과 같다.

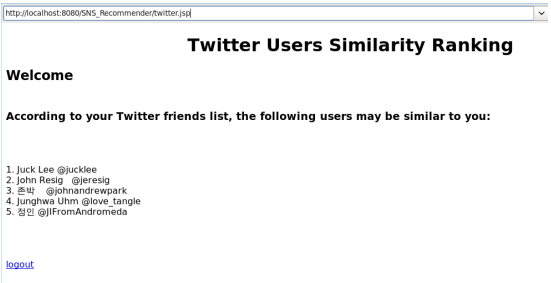

그림 4. 유사성 순위 화면

Ⅴ. 고찰 및 결론

본 논문에서는 트위터 사용자의 계정을 사용하여 유사한 사용자나 정보를 추천하는 유사성 순위 시스템을 제안했다. 제안하는 시스템은 트위터에서 제공하는 OpenAPI나 Twitter4j를 사용하여 데이터를 수집하고 MongoDB로 저장한 뒤 Hadoop과 Mahout을 사용하여 필터링을 진행한다. 그리고 사용자에게 웹브라우저를 통해 정보를 제공한다. 실험에 사용된 시스템은 가상 분산 환경 기반으로 데이터 수집과 필터링을 진행하였다. 실험 결과 처리 데이터 크기가 증가할 수 록 처리 속도는 감소하는 것을 확인하여 효과적인 실험을 위해 물리적인 분산 환경 구축이 필요한 것을 알 수 있었다. 본 시스템은 모듈별로 구성하고 투명성을 제공하여 사용 목적에 맞게 변경하여 사용할 수 있어 다양한 분야에서 응용하여 사용할 수 있을 것으로 예상된다.

향후 연구로는 본 시스템에서 다른 SNS 서비스도 추가하여 사용 범위를 확장할 계획이며 지속적인 연구가 필요하다.

참고 문헌

[1] Kang Ho Lee, “A Study on the Introduction of Twitter According to Its Application Types.” KOCOMA, Vol.37, pp.279-297, 2011.

[2] Wang, Wenbo, et al., “Harnessing twitter ‘big data’ for automatic emotion identification.” 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE, pp.587-592, 2012.

[3] Ritter, Alan, Oren Etzioni, and Sam Clark. “Open domain event extraction from twitter.” Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp.1104-1112, 2012.

[4] Boulmakoul, Azedine, et al., “Mongodb-hadoop distributed and scalable framework for spatio-temporal hazardous materials data warehousing.” The International Environmental Modelling & Software Society, Vol.4, pp.2255-2262, 2014.

[5] Jain, Eeti, and S. K. Jain, “Using Mahout for clustering similar Twitter users: Performance evaluation of k-means and its comparison with fuzzy k-means.” Computer and Communication Technology (ICCCT), pp.29-33, 2014.