스톰을 기반으로 한 실시간 SNS 데이터 분석 시스템

이현경[*] · 고기철[**] · 손영성[***] · 김종배[****]

[*,**,****]숭실대학교 SW특성화대학원

[***]숭실대학교 대학원 IT정책경영학과

E-mail : [*]ketia89@naver.com, [**]jeada4@naver.com, [***]sysgigigi@gmail.com, [****]kjb123@ssu.ac.kr

요 약

광고 효과 분석과 극대화를 위해 기업들이 SNS를 활용하는 비중이 갈수록 높아지고 있다. 특히 광고 효과의 실질적인 효과 분석을 위해 SNS 이용자를 대상으로 한 하둡 기반의 키워드 추출 분석이 곽광을 받고 있다.

기존 하둡 기반 키워드 추출 분석은 저장된 데이터를 Map Reduce 방식으로 처리하는 것이 대부분이다. 이 때문에 정보가 실시간(Real Time)으로 전파되는 SNS의 특성을 온전히 반영하지 못 하는 한계를 가지고 있다.

본 연구에서는 이러한 기존의 하둡 기반 키워드 자동 추출 모델의 한계점을 지적하고, 이를 개선하기 위해 실시간 데이터 분석이 가능한 스톰을 활용하는 모델을 제시하고자 한다.

ABSTRACT

In order to analyze and maximize efficiency of advertise, business put more importance on SNS. Especially, keyword extraction analyses based on Hadoop receive attention.

The existing keyword extraction analyses have mostly MapReduce processes. Due to that, it causes problems data base would not update in real time like SNS system.

In this study, we indicate limitations of the existing model and suggest new model using Storm technique to analyze data in real time.

키워드

하둡, 스톰, SNS, 키워드추출, 마케팅

Ⅰ. 서 론

2010년 이후 전 세계적으로 소셜 미디어 이용자들은 매년 10% 이상의 성장세를 지속해왔다. 특히 SNS(Social Networking System)는 이러한 성장세의 중심에 위치하면서 새로운 광고 플랫폼으로 주목받고 있다. SNS의 대중화와 스마트 미디어의 확산은 광고와 관계 맺음의 혼합되어 있고, 광고와 정보과 혼재되어 있는 융합형 광고가 주요한 형태로 나타나고 있다. 이와 같은 새로운 형태의 광고는 기존의 시장을 분화시킴과 동시에 새로운 시장을 형성함으로써 광고 시장 생태계의 변화를 추동하고 있다. 개인을 포함하는 중소 규모의 광고주들의 광고 시장 진출이 본격화되면서 광고 시장의 저변이 확대되고 있으며, 광고 시장의 전반적인 효율성 또한 향상될 것으로 예상된다. 또한 SNS를 이용한 광고는 기업과 소비자 간의 지속적인 커뮤니케이션을 가능하게 해준다는 점에서 광고주에게는 새로운 대고객 창구로 활용이 가능하며, 이용자에게는 고급 정보를 신속하게 얻을 수 있다는 장점이 있다[1].

이러한 SNS 광고 시장을 기업이 적극적으로 활용함에 따라 하둡 기반의 키워드 자동 추출 시스템이 광범위하게 활용되고 있다. 가장 널리 쓰이는 Map Reduce 기법을 활용한 시스템은 사용자가 원하는 키워드가 포함된 SNS 데이터를 추출해 소비자가 제품 및 기업의 인지도 향상에 미친 실질적인 효과를 분석하는데 초점이 맞추어져 있다. 하지만 이러한 시스템은 실시간으로 데이터와 사용자가 변하는 SNS의 특성을 반영하지 못한다는 단점이 있다. 이와 더불어 데이터의 단순한 분석만을 추구하기에 광고 효과 극대화를 위한 기업의 능동적인 대응에 적합하지 않은 한계를 가지고 있다.

따라서 본 연구에서는 Map Reduce 기법을 사용한 기존 하둡 기반 시스템의 한계를 지적하고, 실시간으로 데이터 분석 및 처리가 가능한 스톰 기법을 사용한 시스템 모델을 제시하고자 한다.

Ⅱ. 관련연구

기존에 진행된 연구에서는 하둡 분산파일시스템(HDFS; Hadoop Distributed File System)과 MapReduce를 포함한 아파치 루씬(Apache Lucene) 프로젝트 중 하나인 하둡 프레임워크(Hadoop Framewok)에 초점을 맞춘다. 하둡 프레임워크는 시스템 요구사양의 제한이 낮으면서도 데이터 신뢰성이 높아 대용량 데이터 처리에 최적화된 파일 시스템으로 평가 받는다. 이러한 하둡 프레임워크의 강점을 살려 보다 많은 데이터를 보다 빠른 속도로 어떻게 처리할 것인지가 기존 연구의 중점적인 과제였다[2].

하지만 이러한 성능과 별개로 MapReduce를 활용한 하둡 프레임워크는 실시간 데이터 처리가 불가능하다는 단점이 있다. 이는 실시간으로 업데이트가 되는 SNS의 특성상 정보의 적절한 활용에 큰 문제점을 초래할 수 있다. 예를 들어 소매 상품을 판매하는 업종의 경우, 상품에 대한 잘못된 정보가 SNS를 통해 널리 퍼질 때 실시간 데이터 분석이 가능하지 않다면 이에 대한 대응이 늦어지는 문제점을 가지고 있다.

이에 대한 대안으로 스톰을 활용해 데이터를 수집 및 분석하고 마케팅을 목적으로 활용하는 BDAS 설계 이외에 스톰 모델에 관한 연구를 찾아보기 힘든 실정이다[3].

Ⅲ. 설 계

본 논문에서는 초단위로 빠르게 변화하는 SNS의 변화에 대응하기 위해, 그림1과 같은 실시간으로 분석이 가능한 시스템인 스톰을 이용한다. 트위터에서 제공하는 Twitter API를 이용하여 트위터 데이터를 수집하고, 수집된 데이터 중 미리 선택한 키워드 별로 정렬된 목록으로 만들어진 튜플, 그리고 이를 연속적으로 수집하는 스트림이 있다. Bolt는 각 키워드 별로 가장 많이 출현한 단어를 카운트 하기 위한 목적이다. 이는 미리 구축해둔 Web Server를 통해 실시간으로 사용자가 검색하여 볼 수 있도록 설계하였다.

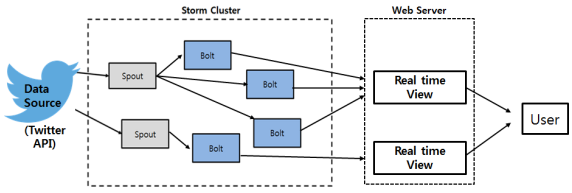

그림 1. SNS 분석설계

제안하는 설계도의 동작단계는 다음과 같다.

1단계 : Twitter API를 통해 수집한다.

2단계 : 수집된 데이터 중에 알고 싶은 키워드를 튜플의 형태로 재구성 한다.

3단계 : 실시간으로 생성되는 튜플들로 구성된 Spout를 통해 Bolt에게 전달한다.

4단계 : Bolt는 전달받은 튜플의 Keyword별 데이터를 취합하고, 동일한 Keyword를 카운트 한다.

5단계 : 웹 상에서 사용자가 원하는 키워드를 검색하면 위의 단계에서 실행한 데이터가 보여진다.

Ⅳ. 결론

본 논문에서는 마케팅을 위해 SNS 데이터를 실시간으로 분석할 수 있는 아파치 스톰 기반 키워드 추출 시스템을 제안하였다. 스톰 기법은 MapReduce 기법에 비해 실시간 데이터 분석 및 처리가 가능하다. 이를 통해 사용자는 실질적인 광고 효과를 면밀히 분석할 수 있고 돌발 상황에 즉각적인 대처가 가능하다는 장점이 있다.

향후 연구 내용으로 스톰 기법을 활용한 키워드 추출 시스템을 SNS 상에 실제 적용시켜, 성능과 정확성을 검증하고 이를 보강하는 연구가 수행되어야 할 것이다.

참고문헌

[1] 성동규, “국내 SNS 광고 현황과 특성 연구”, 한국언론진흥재단 수시 2012-03, pp.12, 2012

[2] 송지훈, 이시진, 박효동 “Hadoop을 이용한 트위터 메시지 분석 시스템 설계”, 한국인터넷정보학회 하계학술대회, 13권1호, 2012

[3] 정이나, 이병관, 박석규, “온라인 마케팅 전략을 위한 SNS와 Web기반 BDAS(Big data Data Analysis Scheme)설계, 한국정보통신학회논문지, 19권1호, pp.141~148, 2015