SNS 환경에서의 광고 오피니언 마이닝 시스템 설계

이수민 유혜진 이아름 방지선 안윤선 김윤희[O]
숙명여자대학교 컴퓨터과학과
2smlee221@gmail.com winnie1219391@gmail.com arsbsp@gmail.com, nin23bix@gmail.com,
ahnysun@sm.ac.kr yulan@sm.ac.kr

A design of Opinion Mining System of Advertisement in SNS Environments

Soomin Lee Hye-Jin Rhu A-Reum Lee Jiseon Bang Younsun Ahn Yoonhee Kim[O]
Dept. of Computer Scinence, Sookmyung Women's University

요 약

SNS(Social Networking System)가 활성화 되면서 많은 사용자들은 실시간적으로 개인적인 의견을 주고받는다. 따라서 SNS 상의 많은 양의 데이터를 분석하여 다양한 새로운 정보들을 생성해 낼 수 있게 되었다. 본 시스템에서는 트위터에 올라오는 제품에 대한 트윗들을 실시간으로 분석하여 해당 제품에 대한 구체적인 광고 반응 평가를 제공하였다. Hadoop을 통하여 많은 양의 데이터들을 실시간으로 분석 하였고 통계적 기법과 오피니언 마이닝을 추가 적용하여 해당 제품에 대한 신뢰도를 분석하였다. 분석 한 신뢰도를 그래프화하여 제공함으로써 광고주들이 실시간적으로 해당 제품의 반응을 파악하기 편리한 시스템을 구축하였다.

1. 서 론

2013년 인터넷이용실태조사[1]에 따르면 인터넷 이용자의 55.1%가 SNS 이용자로 나타났다. 또한 SNS 이용자의 56.2%가 하루에 1회 이상 이용하는 것으로 나타났다. SNS 이용이유로 일상생활에 대한 기록, 개인적 관심사 공유, 전문 정보나, 지식공유가 높은 비중으로 나타난 것으로 보아 SNS 분석을 통해 보다 개인적이고 신속한 정보를 얻을 수 있을 것이다. 따라서 SNS데이터를 이용하여 광고 효과를 분석 할 수 있다고 판단하였다. 기존의 광고효과 측정방법은 체크리스트를 만들어 조사를 하거나 소비자 패널조사 등 사람이 직접 조사하는 방법이 대부분이었다. 이러한 방법은 분석에 소요되는 시간이 오래 걸리고, 높은 정확도를 보장받을 수 없다는 단점이 있다. 따라서, 광고에 대한 효과, 반응을 실시간으로 자동 분석하여 제공되는 프로그램의 필요성이 대두되었다.

본 논문이 웹 시스템은, 이전 논문[2]에서 구현했던 실시간 광고 효과 분석 시스템에, 통계적 기법과 오피니언 마이닝을 추가 적용하여 해당 광고에 대한 평가를 분석하여 광고에 대한 신뢰도를 파악하였다. 이를 통해, 본 웹 시스템은 SNS 상에 광고의 신뢰도 분석과 이에 근거로 단어 카테고리 별 분석을 제공한다. 또한, 동종광고들의 비교 분석 프로그램을 구축하여 광고 반응 및 효과 분석을 사용자가 언제든지 실시간으로 필요한 광고 분석을 할 수 있는 웹 시스템을 구축하였다.

1장의 서론에 이어 2장에서는 관련연구에 대해 서술하였고, 3장과 4장에서는 제안하는 시스템의 기능 설계와 이용방법에 대해 설명하며 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

TV광고의 효과를 간접적으로 계량하기 위한 방법으로 한국방송광고진흥공사(Kobaco)가 프로그램 몰입지수(Program Engagement Index, PEI)를 개발하였다. 이 지표는 TV프로그램의 시청자가 자리를 뜨거나 다른 채널로 돌리지 않았는지, 눈을 떼지 않고 계속 시청했는지 등 TV프로그램의 몰입도를 나타내는 8가지 항목을 측정하여 산출한 것이다. 하지만 이러한 측정 지표는 구체적인 산출 과정에 대한 공개나 검증이 없어 신뢰성에 의문이

[O]교신저자

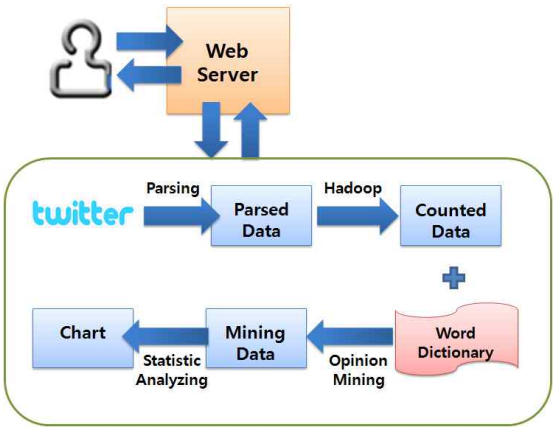
1) 이 논문은 2014년 정부재원(미래창조과학부 여대학(원)생 공학연구팀제 사업)으로 한국연구재단과 한국여성과학기술인지원센터의 지원을 받아 연구되었습니다.

제기되고 있다[3].

현재 SNS 분석 사이트로 소셜메트릭스 인사이트[4] 와 펄스k[5]가 있다. 소셜메트릭스 인사이트는 탐색어 언급도, 긍/부정어 추이 등 다양한 분석을 제공하지만, 검색 시점의 결과만 보여주고 있어 광고효과의 변화 추이는 알 수 없다. 한편 펄스k는 비교분석이나 지역별 분석 등 다양한 분석을 제공하고 있으나 유료이고, 분석 하는데 시간 소요가 있다. 오피니언 마이닝을 통한 텍스트 분석 시스템으로 K-LIWC가 있는데, 글의 구조, 단어 조합의 불균형 등을 통하여 텍스트의 신뢰도를 측정한다. 이를 통해 SNS 데이터 분석과 신뢰도측정이 가능하나, 실시간 분석이 어렵고 체계적인 알고리즘이 부족하다[6].

3. 설 계

본 논문에서 설계한 시스템의 구조는 다음 [그림1]과 같이 하둡을 이용하여 데이터를 수집하여 분산처리 후 오피니언 마이닝과 통계적 기법을 적용하여 웹을 통해 보여주는 형식으로 구성된다. 전체 시스템은 해당 광고의 트윗을 파싱해주는 하둡과 오피니언 마이닝과 통계적 기법을 이용하여 분석해주는 광고 반응 분석기, 사용자에게 분석 결과를 보여주는 웹 인터페이스로 나눌 수 있다. 각 기능에 대한 설명은 다음과 같다.

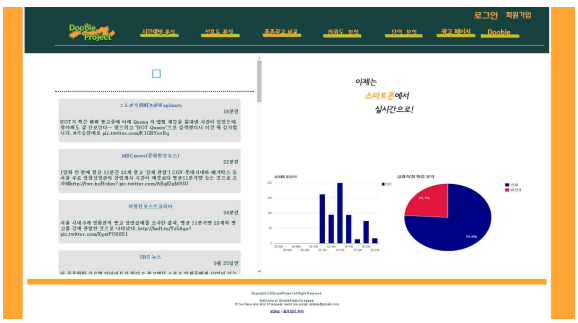


[그림 1] 시스템 구조도

이전 논문[2]에서 구현한 하둡 클러스터에서 파싱해온 데이터에 미리 구축한 단어사전을 토대로 오피니언 마이닝 기법을 추가 적용하여 언급된 문장의 단어 분석을 한다. 본 시스템에서는 긍정어 부정어 유추어 핵심어로 나눈 단어사전을 구축하여 오피니언 마이닝을 적용하여 이를 토대로 각각의 트윗의 단어 구성을 분석한다. 각각의 트윗에서 언급된 단어들은 긍정어 부정어 유추어 핵심어로 마이닝된다. 마이닝된 데이터에 통계적 기법의 식을 적용시켜 광고에 대한 신뢰도를 파악한다. 본 시스템에서는 베이지안 정리를 이용하여 통계적 분석을 한다. 베이지안 정리를 이용하여 단어별 확률을 구한다. 신뢰도 분석 시 마이닝을 통하여 저장해둔 단어 구성에 튜닝을 통해 얻어진 단어별 확률을 적용해 전체적 통계를 구한다. 이를 통해, 본 웹 시스템은 SNS 상에 광고의 신뢰도 분석과 이에 근거로 단어 구성 분석을 도표화 하여 제공한다. 광고주들은 해당제품을 검색하여 웹 서버에서 제공하는 신뢰도 평가와 단어 구성 분석 결과를 얻고 해당 광고에 대한 SNS 사용자들의 인지도 변화와 동종 제품과의 인지도 비교가 가능하다. 본 결과를 통해 구체적이고 신뢰적인 SNS 사용자들의 해당 광고에 대한 반응을 확인할 수 있다.

4. 구 현

먼저, 구현한 웹 서버는 다음 [그림2]과 같다. 실시간으로 파싱된 트윗들은 왼쪽 화면에 띄워져 실시간으로 어떤 내용들이 언급되어지고 있는지 확인이 가능하다. 오른쪽 하단에는 신뢰도와 단어 구성 분포율을 한 눈에 확인 할 수 있는 그래프를 제공한다.



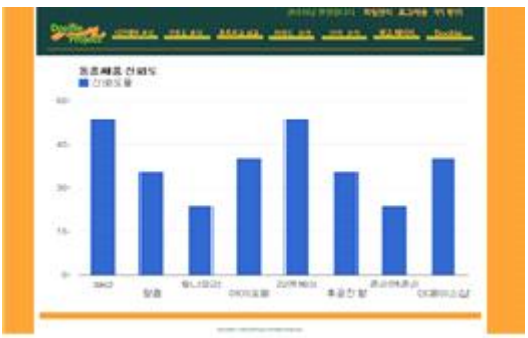
[그림2] 웹 서버 홈페이지

SNS상의 구체적인 반응 분석을 위하여 광고주들은 단어 구성 분석을 참고한다. 단어 구성 분석은 각각의 제품들에 대하여 언급된 트윗들의 긍정어, 부정어, 유추어, 핵심어의 구성 분포율을 보여준다. 또한 동종 제품과의 비교를 할 수 있다. 단어 구성 분석은 다음과 같이 이루어진다. 단어 분석에 앞서 단어 사전을 구축한다. 구축한 단어 사전은 총 1672개의 단어들로, 각각 핵심어, 긍정어, 부정어, 유추어로 나뉘어 데이터베이스에 저장된다. 1시간 별로 crontab을 이용하여 해당 제품에 대해 언급된 트윗들은 하둡으로 파싱되어 모아진다. 모아진 트윗들은 구축한 단어사전을 토대로 오피니언 마이닝을 이용하여 각각 긍정, 부정, 유추, 핵심어로 나누어 단어 구성을 분석한다. 화장품에 대한 단어구성을 검색하면 다음

[그림3]과 같은 결과를 얻는다. 총 8개의 화장품 제품들에 대해서 각각의 SNS상에서 언급되고 있는 트윗들을 바탕으로 그 트윗들의 단어 구성 분석을 한 결과를 보여준다.

[그림3] 동종제품 단어 언급도 그래프

해당 제품의 광고를 통한 제품에 대한 반응을 구체적으로 분석하기 위해서 광고주들은 신뢰도 분석을 참고한다. 신뢰도 분석은 광고를 본 SNS 사용자들의 신뢰도를 계산한 후 동종제품과의 신뢰도를 비교할 수 있다. 신뢰도 분석은 다음과 같이 이루어진다. 단어 언급 분포율을 분석하면서 얻은 마이닝 데이터를 토대로 통계적 기법을 이용하여 신뢰도를 계산한다. 계산된 신뢰도를 토대로 제품별 신뢰도를 분석한다. 화장품의 신뢰도 분석을 검색하면 다음 [그림 4]와 같은 결과를 얻는다.



[그림4] 동종제품 신뢰도 그래프

[그림4]를 보면 총 8개의 제품의 신뢰도를 확인할 수 있다. 그 중 토니모리와 존슨 앤 존슨의 신뢰도가 다른 동종제품들에 비해 낮은 것을 볼 수 있다.이는 [그림3]의 단어 구성 분포율을 보면 토니모리와 존슨 앤 존슨의 경우, 부정적 단어의 분포가 다른 동종제품과 비교해 많은 것을 볼 수 있다. 이에 통계적 기법을 이용하여 계산된 신뢰도가 부정적 단어로 인해 낮아졌다고 볼 수 있다. 반면 신뢰도가 높은 SK2와 리엔케이의 경우, 단어 구성 분포율을 보았을 때 긍정적인 단어의 분포가 다른 동종제품과 비교해 많은 것을 볼 수 있다. 이에 신뢰도가 높은 분포를 띄고 있다. 이를 통해서 신뢰도가 낮은 토니모리와 존슨 앤 존슨의 경우 SNS상에서 부정적인 반응을 보이고 있으며 SK2와 리엔케이의 경우 SNS상에서 긍정적인 반응을 보이고 있음을 확인 할 수 있다.

5. 결론 및 향후 연구

본 논문에서 SNS 데이터를 이용하여 실시간으로 광고효과를 분석해 시각적으로 이해하기 쉽고 자세하게 보여주기 위하여 웹 시스템을 제작하였으며, 오피니언 마이닝을 통해 데이터를 분석하여 광고의 신뢰도를 파악, SNS상의 반응을 파악 할 수 있도록 설계를 확장시켰다. 이 웹 시스템에 대한 평가는 사용자가 얼마나 이용하기 편리한지, 오피니언 마이닝을 통해 구축된 시스템이 얼마나 정확도 높은 분석이 가능한 지가 제시 될 수 있다. 향후에는 웹 시스템 뿐 아니라 접근성이 쉬운 앱으로도 구축을 하여 접근성을 높이는 것이 필요하다. 또한 객관적인 분석을 위하여 SNS 뿐만이 아니라 전체 웹 상의 데이터로 확장시켜 분석하는 것이 필요하다. 또한 현재 분석 가능한 광고 외에 더 다양한 광고 구축과, 오피니언 마이닝에 사용되는 단어사전들의 키워드를 더 다양하게 구축한다면 보다 발전된 광고 분석 시스템이 제작될 수 있을 것이다.

참 고 문 헌

[1] 임재명, 유지열, 장세정, 이정환, 유재민,” 2013년 인터넷이용실태조사” ,2013

[2] 이아름, 방지선, 안윤선, 김윤희, “SNS 환경에서의 지능형 실시간 광고 효과 분석 시스템 설계”, 한국컴퓨터종합학술대회 논문집, 2014

[3] 이승엽, “스마트 광고효과 측정지표 개발 관련 업계 동향”, 동향과 전망:방송•통신•전파 통권 제 66호, 2013

[4] http://insight.some.co.kr/campaign.html

[5] http://www.pulsek.com/pulsek/html/index.html

[6] 김이준, 임지연,” 오피니언 마이닝에서의 텍스트 신뢰도 측정방법” ,2011 한국컴퓨터종합학술대회 논문집,2011