소셜 네트워크에서의 Hadoop 기반 실시간 광고 효과 분석 시스템 설계

방지선, 이아름, 옥윤청, 김윤희
숙명여자대학교 컴퓨터과학부
e-mail : nin23bix@gmail.com, arsbsp@gmail.com, oyc1216@gmail.com, yulan@sm.ac.kr

A Hadoop-based System of Analyzing Real-time Advertisement Effectiveness in Social Network

Jiseon Bang, A-Reum Lee, YoonJung Ock, Yoonhee Kim
Dept. of Computer Science, Sookmyung Women's University

요 약

소셜 네트워크 서비스의 증가로 인해 개인의 관심분야의 수집과 분석이 용이해졌을 뿐만 아니라, 많은 양의 정보를 활용할 수 있게 되었다. 이에 빅 데이터를 이용한 분석이 여러 분야에서 제안되고 있다. 한편, 광고효과 측정 방법에 있어서 빅 데이터 분석은 많은 부분 정확도가 떨어지고, 시간이 오래 걸린다는 단점이 있었다. 때문에 본 시스템에서는 소셜 네트워크에서의 데이터를 파싱하여 TV 광고에 대한 사람들의 반응을 분석하고 그 효과를 그래프로 보여주도록 제작하였다. 본 시스템을 통해 광고효과 분석이 기존보다 빨라졌으며 다양한 방식의 분석이 가능해졌다.[1]

1. 서론

SNS(Social Network Service)의 확산과 동향[1]에 따르면 한국인터넷 진흥원이 실시한 2010년 인터넷 이용실태 조사보고서에서 SNS이용자가 65.7%로 높은 부분을 차지하고 있다고 한다. 또한 SNS 이용자 수는 현재까지 지속적으로 증가해왔으며 앞으로도 증가해 갈 것으로 추정된다. 이와 같이 점차 확산되고 있는 SNS는 신속성, 개인성, 정보의 개방성과 같은 특성을 지닌다. 이에, SNS를 분석했을 때 보다 개인적이고 진실한 정보를 얻을 수 있을 뿐만 아니라 실시간으로 빠르게 갱신되는 정보들을 분석할 수 있다고 추측할 수 있다. 따라서 SNS를 분석함으로써 실시간으로 유효한 데이터를 통해 광고 효과를 분석 할 수 있을 것이다.

SNS는 개인의 관심분야를 수집·분석하는 데에 많이 이용되고 있고 Social big data, 금융혁신의 새로운 도구[2]에서도 국내 금융회사가 보유 고객에 대한 내부 정보와 SNS를 통한 외부의 데이터를 통합 분석하여 고객 맞춤형 마케팅 전략에 적극적으로 활용할 필요성이 있다고 밝혔다. SNS의 빅 데이터를 실시간으로 분석하기 위해서는 SNS의 정보를 얻어오고, 빅 데이터를 분석하기 위한 별도의 시스템이 구성되어있어야 한다. 따라서 기존에 존재하지 않던 실질적인 시스템이 개발되어야 하며, 해당 시스템은 기존의 광고 분석 기법에 의거하여 SNS의 정보를 분석하고 처리하여 타당한 결론을 낼 수 있어야 한다. 이를 위해서, 본 논문에서 제안하는 시스템은 하둡을 이용하여 빅 데이터를 빠르게 분석한다. 빅 데이터는 그 방대한 양으로 인하여 기존 시스템으로는 분석에 시간 소요가 매우 많이 든다는 단점이 있다. 하지만 하둡을 이용하면 빠르게 빅 데이터를 분석할 수 있기 때문에, 이를 이용하여 광고 효과 분석에 이용하면 분석 시간을 크게 단축할 수 있다.

한편, 제안하는 시스템은 광고효과 측정방법에 있어서도 새로운 접근법을 제시하고 있다. 기존의 광고 효과 측정방법은 일일 후 회상 조사(DAR)와 같은 설문조사나 광고 인지율 조사가 대부분이었다[3]. 이는 정확도가 떨어지고 실제 광고 직후 반응을 얻기가 힘들다. 이에, 광고에 대한 반응을 실시간으로 얻을 수 있는 프로그램의 필요성이 대두되었다. 본 논문에서 제안하는 시스템의 목적은 실시간으로 광고 전후의 소셜 네트워크에서의 언급도를 비교하여 광고 반응을 알 수 있도록 분석하고 광고효과를 판단하는 데 있다. 이 시스템은 SNS 빅데이터를 이용한 프로그램이라는 점에서 광고 분석 기법의 새로운 전환점이 될 수 있을 것이라 기대한다. 기존의 방식보다 더 빠르면서도 더 정확한 처리를 가능하게 하는 것이 제안하는 시스템 설계의 목적이다.

기존의 광고 효과는 개별 광고에 대해서만 조사를

[1] 본 연구는 2013 년도 정부(미래창조과학부)재원으로 한국과학창의재단 청년 과학융합 창업 아이디어 창출활동 지원사업으로 수행되었음. 과제번호: 2013AAC0016

할 수 있었다. 하지만 본 논문에서 제안하는 시스템은 개별 광고뿐만이 아니라 같은 주제에 대한 여러 광고를 함께 분석할 수 있다. 기존에 한 번의 조사를 통해 한 개의 광고만을 조사할 수 있었던 것에 비하여 제안하는 시스템이 갖는 장점이라고 할 수 있다. 이 기능을 통해 특정 주제에 대해서 어떤 기법의 광고가 가장 높은 인지도를 내는지, 많은 긍정적 반응을 이끌어 내는지를 알아낼 수 있다. 또한 특정 광고에 대해서 동종광고를 비교 분석할 수도 있다. 이를 통해 타 광고에 비해서 분석하려는 광고가 얼마나 많이 시청되는지 또한 알 수 있다.

제안하는 시스템에서는 트위터와 광고의 데이터를 수집하여, 이를 토대로 세 가지 시나리오에 대해 그래프를 보여준다. 또, 이용자가 광고주일 때에 한하여 새로운 광고를 만드는 데 도움을 주고, 루머가 발생했을 때 알려주는 시스템을 소개한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대하여 설명하고, 3장에서는 시스템 설계와 구현에 대해 설명한다. 4장에서는 분석 결과를 기술하며 5장에서 결론을 맺는다.

2. 관련연구

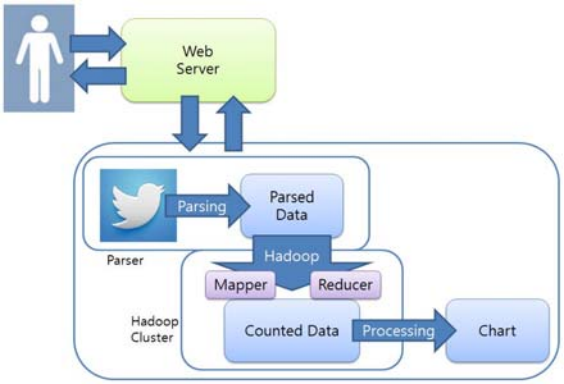
본 논문에서 제안하는 시스템에서 다루는 TV 광고에 대한 광고효과 분석 기법으로는 시청률조사가 있는데, 시청률조사기법은 각 가정에 셋톱박스를 설치하는 방식으로 측정한다. 하지만 조사 기관이 민간기업이고, 그 수가 너무 적어 독점에 대한 문제제기가 있어왔다. 또한 조사 기준이나 절차에 있어서도 문제점이 나타났는데, 구체적으로, 대표성이 미흡하고 비일관적이라는 것이다. 또한, 통계적 오차 범위까지 벗어나, 신뢰도에 있어서도 의심을 받고 있다[4].

한편, 현재 SNS 정보를 수집, 분석하여 보여주는 사이트가 존재하는데, 트윗트렌드[5]와 소셜메트릭스[6]가 그것이다. 하지만, 트윗트렌드의 경우 분석결과가 매우 적어 정확한 분석이 어려웠으며, 소셜 메트릭스의 경우 다양하게 분석이 가능했으나, 여러 광고에 대한 비교는 불가능했다. 또한 두 사이트 모두 분석에 어느 정도의 시간 소요가 있었다.

하둡을 이용한 소셜 네트워크 분석에 대한 논문은 2012 년부터 발표되었고 지난 해인 2013 년에도 발표되었다[7], [8]. [7]번 논문은 하둡을 이용하여 소셜 네트워크 데이터를 이용한 분석이 가능함을 제시한 논문으로, 하둡 시스템과 한글 형태소 분석 프로그램을 도입한 시스템 모델을 설계함으로써 하둡이 SNS 정보를 빠르게 처리할 수 있다는 가능성을 열어 주었다. 하지만 하둡으로 데이터를 처리했을 때, 속도 개선 이외의 다른 성능에 대한 설명이 없었고, SNS 분석에 있어 구체적인 결과에 대한 설명도 미흡했다. [8]번 논문은 한 층 더 발전된 형태로, 직접 크롤링을 통해 데이터를 전달받고, 하둡 시스템을 이용하여 광고 효과를 그래프와 수치로 분석하였다. 이 시스템은 하둡을 이용하여 광고효과를 분석하는 프로그램을 처음 만들었다는 점에서 의의가 있었다. 본 논문은 이 논문의 후속논문으로, 광고 효과를 언급도 이외의 여러 항목에 대해 분석하고, 단일 광고뿐만 아니라 여러 광고를 비교할 수 있도록 하였다.

3. 설계 및 구현

본 논문에서 제안하는 시스템의 형식은 (그림 1)과 같다. 전체 시스템은, 트위터에서 사용자들이 작성한 글 내용을 시간정보와 함께 수집하는 파서, 이 정보를 토대로 조건에 맞는 글의 개수를 계산하는 하둡 프로그램, 이렇게 계산된 정보를 바탕으로 그래프를 보여주는 웹 인터페이스로 이루어져 있다. 각 기능의 자세한 설명은 아래와 같다.



(그림 1) 시스템 구조도

시스템은 크게 데이터 파서(Parser), 하둡 클러스터(Hadoop Cluster), 그리고 웹 서버(Web Server)로 구성되어 있다. 파서는 트위터 글과 광고정보 사이트 데이터를 수집한다. 두 정보는 각각 다른 방식으로 수집되고 저장된다. 자세한 설명은 3.1절에서 다룬다.

하둡 클러스터는 트위터 데이터의 파싱 결과를 세 가지 시나리오 별로 조건에 맞는 항목의 개수를 계산한다. 여기에서 세 가지 시나리오에 대한 설명은 다음과 같다.

- 해당 광고가 어떤 시간대에 방송되는 것이 가장 효과적인지를 알고 이후 광고 방송 시 활용하기 위해 특정 광고에 대한 시간대별 분석을 실시한다.
- 해당 광고가 효과적인 광고인지 여부를 측정하기 위하여 특정 광고에 대한 긍정적/부정적 반응 정도를 분석한다.
- 동종 광고 중에서 어떤 광고가 가장 효과적인지를 측정하기 위하여 동종 광고의 긍정적/부정적 반응 정도를 분석한다.

한편, 이용자가 광고주인 경우에는 다른 두 가지 기능을 제공하는데, 광고 제작 시 활용할 수 있도록 좋은 평가를 받은 광고를 추천해 주는 기능, 루머 발생시 초기 대응을 위해 미리 알려주는 기능이 그것이

다. 이 두 가지 기능을 제공하기 위해, 좋은 평가를 받은 광고와, 루머 글의 개수를 계산한다. 이렇게 계산된 값은 각각 따로 텍스트파일로 저장된다. 자세한 내용은 3.2 절에서 소개한다.

웹 서버는 이렇게 계산된 값을 가지고 이용자의 주문에 따라 도표화 하여 보여준다. 이때 광고 정보 사이트의 파싱 정보가 함께 들어가게 된다. 자세한 내용은 3.3 절에서 소개한다.

3.1 데이터 수집

TV 광고에 대한 트위터 사용자들의 언급 현황과 선호도를 분석하기 위해서는 TV 광고와 트위터 두 가지의 데이터가 필요하다. 파싱에는 Jericho, Jsoup, HTML Parser 를 이용하여 두 정보를 각각 수집하였다.

제안하는 시스템에서 TV 광고 정보를 수집하는 사이트는 TVCF 라는 사이트로, 1950 년부터 현재까지 국내외 TV 광고를 카테고리에 따라 분류해 놓았다[9].

먼저 트위터 데이터의 경우, 트위터 글을 빠르게 업데이트 하기 위해 한 시간에 한 번씩 crontab 을 이용하여 파싱 한다. 파싱은, 이후에 하둡에서 처리를 용이하게 하기 위한 형식으로 작성한 후, 텍스트 파일에 저장한다. 트위터 정보는 (그림 2)와 같이 사용자가 작성한 글을 수집하였다.

(그림 2) 트위터 정보 수집 예

광고정보 사이트 데이터의 경우는, 웹 페이지에서 도표화 할 때 광고가 처음 방영된 날짜와, 광고의 카테고리나 회사정보 같은, 광고의 기본적인 사항을 표시하기 위해 필요하다. (그림 3)은 광고 정보 사이트에 등록된 광고 중 하나이다. 데이터 수집은 각 광고의 정보를 이용하였다.

(그림 3) TVCF 광고정보 수집 예

광고정보 사이트에서 수집한 데이터 중 광고명과 카테고리 키워드는 트위터 데이터 키워드와 매핑된다.

3.2 기능

본 논문에서 제안하는 시스템은 광고의 효과를 보다 구체적이고 정확하게 분석하기 위하여 다섯 가지 기능을 구현하였다. 그 중, 시간대별 언급도 분석, 선호도 분석, 동종광고 분석은 로그인을 한 모든 사용자가 사용할 수 있는 기능이며, 광고 추천 기능과 루머 알림 기능은 광고주로 등록된 사용자가 사용할 수 있는 기능이다.

다섯 가지 기능을 구현하기 위해서 제안하는 시스템에서는 하둡을 이용하였다. 하둡은 파싱을 통해 얻은 정보 중에서 조건에 맞는 항목의 개수를 세어준다. 하둡 프로그램은 맵과 리듀스 과정을 통해 이루어지는데, 트위터 데이터를 입력 값으로 넣으면, 맵에서는 다섯 가지 기능에 따라 시간이나 조건 별 항목의 값이 value가 되고, 조건의 이름이 key값이 되도록 하였다. 리듀스 과정에서는 이렇게 생성된 key와 value값을 통해 실제 개수의 계산이 이루어진다. 맵리듀스 과정을 통해 얻어진 정보는 key와 value값으로 정렬되어 지정된 경로에 텍스트 파일로 저장되도록 하였다.

첫 번째 기능은 시간대별로 광고의 언급도를 도표화하는 것이다. 이는 언급도라는 지표를 통해 얼마나 많은 사람들이 해당 광고에 대해서 알고 있는지, 그 인지도를 나타내기 위하여 설계 하였다. 제안하는 시스템에서는 이를 날짜 별, 시간 별로 그래프화 하였는데 시간의 경우 두 시간 단위로 하여 정확도를 높였다.

두 번째 기능은 트위터에 올라와 있는 글의 선호도를 그래프화 하는 것이다. 여기에서 선호도는 긍정적, 부정적 반응을 토대로 구현하였다. 선호도라는 지표는 단순 언급도 만으로는 해당 광고가 얼마나 긍정적인 여론을 형성하는 지 알기 어렵다는 판단으로 설계하게 되었다.

세 번째 기능은 동종 광고 중 어떤 광고가 더 선호도, 비선호도가 높은지를 그래프화 하는 것이다. 이를 통해, 같은 업종에서 어떤 광고가 더 긍정적인 반응을 이끌어냈는지 분석하였다

광고 추천 기능은 사용자가 광고주일 경우에 이용할 수 있는 기능으로, 광고주가 광고를 제작할 때 더 효과적인 광고를 제작하는 데 도움을 주기 위해, 사용자가 입력한 카테고리 내에서 언급도가 가장 높은 광고를 보여준다. 사용자가 새로운 광고를 제작할 때, 어떤 광고가 가장 인기가 많았는지를 보여주어, 제작에 참고 할 수 있도록 하였다.

루머 알림 기능 역시 사용자가 광고주일 경우에 한하여 이용할 수 있는 기능으로, 사용자가 미리 입력한 키워드가 다수 트위터에 등록되는 경우, 사용자가 이에 상황에 빠르게 대응하게 하기 위해 설계되었다. 사용자에게 메일을 발송하여, 사용자가 루머 발생시에 조기에 대응하여 큰 피해를 입기 전에 미리 대처할 수 있다.

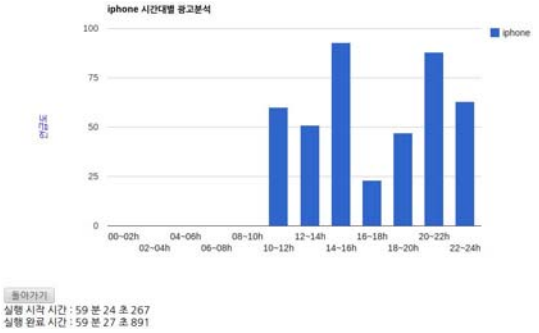
3.3 도표화

사용자는 웹을 통해 자신이 원하는 정보를 그래프로 확인할 수 있다. 제안하는 시스템에서는, 하둡을 이용해 얻은 결과값을 바탕으로, 구글 차트를 이용하여 이를 그래프로 그려 사용자에게 보여주도록 하였다. 이는 사용자에게 시각적 편의성을 도모하기 위해 설계했다. 사용자가 앞서 언급한 다섯 가지 기능 중 하나에 대한 결과를 요청하면, 그에 대한 그래프가 화면에 출력되도록 하였다.

4. 결과

본 논문에서 제안한 시스템은 하둡이 설치된 Master 컴퓨터 한 대와 Slave 컴퓨터 세 대에서 작동한다. 총 네 대의 컴퓨터의 운영체제는 Ubuntu 12.04 이고, CPU는 Intel(R) Core(TM) i7 CPU 870 @ 2.93GHz이며, 8GB RAM을 사용하였다. 설치된 하둡의 버전은 Hadoop-1.2.1이다.

본 논문에서 제안하는 시스템의 기능은, 시스템에서 제공하는 다섯 가지의 기능 중 원하는 항목을 사용자가 요청하면, 1시간마다 수집하는 트위터 데이터를 통해 그래프를 그려 보여주는 것이다. 분석 결과는 (그림4)및 (그림5)와 같다.



(그림 4) 언급도 분석 결과 화면

(그림 5) 선호도 분석 결과 화면

(그림 4)와 (그림 5)는 3월 19일부터 28일까지 수집한 데이터를 통해 도출한 그래프이다. (그림 4)는 iphone의 언급도를 시간대별로 표시한 것으로, 10시부터 24시 사이에만 언급이 있었으며, 14~16시에 가장 높은 반응이 나타났음을 알 수 있다. (그림 5)는 SKT의 선호도를 나타낸 그래프로, 선호도가 75.9%로 더 높았음을 알 수 있다. 각 그래프 아래에는 실행 시작 시간과 완료 시간을 표시하여, 분석에 소요된 시간을 알 수 있게 하였다.

루머 알림 기능의 경우, 메일을 발송하는 파일을 실행파일로 만든 뒤, 스크립트 파일을 이용하여 조건에 해당되면 실행시키도록 하였다. 그 결과, 키워드가 100건이 넘는 경우 정해진 시간에 메일이 발송됨을 확인하였다.

제안하는 시스템에서 15.7MB의 총 92,809개 트위터 데이터를 수집하여 하둡을 실행했을 때, slave를 하나만 두었을 때는 47초의 분석 시간이 소요되었으며, slave를 3개 두었을 때는 35초의 분석시간이 소요되었다.

5. 결론

본 논문에서는 SNS 에서 추출한 데이터를 바탕으로 사용자의 요구가 있을 때마다 하둡이라는 분산 시스템을 통해 실시간으로 처리를 하여 원하는 결과를 그래프로 보여주는 프로그램을 제작하였다.

이 시스템에 대한 평가는 사용자의 요구가 있을 때, 미리 저장된 데이터에서 얼마나 빠른 시간 내에 처리하여 도표화 하는지, 그리고 얼마나 많은 정보를 보여 줄 수 있는지로 제시될 수 있을 것이다. 향후 연구에서는 트위터 데이터뿐 아니라 다른 SNS 정보를 이용할 수 있다면, 더욱 정확한 분석이 가능할 것으로 예상된다. 또, 다섯 가지의 기능 외에 더 다양한 기능을 추가한다면 더욱 발전된 광고 분석 시스템이 제작될 것으로 기대된다.

참고문헌

[1] 이진형, “SNS(Social Network Service)의 확산과 동향”, 인사관리 통권271호, 2012

[2] 하나금융경영연구소, “Social big data, 금융혁신의 새로운 도구”, 2012

[3] You-Jae Lee, “Measuring Advertising Effect”, http://youjae.com/data/cdata3/20/ad13.pdf

[4] 배효승, 신소연, 이상우, “IPTV 셋톱박스 로그분석을 통한 시청률 연구”,방송문화연구 제24권 제1호, 2012

[5] http://tweetrend.com/

[6] http://insight.some.co.kr/

[7] 송지훈, 이시진, 박효동, “Hadoop을 이용한 트위터 메시지 분석 시스템 설계”, 한국인터넷정보학회 2012년도 하계학술발표대회 논문집, 2012

[8] 허서연, 김윤희, “하둡을 이용한 소셜네트워킹의 TV광고효과 분석 시스템 설계, 인터넷정보학회논문지 제14권 제6호, 2013

[9] http://www.tvcf.co.kr/