ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

하둡을 이용한 소셜네트워킹의 TV광고효과 분석 시스템 설계

A Design of Analysis System on TV Advertising Effect of Social Networking Using Hadoop

허 서 연[1] 김 윤 희[1*]
Seoyeon Hur Yoonhee Kim

요 약

빅데이터가 화두가 되면서, 그 대표적인 예인 SNS을 이용한 서비스 개발도 활기를 띠고 있다. SNS는 기존 매체와는 다르게 실시간으로 의견을주고받는 하나의 장으로 확장되었고, 다양하고 많은 개인들의 의견을 분석하고자 하는 서비스들도 늘어나고 있다. 한편, 매체가 다양화되면서, TV광고계에서도 광고에 대한 의견의 확보와 분석에 새로운 접근방법이 필요해졌다. 이에 본 연구에서는 TV광고의 효과를 트위터 데이터를 기반으로 분석하며 특히 하둡을 이용하여 트위터 데이터와 같은 빅데이터를 저장 및 분석하도록 하는 LiveAD라는 시스템을 설계 및 구축하여, 트위터를 대상으로 TV광고 분석을 빠르게 수행할 수 있음을 보여주었다.

☞ 주제어 : 하둡, 광고효과분석, TV광고, 소셜 네트워크 서비스, 트위터

ABSTRACT

As 'Big data' has been one of challenging issues, development of new services using Social Network Service (SNS) which is its typical example became active. SNS has developed as a media where everyone communicates at real time and the number of SNS opinion analyzing services is increasing. Meanwhile, new approach to acquire and analyze twitter data becomes necessary in TV advertisement system. This paper proposes LiveAD system, which store and analyze big data such as twitter data as well as analyze TV advertising effect based on twitter data. As a proof of concept, the proposed system has been implemented collecting and analyzing twitter data using Hadoop. The result of collected information over the system increases the chance of analyzing TV advertising effect on twitter in real-time.

☞ keyword : Hadoop, Analysis on Advertising Effect, TV Advertisement, Social Network Service, Twitter

1. 서 론

SNS 시장이 커져가면서 이를 활용하기 위한 방안에 대한 논의 역시 활발해지고 있다. 대표적인 SNS 매체 중 하나인 트위터는 단문으로 된 의견을 공개하거나 공유하는 특성과 스마트폰의 보급으로 모바일 환경에서의 접근이 가능해지면서, 새로운 정보와 의견을 빠르게 전하는 하나의 매체로 자리잡았다. 이미 트윗트렌드와 같이 트위터를 기반으로한 분석 서비스들이 등장하며 트위터의 인기와 활용도를 알렸다. SNS의 대표적인 특징은 신속성, 개인성과 정보 개방성이다 [1]. 즉, 이러한 특성으로 인해 개개인의 의견이 공개되어 트윗은 제품 • 기업 선호도 및 트렌드 분석에 중요한 자료가 된다. 트윗 각각은 140자의 적은 양이다. 하지만 트윗은 실시간으로 전세계적으로 생성된다는 점에서 그 양이 많고 다루기 어려워 빅데이터라고 할 수 있다. 실제 7주년을 맞아 트위터에서 발표한 자료에 따르면 하루에 전송되는 트윗 수가 4억 건이 넘는다고 발표했다 [2]. 빅데이터는 기존의 방식으로는 다루기 어렵고 처리하는데도 효율이 떨어져, 트윗을 처리하는 데에 기존 방법이 아닌 새로운 방식이 필요해졌다.

한편, 광고는 기업과 소비자를 이어주는 창이다. 그 중에서도 우리가 가장 쉽게 접하는 광고는 바로 TV에서 프로그램 사이사이에 나오는 CF영상들이다. TV는 드라마, 뉴스, 예능 등 다양한 컨텐츠들을 제공하며, 보급률과 신뢰도가 높다. 이와 같은 매체 특성에 의해, TV광고는 많은 시청자들에게 노출될 수 있어 그 효과가 좋으나, 반대로 노출 범위가 넓기 때문에 매체의 특성상 그 반응을 즉각 측정하기 어렵다. 그럼에도 불구하고, 방영 후 반응을

1 Dept. of Computer Science, Sookmyung Women's Univ., Seoul, 140-742, Korea
* Corresponding author (yulan@sm.ac.kr)
[Received 1 August 2013, Reviewed 4 August 2013, Accepted 31 November 2013]
☆ 2013년도 정부(미래창조과학부)재원으로 한국과학창의재단 이공계우수연구(HCR) 지원과제로 수행하였음.

조사하는 일은 매우 중요하다. TV광고는 비용이 크며, 기대효과가 클 뿐만 아니라 광고계의 빠른 흐름을 파악하는 것은 보다 효과적인 광고 제작을 가능케 하기 때문이다. 위와 같은 이유로, 이전부터 광고 효과를 분석하는 다양한 기법들이 시도되었다 [3]. 특히 설문 기법이 주로 쓰이고 있으며 이를 위해 자료를 수집하고, 분석하는 절차를 자동이 아닌 수동으로 처리하고 있다.

본 논문에서는 트위터의 장점을 이용해 광고 효과 분석을 더욱 효율적으로 시도하기 위해 하둡을 이용하여 빠르게 트위터를 분석하는 시스템에 대해 설명한다. 위에서 언급했듯이, 트위터는 사람들의 의견에 대한 많은 데이터를 가지고 있기 때문에, 이 데이터를 광고 효과 분석에 이용할 수 있다. 하지만 트위터의 데이터양은 너무 많아 기존 방법으로는 처리하기 어렵고, 기존 TV광고 분석 기법은 시간과 비용이 상당히 소요된다는 단점이 있다. 이러한 문제와 빅데이터 분석에 걸리는 시간을 효과적으로 단축하기 위해, 본 논문에서는 하둡을 이용하였다. 이전에도 트위터 서비스는 존재해왔으나, 상용화된 기존 트위터 서비스는 트윗 검색, 트위터 내 트렌드 분석 등 트위터 내에 한정된 경우가 많았다. 그러나 본 연구는 트위터 자체가 아닌 트위터와 광고를 결합하여 분석하는데 초점이 맞추어져 있다. 본 연구의 목적은 광고와 트위터 데이터를 수집하며, 수집한 데이터를 토대로 광고의 효과를 트위터 상에의 언급도로 측정하는 것이다. 본 논문에서는 광고와 트위터의 데이터를 수집하고, 트위터에서의 시간별 광고의 언급도를 측정하는 시스템인 LiveAD에 대한 설계와 구축에 대해 기술한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 설명하며, 3장에서는 시스템 설계에 대해, 4장에서는 시스템 구현에 대해 설명한다. 5장에서는 분석 결과를 기술하며, 6장에서는 결론을 맺는다.

2. 관련 연구

광고 효과의 측정이란 광고에 의해 광고목적이 어느 정도 달성되었는가를 판정하기 위해 광고내용이 전달된 후 수용자에게 미친 영향을 과학적인 방법으로 측정하는 것이다[4]. 방송 후, 방송광고에 대한 광고 효과를 측정하는 기존 방식은 정성적 접근과 정량적 접근, 두 가지 방식이 있다. 먼저 정성적 접근방식은 일일 후 회상조사(Day-After Recall, DAR), 테스트 마켓(Test Market), 단일원천 추적조사(Single-source Tracking Test)가 있다 [3]. 이 중에서 일반적으로 쓰이는 방법은 일일 후 회상조사로 버크 테스트(Burke Test)라고도 불린다. 광고가 방영되고 하루가 지난 다음 집행된 프로그램을 시청한 사람들을 대상으로 전화면접을 실시한다. 응답자가 광고를 기억할 경우와 그렇지 않은 경우를 나누어 조사한다. 그러나 버크 테스트는 응답자의 불확실한 기억 또는 설문자의 도움에 의존하기 때문에, 응답 결과가 정확하지 않다는 단점이 존재한다[5]. 또한 조사기간이 오래 걸리고 비용 또한 크다는 문제가 존재한다.

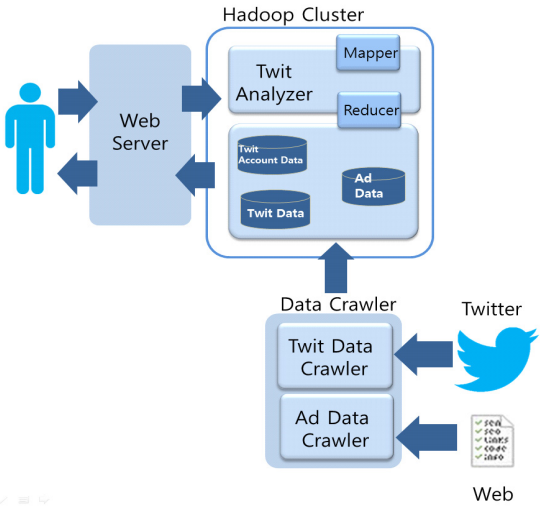
TV 방송 광고의 효과를 측정하는 정량적인 수단은 지표를 이용하는 것이다. 여기에는 시청률에 기반한 CPRP와 GRPs라는 두 가지 지표가 있다. CPRP는 Cost Per Rating Point의 약자로, 1%의 시청률을 얻기 위해 투입된 광고비의 규모를 의미한다. CPRP는 광고비를 시청률로 나눈 값으로 광고의 효율성을 측정하는 지표이다. 한편, GRPs는 Gross Rating Points의 약자로 특정 기간의 시청률의 합이다 [6]. 그러나 두 가지 지표 모두 시청률에 기반하여 측정하므로, 현재의 다매체 다채널 환경에 맞지 않다는 문제가 있다 [7].

한편 트위터를 하둡을 이용하여 분석하는 시스템에 대한 시도가 있었다. 그 연구에서 하둡을 이용하여 트위터 메시지를 분석하는 것을 제안하였으나 시스템의 설계에 그치고 활용효과를 언급하지 않았다 [8].

3. 설 계

본 연구에서 제안한 LiveAD는 그림 1과 같이 하둡을 이용하여 데이터를 수집하여 분산저장 및 분산처리 후 웹을 통해 보여주는 형식으로 구성된다. 전체 시스템은 하둡을 이용하여 광고데이터와 트위터를 매일 자동 수집하는 데이터 크롤러, 특정 광고를 언급하는 트위터를 분석해주는 광고 효과 분석기와 사용자에게 분석 결과를 보여주는 웹 인터페이스로 나눌 수 있다. 각 기능의 자세한 설명은 다음과 같다.

시스템은 크게 데이터 크롤러(Data Crawler), 하둡 클러스터(Hadoop Cluster) 그리고 웹 서버(Web Server)로 구성된다. 데이터 크롤러는 트위터와 TV광고 데이터를 수집하며 자세한 내용은 3.1절에서 설명한다. 하둡 클러스터는 데이터베이스와 트윗 분석기(Twit Analyzer)로 구성된다. 데이터베이스에는 수집된 트위터와 TV광고 데이터가 저장되며, 트윗 분석기는 광고 효과를 맵과 리듀스를 통해 분석한다. 자세한 분석 내용은 3.2절에서 소개한다.



(그림 1) 시스템 구조도
(Figure 1) System Architecture

본 시스템의 구조를 통해 분산 데이터베이스는 크롤러와 분석기가 공유하도록 하여, 데이터를 수집하는 동안, 분석 서비스를 이용할 수 있도록 하였다.

3.1 데이터 크롤링

TV 광고에 대한 의견 분석을 위해서는 TV 광고와 트위터의 두 가지 데이터가 필요하다. 데이터 크롤링에서는 두 가지 데이터를 각각 웹 크롤링을 통해 수집한다. 이러한 웹 크롤링을 통해 필요한 데이터를 자동적으로 업데이트할 수 있다. 먼저 TV 광고에 대한 데이터는 TVCF(tvcf.co.kr)에 공개된 동영상, 이미지, 텍스트 데이터를 이용한다 [9]. 광고 데이터는 표 1과 같이 데이터의 형태에 따라 미디어와 텍스트로 나누고, 텍스트 데이터는 광고 기본 데이터와 상세 데이터로 나뉜다. 광고 기본 데이터는 광고들을 구분 짓기 위해 쓰이고, 광고 상세 데이터는 트위터리안을 대상으로 한 광고의 영향력 측정을 위한 지표로 사용된다. 위 사이트에서는 매일 전날 첫 방영된 광고들에 대한 데이터를 제공한다. 새로 방영된 광고를 추가하여, 광고 데이터의 분류에 알맞은 데이터를 파싱하여 광고 데이터베이스에 저장한다.

한편 트위터 데이터는 트위터 계정과 트윗 데이터를 파싱하여 분산 데이터베이스에 저장한다. 표 2는 트위터에서 수집하는 데이터를 계정과 트윗에 대한 데이터로 나누어 정리한 것이다. 트위터 데이터는 트위터 계정 데이터과 사용자가 올린 트윗, 두 가지로 이루어져 있다. 트위터 계정 데이터를 수집할 때, 사용자의 기본 데이터인 계정URL과 이름이외에도 사용자의 상세 데이터인 사용자의 트윗 수, 팔로워 수, 위치와 언어 데이터를 수집한다. 사용자의 상세 데이터는 광고 효과 분석 시, 사용자의 영향력을 측정하기 위해 사용된다. 트윗 데이터는 트윗 내용, 리트윗과 링크 여부와 함께 리트윗, 관심 글 수를 함께 수집하여 트윗의 파급력을 측정한다. 위에서 언급한 광고와 트위터에 대한 데이터는 자동으로 업데이트 하도록 하여 최신으로 유지시킨다. 또한 위 데이터들은 분산 저장하여 이후 분산 처리를 통해 분석할 수 있도록 한다.

(표 1) 광고 데이터의 분류
(Table 1) Classification of Advertisement Data

형태	분류	데이터명
미디어	동영상	광고 동영상
	이미지	썸네일 이미지
텍스트	광고기본 데이터	광고 이름
		광고 구분코드
		광고 TV온에어 날짜
	광고 상세 데이터	광고대행사, 브랜드회사, 상품명, 모델, 태그, 대, 관련 뉴스 기사

(표 2) 트위터 데이터의 분류
(Table 2) Classification of Twitter Data

형태	분류	데이터명
텍스트	트위터 계정 데이터	계정URL
		사용자ID, Name, 위치, 언어
		팔로워 수, 트윗 수
	트윗 데이터	트윗 내용, 리트윗 및 공유 링크
		리트윗 수, 댓글 ,관심글 수

3.2 광고 효과 분석

TV광고의 효과는 광고 방송 전후의 언급도로 측정할 수 있다. 여기서의 언급은 상품의 이름뿐만이 아니라 모델, 회사, 광고 카피 등 키워드를 포함하여 세도록 한다. TV광고를 언급하는 트윗의 수를 고려하는 것 뿐만이 아니라 리트윗을 이용하여 얼마나 광고에 대한 언급이 퍼져가는지를 측정하도록 한다. 또한 이러한 변화를 시간별로 나타내며 광고가 TV에 방영된 전후의 변화를 측정하

여, 정량적으로 그 수치를 알아내도 한다.즉, TV광고의 효과 분석은 TV광고 방영일을 기준으로 전후의 언급도와 파급력의 변화를 분석하도록 한다.

광고 효과의 측정은 언급도와 파급력의 변화율을 기준으로 한다. 언급도(R)는 각 시간대 별 전체 트윗 중 광고에 관련된 키워드를 언급하고 있는 트윗의 수를 나타낸다. 파급력(S)은 각 시간대 별 전체 트윗 중 광고에 관련된 키워드를 언급하는 트윗을 리트윗한 수로 측정한다. 언급도의 변화율(D_R)은 아래 수식과 같이 광고 전후 언급도의 차를 광고 전 언급도로 나눈 값을 백분율(%)로 표현한 값이다. 파급력의 변화율(RS) 역시 위 방법과 동일하게 계산한다.

$$D_R(\%) = \frac{R_B - R_A}{R_A} \times 100$$

$$D_S(\%) = \frac{S_B - S_A}{S_A} \times 100$$

$$R_A = \sum_{i=0}^{A-1} R_i \, , R_B = \sum_{i=A}^{n} R_i$$

$$S_A = \sum_{i=0}^{A-1} S_i \, , S_B = \sum_{i=A}^{n} S_i$$

R_i = (i번째 시간대의 언급도)
S_i = (i번째 시간대의 파급력)
$i = 0, 1, \ldots, A, A+1, \ldots n$
(A는 광고방영일)

(그림 2) 광고 언급도 및 파급도의 변화율 계산
(Figure 2) The rate of change of ads' reference and spread effect

사용자는 웹을 통해 시스템에 광고 분석 요청을 한다. 이후, 시스템은 광고 데이터베이스에서 상품 이름, 모델, 회사와 같은 일반적인 데이터와 광고에 대한 키워드를 추출한다. 추출한 키워드들을 바탕으로 저장된 트윗 메세지에서 해당 광고를 언급하고 있는 트윗을 걸러낸다. 걸러낸 트윗은 광고가 TV에 방영된 시점을 전후로 시간대별로 정리된다. 각 시간대의 시간당 트윗의 수와 시간당 리트윗 수를 계산하여 광고의 방영 후 언급도와 파급력을 측정하도록 한다. 이렇게 측정된 데이터는 웹 서버로 보내 사용자가 확인할 수 있도록 한다.

4. 구 현

본 시스템의 구현은 광고 및 SNS데이터 프로파일링, 광고 효과의 분석 및 인터페이스 부분으로 나눠진다. 각 부분에 대하여 설명하면 아래와 같다.

4.1 광고 및 SNS데이터 프로파일링

본 시스템에서 수집할 데이터는 광고와 트위터 두 가지가 있다. 두 가지 광고 모두 HTML Parsing을 통해 수집한다. HTML Parser로는 Jericho HTML Parser를 이용한다. Jericho HTML Parser는 open source로 제공되는 Java library로 웹에서 데이터를 추출한다 [10]. 광고와 트위터, 두 가지 데이터 수집의 구현은 아래와 같다.

4.1.1 광고 데이터 수집

TVCF라는 웹 페이지에서 공개하는 광고 데이터를 HTML Parsing을 통해 수집하고 저장한다. TVCF 웹 페이지에서는 아래 그림의 형태로 TV광고의 데이터를 제공한다. 광고마다 광고를 구분하는 광고 코드(code)가 존재하고, 광고 데이터를 제공하는 URL 뒤에 광고 코드 부분에 광고 코드를 입력하면, 해당 광고의 기본 데이터와 상세 데이터를 확인할 수 있다. 화면 좌측에는 광고의 동영상과 제목(소제목 포함), 분류, 방영 시작일과 같은 기본 데이터가 있다. 창의 우측에는 비슷한 TV광고, 국내 경쟁광고, 해외 경쟁광고 및 모델, 대행사와 같은 상세 데이터를 제공한다. 그 하단에는 TVCF에서 분류한 광고의 태그 데이터, 콘티, 관련 뉴스기사에 대한 데이터가 있다. HTML 코드에서 각 부분에 해당하는 데이터를 찾아서 파싱하여 광고 데이터베이스에 저장한다.

(그림 3) TVCF에서 제공하는 광고
(Figure 3) Ad information from TVCF

그림 3은 TVCF에서 수집하는 데이터가 있는 화면이다. 각 부분에 해당하는 HTML Code에서 데이터를 추출하여 광고 데이터베이스에 저장한다. 매일매일 TVCF에는 전날 첫 방영된 TV광고에 대한 정보가 추가된다. 매일 새로 추가된 광고들에 대한 수집을 담당하는 프로그램을 크론 테이블(Cron Table)에 추가시켜 매일 0시마다 수행한다. 이를 통해 매일 광고를 자동으로 수집한다.

4.1.2 트위터 데이터 수집

트위터 데이터베이스는 트위터 계정 데이터와 트윗 데이터로 이루어진 트위터 데이터를 저장한다. 트위터 데이터는 트위터 URL 뒤에 계정명을 붙인 트위터 계정 페이지에서 HTML 파싱을 통해 수집한다.

(그림 4) 트위터 계정 페이지
(Figure 4) Twitter Account Page

트위터 계정 페이지를 살펴보면 위 그림 4와 같다. 그림 4는 트위터 계정 페이지 좌측의 로그인 창 등의 메뉴를 제외하고 우측의 계정에 대한 기본 데이터 및 상세 데이터, 트윗을 포함하는 부분만을 캡쳐한 것이다. 트윗 계정 페이지의 우측 상단에는 계정데이터가 있다. 계정 데이터는 계정의 프로필 사진, 이름, 계정명과 같은 계정 기본 데이터와 트윗 수, 팔로워 수, 팔로잉 수와 같은 계정 상세 데이터를 포함한다. 계정 데이터는 HTML 코드에서 module profile-card component profile-header profile-page-header 클래스에 담겨 있다. 트윗 계정 페이지의 우측 하단에는 계정의 사용자가 올린 트윗들이 시간 순서대로 나타난다. 각 트윗은 content 클래스에 담겨있다.

트위터 계정 페이지에서 트윗의 내용과 추가적인 트위터 계정을 추출하면서 트윗과 트위터 계정을 수집하는 방식은 아래와 같다.

```
1.  Store Seed Twitter Accounts into Account Table
2.    For each Twitter Accounts{
3.       Read HTML Code from the twitter account
      // Collect Twits
4.       Parse Twits from HTML Code
5.       For each Twit Data{
6.          Save Parsed Twit Data into Twit Table
      }

      // Add new Twit Accounts
7.       Parse  Twit Comments from HTML Code
8.       Extract Twit Accounts that Commented
9.       If  (there aren`t the Twit Accounts  in Twit
Accounts Table)
         Save the Twit Accounts into Twit Account Table}
```

먼저 분산 저장된 트위터 계정 테이블에서 트위터 계정 데이터를 불러온다(1). 각 트위터 계정에 대해, 트위터 계정 주소에 해당하는 웹 페이지 HTML코드를 읽어온다(2). 트위터 계정 주소란 트위터 주소(twitter.com/) 뒷 자리에 계정 명을 추가한 것이다. 각 트위터 계정 주소에서 받아온 HTML에는 최대 20개의 트윗이 저장되어 있다. 또한 각 트윗에는 다른 사용자들의 댓글이 저장되어 있다. 해당 페이지에서 HTML 코드를 가져와 분석할 준비를 한다(3). 먼저 트윗 데이터베이스에 저장할 트윗을 모은다. 트윗부분을 포함하는 content 클래스를 파싱하여(4) 각 트윗 별로 트윗 URL, 트윗 내용, 트윗을 올린 계정, 포함된 링크, 트윗 게시 시각, 언어, 리트윗 수 그리고 관심글 수를 트윗 테이블에 저장한다(5). 여기에서 트윗 URL은 트위터 계정 주소 아래에 트윗마다 가지고 있는 고유의 ID를 추가한 URL이다. 언어 정보는 한국어와 영어를 쓰는 사용자를 대상으로 하기 위해 수집한다(6).

다음으로 트위터 계정 테이블에 새로이 추가될 새로운 트위터 계정들을 추가한다. 새로운 트위터 계정들은 트윗에 댓글을 단 계정들이다. 이 계정들은 현재 트위터 계정과 교류를 쌓고, 활동중인 계정으로 추정할 수 있다. 각 트윗 페이지는 댓글을 단 트위터 계정과 그 내용을 공개하고 있다. HTML 코드를 파싱하여(7) 트위터 계정을 추출한다(8). 이렇게 얻은 트위터 계정들이 기존의 데이터베이스에 존재하지 않는 계정일 경우(9), 이러한 계정들

을 트위터 계정 테이블에 저장한다(10). 이렇게 새로운 계정을 추가하고, 전 과정을 반복하여, 기존계정에서 추가된 트윗과 새로운 계정의 트윗 모두를 계속적으로 수집한다.

4.1.3 SNS데이터 저장소

본 연구에서는 수집한 트위터 데이터를 분산으로 저장하여, 광고 효과 분석과 같은 분산 처리 작업을 수행하기 위해 HBase를 활용하여 두 개의 테이블을정의하였다. 하나는 트위터 계정(Twitter Accounts) 테이블이고, 또 다른 하나는 트윗(Twits) 테이블이다. 트위터 계정 테이블에는 트위터 계정의 기본 및 상세 데이터가 저장된다. 트윗 테이블에는 트위터 기본 및 상세 데이터가 저장된다 (표2 참조).

4.2 광고 효과의 분석

사용자는 웹을 통해 분석을 요청하여, 분석하고자 하는 광고에 대한 트위터에서의 언급도와파급력을 하둡의 맵리듀스(MapReduce)를 통해 빠르게 분석할 수 있다.

선택한 광고 효과의 분석은 광고의 키워드를 추출, 필터링, 맵리듀스를 통해 시간대별 언급도 및 파급력을 계산하여 이루어진다. 먼저 광고 데이터베이스에서 모델, 브랜드, 품목, 광고 제목 등을 가져온다. 가져온 데이터들이 광고의 키워드가 된다. 한편, 트윗 테이블에서 해당 광고의 키워드를 언급하는 트윗만을 필터링한다. 필터링은 HBase의 라이브러리에서 제공하는 여러 필터를 조합하여 키워드를 하나라도 포함하는 트윗만을 추출한다. 필터링 과정은 맵(Map)과정에서 키워드 포함 유무 판별을 하지 않도록 하여, 분석 시간을 크게 줄이게 된다. 이렇게 추출된 트윗들은 맵리듀스 단계로 넘어가게 된다.

먼저 언급도를 계산하는 맵리듀스 프로그램을 살펴보면 다음과 같다. 맵리듀스 단계에 들어가는 데이터들은 모두 광고의 키워드를 언급한 트윗이다. 맵(Map) 단계에서는 이러한 트윗들이 만들어진 시간대를 key로 하고, value를 개수로 하고, 리듀스(Reduce) 단계에서 value를 합하여, 시간대별 언급도를 계산하게 된다.

파급력을 계산하는 맵리듀스 프로그램은 아래와 같다. 맵(Map) 단계에서는 이러한 트윗들이 만들어진 시간대를 key로 하고, value를 해당 트윗의 리트윗 수로 하여, 리듀스(Reduce) 단계에서 value를 합하여, 시간대별 파급력을 계산하게 된다.

위에서 계산한 자료는 4.2절에서 언급한 언급도(R)와 파급력(S)을 의미하며, 4.2절에서 언급한 수식에 따라 언급도의 변화율(DR)과 파급력의 변화율(DS)을 계산한다

4.3 사용자 인터페이스

사용자는 웹을 통해 분석을 요청하고, 언급도와 파급력 중심의 분석 결과를 확인한다. 웹 페이지는 광고 효과 분석을 요청하는 'Create Report' 메뉴와 분석 결과를 확인하는 'Browse Report' 메뉴로 구성된다.

(그림 5) 광고 효과 분석 화면

(Figure 5) Screenshot of analysis on a TV Advertising effect

웹 페이지를 통해, 사용자는 분석하고자 하는 광고를 입력하고, 광고 정보와 트위터 기반의 분석 결과를 얻을 수 있다. 'Create Report' 메뉴에서는 광고 효과 분석을 요청하는 작업을 수행한다. 검색창에 광고 관련 키워드를 입력하여 광고를 검색한 후, 원하는 광고를 선택하면 요청이 완료된다. 검색 시, 사용자가 입력한 키워드는 광고의 제목과 상품명과 같은 광고 기본 데이터와 광고 모델과 대행사와 같은 상세 데이터에서 검색하여 분석하고자 하는 광고를 찾게 된다. 시스템에서는 사용자의 요청에 따라 해당 광고의 광고 효과를 맵리듀스를 통해 분석한다.

분석이 완료된 경우, 'Browse Rep ort'메뉴에서 원하는 광고의 기본 및 상세 정보와 분석 결과를 볼 수 있다. 원하는 광고를 선택하면 광고의 광고의 제목, 분류, 광고 방영 날짜, 광고 영상과 같은 기본정보와 광고 모델, 광고 관련 태그와 뉴스와 같은 상세정보와 함께 광고 효과 분석 결과를 그래프형태로 보여준다. 광고 분석 결과는 광고 상영일을 기준으로 전후 3일동안 시간대에 따른 언급도와 파급력에 대한 그래프를 보여준다. 그래프는 HighChart에서 제공하는 라이브러리를 사용하여 구현하였다 [11].

5. 실험 및 결과

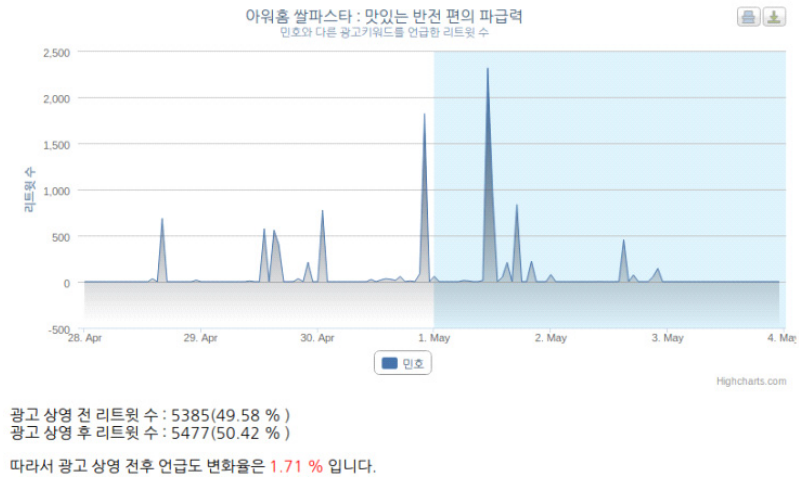
본 연구에서 제안하고 있는 LiveAD 시스템의 구현은 하둡이 설치된 2대의 컴퓨터에서 이루어 진다. 1대의 서버는 하둡의 Master server가 되고, 나머지 1대의 서버는 하둡의 Slave server가 된다. 각 서버의 사양은 아래와 같다. 운영체제는 Ubuntu 12.04 Server를 이용하였다. 설치된 하둡의 버전은 hadoop-1.0.4이며, HBase의버전은 0.94.2이다.

본 시스템은 학교 내 연구실에 설치되어 2013년 3월부터 2013년 5월까지, 광고 텍스트 데이터 약 4MB와 트위터의 텍스트 데이터 약 6GB를 축적했다. 본 시스템의 주요 기능은 광고 및 트위터 데이터를 수집하는 것과 사용자의 요구에 따라 그 효과를 분석하는 기능이다. 특히, 본 시스템의 목적은 축적된 광고・트위터 데이터를 빠르게 분석 처리하는 데 있다.

트위터 데이터의 분석은 2013년 5월 1일 TV에 첫 방영된 '아워홈 쌀 파스타: 맛있는 반전'광고를 대상으로 실시하여, 이 광고의 언급도와 파급력을 조사하였다. 위 광고에서 광고 이름과 모델, 상품명, 브랜드에서 추출한 키워드는 '민호, 아워홈, 샤이니, 맛있는반전, 아워홈 쌀파스타, 쌀파스타'와 같다. 추출한 키워드가 광고를 언급한다고 가정하여 위 키워드를 언급하는 트윗 메세지를 맵리듀스를 이용하여 분석하였다. 광고 효과 분석에는 언급도와 파급력을 계산하는 2개의 잡(Job)이 동시에 수행되었다. 먼저 데이터는 두 개의 잡은 트위터 데이터를 대상으로 하여, 키워드를 포함하는 트윗만을 맵리듀스 작업을 수행하도록 하는 필터링 작업을 거쳤다. 필터링 작업 결과 74,906 바이트로 처리할 데이터를 줄였다. 한편 언급도를 계산하는 잡은 1분 18초, 파급력을 계산하는 잡은 1분 7초의 시간이 소요되어, 결과적으로 분석의 총 수행시간은 두 개의 잡 중 긴 작업시간인 1분 18초이다.

광고 분석 결과 언급도의 변화는 18.48%로 크지 않으나 언급도가 증가하였고, 파급력의 변화는 1.71%로 거의 변하지 않았다. 두 그래프의 모형으로 보아, 광고 후, 일시적인 큰 관심을 받았지만 이내 소강된 것으로 풀이된다.

(그림 6) 아워홈 광고의 언급도의 변화를 그래프로 표현한 화면
(Figure 6) Referring rate graph of the TV Ad, 'Our Home'



(그림 7) 아워홈 광고의 파급력의 변화를 그래프로 표현한 화면
(Figure 7) Spread effect graph of the TV Ad, 'Our Home'

6. 결 론

본 연구에서는 분산 데이터베이스인 HBase를 이용하여 TV광고와 트위터를 매일 자동으로 수집하며, 대표적인 분산처리 플랫폼인 하둡을 이용하여 수집한 TV광고와 트위터 데이터를 이용하여 SNS에서의 TV광고의 영향력을 실시간에 가깝게 빠르게 측정하는 광고효과 분석 시스템인 LiveAD 시스템을 개발하였다.

이 시스템을 평가하기 위해서는 속도와 정확성이라는 두 가지의 지표가 있다. 속도의 면에서는 시스템의 크기, 데이터의 양에 따라 다양한 차이가 존재하지만 데이터 양 대비 처리 시간 등으로 측정할 수 있을 것이다. 정확성의 면에서는 광고의 효과가 얼마나 잘 측정되었는지로 알 수 있다. 특히 본 연구에서 개발한 광고 효과 분석기는 광고의 영향력을 수치와 그래프로 측정하는 데에 초점이 맞춰졌다. 향후 연구 내용으로 수치뿐만이 아니라 광고에 대한 긍정 혹은 부정에 대한 의견, 이에 따른 구매효과 등의 의견을 분석할 수 있다면, 광고 효과 분석의 진정한 목적을 달성할 수 있을 것이다. 위에서 설명한 트위터리안의 의견 분석을 통해 단순히 광고 효과의 측정을 넘어 어떻게 광고를 발전시킬지도 분석할 수 있는 툴로 발전할 수 있을 것으로 기대된다.

참 고 문 헌(Reference)

[1] Jin-Hyung Lee, "Trend and Expantion of SNS", http://www.kca.kr/open_content/bbs.do?act=file&bcd=radiotrends&msg_no=10462&file_no=1].

[2] ZDNet Korea, "7-years-Twitter …Everydy-Twits 400 million", http://www.zdnet.co.kr/news/news_view.asp?artice_id=20130321205519

[3] You-Jae Lee,"Measurement of Advertising Effect", http://youjae.com/data/cdata3/20/ad13.pdf

[4] Naver encyclopedia, "Measuring Advertising Effect", http://terms.naver.com/entry.nhn?cid=512&docId=51147&mobile&categoryId=512

[5] Joseph Raymond Roy, "Burke Test", Marketing Metrics Made Simple, http://www.marketing-metrics-made-simple.com/burke-test.html

[6] AC Nielsen Media research, "Analysis of Public TV Advertising Effect and its Factor", KOBACO, 2002.04, http://www.kobaco.co.kr/information/adinfo/UploadFile/0204_132.pdf

[7] Jun-Cheon Jang, "CPRP based on Viewer Ratings has Limits to Measurement of Engagement of Viewers", KOBACO, 2009.01, http://www.kobaco.co.kr/information/adinfo/html/200901/054_%BA%D2%C8%B2%B1%E2TV.htm

[8] Ji-Hoon Song, Si-Jin Lee, Dong-Hyo Park, "Twitter message analysis system design using Hadoop", A proceedings of Korean Society for Internet Information 2012 Summer Conference, p.169-170, 2012.6.

[9] TVCF, a Web site which provide TV advertisement information, http://www.tvcf.co.kr/.

[10] Jericho HTML Parser, http://jericho.htmlparser.net/docs/index.html

[11] HighChart, http://www.highcharts.com/

◐ 저 자 소 개 ◑

허 서 연

2009년~현재 숙명여자대학 컴퓨터과학과 재학
관심분야 : 클라우드 컴퓨팅, 보안
E-mail : cloverkid90@gmail.com

김 윤 희

1991년 숙명여자대학고 전산학과 졸업(학사)
1996년 시라큐스대학교 대학원 컴퓨터과학과 졸업(석사)
2000년 시라큐스대학교 대학원 컴퓨터과학과 졸업(박사)
2001년 로체스터공대 컴퓨터공학과 조교수
2001년~현재 숙명여자대학교 컴퓨터과학과 교수
관심분야 : 문제 풀이 환경(PSE), 과학 워크플로우 관리, 클라우드 컴퓨팅
E-mail : yulan@sm.ac.kr