Google Analytics API를 연동한 R 프로그래밍 데이터 시각화

안장근 · 장시웅

동의대학교

Data Visualization of R Programming using Google Analytics API

Jang-Keun Ahn* · Si-Woong Jang*

*DONG-EUI University

E-mail : jangkeun@hanamil.net, swjang@deu.ac.kr

요 약

최근 IoT 기술발달로 인한 스마트폰 및 대용량 미디어기기 사용증가로 인터넷 네트워크 사용량이 폭발적으로 증가되고 있고, 이러한 데이터 사용량 급증으로 대량의 데이터를 지칭하는 빅데이터 수집 및 분석에 많은 기업과 정부가 주목하고 있다. 빅데이터는 기존에 없던 새로운 데이터의 구축이 아니며, 그동안 축적된 다방면의 방대한 데이터의 집합이라 할 수 있다. 빅데이터의 이용 및 분석에 대한 기업·정부·학계의 수요는 증가하고 있지만, 고난도의 빅데이터 분석을 위한 인프라 구축이 선결과제이어서, 이러한 인프라구축 비용 때문에 빅데이터 분석이 일선 산업분야에 바로 적용하는데 많은 장애요인이 되어 데이터 분석가들의 빅데이터 분석에 애로사항으로 존재하고 있다.

이러한 어려움을 해소하기 위한 방안으로 새로운 인프라 구축 없이 Google Analytics API를 연동한 R 프로그래밍의 데이터 시각화를 활용한 데이터 분석 방안을 제시하고자 한다.

본 연구에서는 구글 애널리틱스 API를 연동하여 사용자 웹사이트의 사용자접속, 사이트운영, 이벤트 발생 등의 데이터를 R 프로그램을 활용하여 사이트 현황을 데이터 시각화로 분석하고 운영중인 웹사이트에 적용 가능한 콘텐츠 개발 방안에 대해 연구하였다.

키워드

애널리틱스, 빅데이터, R프로그램, 데이터 시각화

Ⅰ. 서 론

최근 ICT기술의 발전 및 트위터, 페이스북, 유트브 등 쇼셜 미디어 사용이 증가하면서, 기존의 PC 사용 위주에서 이동과 휴대가 가능한 스마트 기기로 인터넷 사용이 증가함에 따라 비정형 데이터의 사용과 데이터의 축적이 급격이 증가하고 있으며 이를 통해 축적된 방대한 양의 빅데이터를 활용하기 위한 관심이 뜨겁다. 현대 정보화 사회의 모든 분야에서 빅데이터를 수집하고 분석하여 분석한 데이터를 시각화 기법을 동원한 최적의 의사결정을 도출하는 작업이 핵심 이슈로 부각되고 있다. 그만큼 빅데이터 활용 및 분석의 중요도는 현대사회의 IT 트렌드를 넘어 정치, 경제, 사회, 문화, 과학 등 다방면의 문제점을 해결하는 수단으로 부상되면서 새로운 가치와 수익을 창출할 것으로 기대가 크지만, 빅데이터의 활용 및 분석을 담당할 전문 인력이 아직까지는 부족하고 빅데이터의 분석 결과를 쉽게 이해할 수 있도록 시각적 수단을 활용하여 나타내는 방법은 아직 보편화되지 않고 있으며, 현재까지 많은 연구가 진행되고 있는 상황이다. 본 논문에서는 데이터 분석을 위한 시각화 방안을 제시하기 위해 가상의 웹사이트를 구축하고, 웹사이트 사용자접속 및 이용 현황 등의 웹로그 데이터를 구글 애널리틱스 API 연동을 통해 웹로그 데이터의 분석항목을 추출하여 빅데이터 분석프로그램인 R 프로그램을 활용하여 데이터 분석 및 시각화 방안을 구현한다.

Ⅱ. 기존연구

빅테이터 처리 및 분석 기술은 분석 인프라 기술 및 분석기법으로 나눌 수 있다.

빅데이터 분석 인프라 기술로는 표 1.와 같이 Hadoop, NoSql, Map-reduce, R 프로그램, 구글 BigQuery 등 있으며, 빅데이터 수집 및 분석 인프라 시장의 지배적 기술은 아직 존재하지 않는다. 따라서 빅데이터 분석 시장의 절대적인 강자가 되기 위해 인프라 기반 기술의 연구는 더욱 활발히 진행될 것이다[1-5].

표 1. 빅데이터 분석 인프라 기술

분석 인프라 기술	내용
Hadoop	· 정형/비정형 빅데이터 분석 · 하둡 분산 파일 시스템인 HDFS (Gadoop Distributed File System)을 활용한 분산 자원 관리
NoSql	· Not-Only Sql, No SQL · 대표적인 솔류션 Cassandara, Hbase, MongDB · 스키마 고정 되어 있지 않음 수평적 확장 용이
Map-reduce	· 분산 대용량 데이터 처리 프레임 워크 · Map, Reduce라는 간단하고 추상화된 기본 연산으로 병렬 처리 지원
R 프로그램	· 통계계산 및 데이터 분석 시각화 강점 · Java, C, Python 등 타 프로그래밍 언어와 연결 용이
구글 BigQuery	· 페타 바이트급의 데이타 저장 및 분석을 위한 크라우드 서비스 · 기존 SQL 언어 사용

빅데이터 분석 기법으로는 표 2.와 같이 Text Mining, Opinion Mining, Social Network Analytics, Cluster Analysis, WebLog Analytics 등이 있으며, 빅데이터 분석 인프라기술을 활용한 다양한 빅데이터 분석의 혁신적인 분석기법이 연구되고 있으며, 이를 검증하기 위한 연구도 활발히 진행되고 있다[1-4].

표 2. 빅데이터 분석 기법

분석 기법	내용
Text Mining	· 정형·비정형 텍스트 데이터에서 자연어 처리 · 유용한 정보 추출 및 가공
Opinion Mining	· 정형·비정형 텍스트의 긍정, 부정, 중립의 선호도 판별 · 시장규모 예측, 소비자 반응, 입소문 분석에 활용
Social Network Analytics	· 소셜 네트워크의 연결구조 및 연결강도를 바탕으로 영향력 측정 · 입소문의 중심이나 허브역할의 사용자 검색에 활용
Cluster Analysis	· 특성을 가진 개체를 군집형태로 분석 · 관심사, 취미에 따른 군집 사용자 분석 활용
WebLog Analytics	· 웹서비스의 방문, 이벤트 발생 등의 웹로그 분석 · 온라인 쇼핑 및 웹사이트 분석 및 관리 등 활용

III. R 프로그래밍 데이터 시각화

3.1 웹로그 분석

본 논문에서는 구글 애널리틱스 웹로그 데이터의 추출 및 분석을 위해 웹로그 분석 기법을 활용하였다. 웹로그 분석은 웹사이트 방문자 웹 사이트 접속시 생성되는 로그파일(Log file)을 분석하는 것이며, 웹서비스 방문경로, 유입경로, 지역적 위치, 검색 키워드 등 해당 웹 사이트에 방문하는 순간부터 이탈 후 행동에 대한 데이터 분석을 의미한다. 구글에서 제공하는 구글 애널리틱스 웹로그 분석도구는 웹사이트 방문자의 사이트 활동사항, 유입경로, 잠재고객의 성향분석과 타켓팅에 특화된 광고상품을 연동할 경우 방문자의 연령이나, 성별, 관심분야 등 최신 분석 자료도 제공한다. 다음 그림 1은 구글 애널리틱스의 로그분석 화면이다.

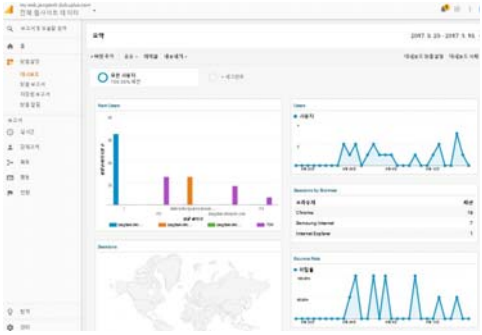

그림 1. 구글애널리틱스

3.2 구글 애널리틱스 아키텍처

웹로그 분석을 위한 데이터베이스 구축은 구글 개발자 콘솔(Google Develoers console)에 사용자 등록 후 분석 대상 웹서버의 URL을 구글 애널리틱스에서 제공하는 configuration 과정을 진행하면, 구글 애널리틱스 서버는 실질적인 트래킹이 가능한 트래킹 코드를 그림 2와 같이 생성한다. 그림3은 트래킹 코드를 분석 대상 웹사이트 소스에 추가하면 일반유저가 분석 대상 웹서버에 접속시 부터 유저의 분석 데이터는 유저의 브라우저에 HTML문서를 제공하면서 추가한 트래킹 코드를 함께 전달한다. 유저의 브라우저는 이 트래킹 코드를 다운로드 받은 다음 구글 애널리틱스 서버쪽에 ga.js파일을 전달받아 구글 애널리틱스 서버에 사용자 정보데이터를 구글애널리틱스 서버에 웹비콘 이미지(UTM.gif)를 파싱한 후 웹로그 데이터베이스에 사용자분석 데이터를 저장하고, 사용자분석 데이터 정보를 기반으로 분석 보고서를 제공한다.

```
<!-- Global Site Tag (gtag.js) - Google Analytics -->
<script async src="https://www.googletagmanager.com/gtag/js?id=UA-105383898-1"></script>
<script>
  window.dataLayer = window.dataLayer || [];
  function gtag(){dataLayer.push(arguments);}
  gtag('js', new Date());

  gtag('config', 'UA-00000000-1');
</script>
```

그림 2. 웹사이트 추적코드

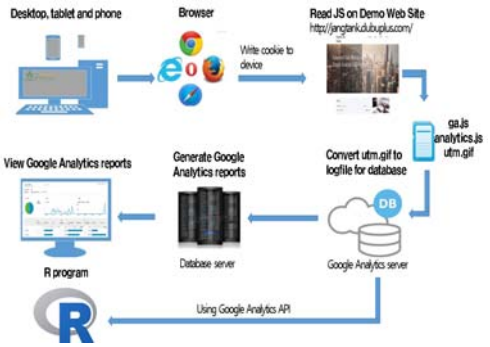

그림 3. 구글 애널리틱스 아키텍처

3.3 R 프로그램을 구글애널리틱스 API 연동

소프트웨어 구성은 R Studio를 통해 코딩하였고 통계분석을 강화하기 위해 구글애널리틱스 API를 연동하였다. R 프로그램에서 구글 애널리틱스에 데이터를 추출해오는 라이브러리는 RGoogleAnalytics, RGA, ganalytics, GAR 등이 있으며, 본 논문에서는 RGA 라이브러리를 사용하여 분석하였다. 그림 4와 같이 RGA라이브러리는 (A Google Analytics API Client for R) 구글 애널리틱스 API를 사용할 수 있도록 구성된 R 프로그램 패키지이며, 주요기능은 구글 개발자 콘솔 인증지원, 구글 애널리틱스 API 기능을 연동하여 사용할 수 있도록 구성된 라이브러리 패키지다. RGA 라이브러리 호출이 완료되면 구글 사용자 인증 및 구글애널리틱스 API 사용을 위해 구글 개발자 콘솔(Google Develoers console)을 활용하여 Analytics API를 활성화한 후 구글 클라우드 프로젝트 인증인 Client.id, Client.secret, Google Analytics ids를 발급받아 R 프로그램 소스에 삽입 후 사용한다. authorize() 함수는 RGA 라이브러리가 구글 애널리틱스 Data에 접근할 수 있도록 권한을 부여하기 위한 함수이며, list_profile() 함수로 접근할 사이트 ID를 검색하여 데이터를 추출한다. 그림5와 같이 API 연동 확인 및 데이터 검증을 위해 ga_profile에 구글 애널리틱스 사용자 ID 및 사용자 설정정보를 추출한 데이터를 확인할 수 있다.

```
## RGA 라이브러리 설치 및 호출
If( " RGA " %in% installed.packages() == FALSE)
  install.packages("RGA")
Library(RGA)

## 인증서 저장 위치 설정
Setwd( " C:/data/work " )

# GA 계정 정보 등록
client.id <- "245367242808 9a7t0gsf8
              nljopk.apps.googleusercontent.com"
client.secret <- "QuTOpEBWhXDI8dzAnfD"
ga_token <- authorize(client.id = client.id,
                      client.secret = client.secret)

# 토큰 정보 저장
save(ga_token, file="./ga_token")
ga_profile <- list_profiles(token = ga_token)
load("ga_token")
source("ga_conf.R")
```

그림 4. 구글 애널리틱스 API 연동 Sorce Code

	id	accountId	webPropertyId	internalWebPropertyId	name	currency	timezone	websiteUrl
1	158856637	105383898	UA-105383898-1	157359921	my ac	KRW	Asia/Seoul	http://jangtank.dubuplus.com
2	158873979	105383898	UA-105383898-1	157359921	전체 웹사이트 데이터	KRW	Asia/Seoul	http://jangtank.dubuplus.com
3	158904440	105383898	UA-105383898-1	157359921	보고서 대한민국	USD	Asia/Seoul	http://jangtank.dubuplus.com

그림 5. ga_profile 데이터

3.4 웹로그 데이터 분석 및 시각화

웹사이트 추적코드를 삽입한 데모사이트에 사용자가 접속하면, 사용자 추적 정보가 구글 애널리틱스 서버에 저장된다. 구굴 애널리틱스 서버는 웹사이트 트래픽 추적 분석을 넘어, 수많은 데이터 소스를 엮어 하나의 데이터 분석 플랫폼을 만들 수 있다. 표 3과 같이 구글 애널리틱스 API 및 GA를 활용하면 웹 추적코드에서 발생한 데이터를 통합해 하나의 보고서로 만들 수 있고, 구글 애드워즈, 애드센스, 웹마스터 도구(Search console), YouTube, 이메일 뿐 아니라, 구글 설문지 온라인 데이터도 통합이 가능하며 구글 외의 제품도 데이터를 통합할 수 있다.

표 3. 구글 애널리틱스 API 내용

Google Analytics API	내용
Management API	계정, 웹 속성 및 세그먼트에 대한 구성 데이터 연동
Core Reporting API	맞춤 보고서를 생성을 위한 기능 연결
Multi-Channel Funnels Reporting API	사용자의 목표 및 트래픽 소스 경로 등 연결
Real Time Reporting API	실시간 리포트 생성
Metadata API	API 측정 기준 및 측정 항목 등 메타정보 연결

그림 6과 같이 구글 애널리틱스 API을 연동하여 데모사이트에 접속한 로그데이터를 추출한다. 추출데이터 항목은 접속일자, 접속기기, 접속자수, Sessions, 페이지뷰 이다. 추적코드 및 로그데이터는 구굴 애널리틱스 서버에 데이터베이스로 저장되어 있으며 R 프로그램을 이용하여 관련 데

이터베이스를 Query을 통해 추출할 수가 있다[6-8].

구글 애널리틱스 API 사용을 위해서는 구글 애널리틱스에서 제공하는 BigQuery 항목 중 디멘전(Dimensions)과 메트릭스(Metrics)를 이용하여야 한다. 디멘전은 다차원 데이터에서 심층 비즈니스 분석이 가능하도록 데이터를 구성하는 기준 정보 구조를 의미한다. 즉 데이터 분석가 입장에서의 데이터분석을 진행하기 위한 여러가지 과정 정보라고 할 수 있다. 매트릭스란 데이터분석을 진행하고자 하는 속성들에 대한 측정 가능한 값을 나타내는 것이며, 디멘전의 특성에 대한 측정할 수 있는 수치화 표현 값을 말한다[9].

추출데이터의 쿼리항목은 변수 profileId에 구글에서 발급받은 API ids 값을 할당하고, 추출할 데이터의 시작일, 종료일을 지정한다. 측정값인 메트리스에는 사용자, Sessions, 페이지뷰 항목을 기제하고, 측정항목인 디멘전에는 접속 날자 및 접속기기정보를 추출하도록 하였다

추출된 데이터의 시각화 표현을 위해 접속기기 정보를 바탕으로 날자별 접속현황을 시각화 하였다.

```
if("ggplot2" %in% installed.packages() == FALSE) install.packages("ggplot2")
library(ggplot2)
if("dplyr" %in% installed.packages() == FALSE) install.packages("dplyr")
library(dplyr)

ga.df <- get_ga(profileId = id,
                start.date = "2017-09-10", end.date = "2017-09-21",
                metrics = c("ga:users", "ga:sessions"," ga:pageviews"),
                dimensions = c("ga:date","ga:devicecategory"),
                sort = "ga:date", filters = NULL,
                segment = NULL, samplingLevel = NULL, start.index = NULL,
                max.results = 10000, include.empty.rows = NULL,
                fetch.by = NULL, ga_token)

ggplot(data = ga.df, mapping = aes(x = date,
       y = sessions, fill=devicecategory, group = devicecategory,
       colour = devicecategory)) +
  geom_bar(stat="identity") + scale_fill_hue(l=80) +
  geom_line() + geom_point(size=3, colour="#CC0000") +
  facet_wrap(~ devicecategory) +
  labs(title = '사용자 접속 현황', x = '웹사이트 접속수',
       y = '접속일자') +
  theme_bw()
```

그림 6. 데이터 추출 및 시각화 Sorce Code

그림 7. 추출 데이터의 시각화 표현

그림 7은 데모사이트 접속현황을 데스크톱 및 모바일 부분으로 나누어 시각화로 표현하였다. R 프로그램을 사용하여 막대그래프 및 선 그래프 등 많은 시각화 방법으로 데이터들의 가독성을 높일 수 있다[5-7].

Ⅳ. 결 론

본 논문에서는 빅데이터 분석 프로그램인 R 프로그램과 웹로그 분석도구로 활발히 사용되고 있는 구글 애널리틱스 API 연동을 통해 추출데이터 분석 및 시각화 방안을 구현하였다.

본 논문의 결과로 특정 웹서비스의 로그데이터를 추출하여 R 프로그램을 활용한 다양한 분석이 가능할 것이며, R 프로그램의 강력한 분석기법을 활용한다면 빅데이터 분석의 최종목표인 미래의 합리적이고 최적의 의사결정에 도움을 줄 수 있을 것이다.

참고문헌

[1] J.Manyika, M.Chui, B.Brown, J.Bughin, R.Bobbs, C.Roxburgh, and A.Byers, “Big Data The Next Frontier for Innovation, Competition, and Productivity”, Techical Report, McKinsey Global Institue, p6~p8, 2011.

[2] Philip R, “Big Data Analytics”, TDWI Best Practices Report, p1~p35, 2011.

[3] Nodar Montselidze, Alex Kuksin, “Hadoop Integrating with Oracle Data Warehouse and Data Mining”, Journal of Technical Science and Technologies, p21-25, 2014.

[4] 이후영, “웹 애플리케이션 기반의 빅데이터 분석 시스템 구현에 관한 연구”, 공주대학교 대학원 멀티미디어공학과 석사학위논문, p7~p20, 2017.

[5] 이은경, “R을 이용한 빅데이터 분석 : 데이터의 다차원 처리 및 시각화”, 이화여자대학교 대학원 석사학위논문, p7~p20, 2014.

[6] 김희주, “하둡에서 데이터접근 제어 설계 및 구현”, 강원대학교 대학원 이학석사학위논문, p8~p30, 2014.

[7] 박준형, “빅데이터 처리를 위한 R 병렬 패키지에 관한 연구”, 한남대학교 대학원 컴퓨터공학과 석사학위논문, p10~p14, 2017.

[8] 박용민, “R을 활용한 대용량 데이터의 처리 및 병렬 컴퓨팅에 대한 연구” p4~p10, 2013.

[9] Jordan Tigani, Siddartha Naidu, “Google BigQuery Analytics”, 에이콘, p230~240, 2016.