creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

[UCI]I804:21010-200000016993

공학석사 학위논문

Google Analytics API를 이용한 빅데이터 구축 및 시각화

지도교수 장시웅

이 논문을 공학석사 학위논문으로 제출함

2018년 2월

동의대학교 대학원

IT융합학과

안 장 근

공학석사 학위논문

공학석사 학위논문

Google Analytics API를 이용한 빅데이터 구축 및 시각화

지도교수 장시웅

이 논문을 공학석사 학위논문으로 제출함

2017년 12월

동의대학교 대학원

IT융합학과

안 장 근

공학석사 학위논문

dCollection @ dau

안장근의 공학석사 학위논문을 인준함

주 심 안 귀 임 인

위 원 장 시 웅 인

위 원 정 덕 길 인

2017년 12월

dCollection @ dau

Google Analytics API를 이용한 빅데이터 구축 및 시각화

안 장 근

동의대학교 대학원 IT융합학과

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 위해서는 비정형 및 정형 빅테이터의 수집을 위한 IoT 센서 디바이스, 수집된 데이터를 저장할수 있는 서버장비, 빅데이터 수집분석을 위한 어플리케이션 소프트웨어 등 인프라 구축이 선결과제이어서, 이러한 인프라구축 비용 때문에 빅데이터 분석이 일선 산업분야에 바로 적용하는데 많은 장애요인이 되어 산업 및 학계에서 빅데이터 수집 및 분석진행에 애로사항으로 존재한다

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 운영 중인 웹사이트에 적용 가능한 콘텐츠 개발 방안에 대해 연구한다.

dCollection @ dau

목 차

Ⅰ. 서 론 ····· 1

Ⅱ. 관련 연구 ····· 2
2.1 빅데이터 수집 및 분석 기술 ····· 2
2.1.1 빅데이터 처리 인프라 기술 ····· 3
2.1.2 빅데이터 분석 기법 ····· 5
2.2 빅데이터 수집 및 분석 기술의 국내·외 개발 현황 ····· 6
2.2.1 빅데이터 수집 및 분석 기술의 국내 개발 현황 ····· 6
2.2.2 빅데이터 수집 및 분석 기술의 국외 개발 현황 ····· 8
2.3 국내 빅데이터 분석 활성화를 위한 과제 ····· 10
2.4 기존 연구의 차별성 ····· 12

Ⅲ. Google Analytics API를 연동한 R 프로그래밍 데이터 시각화 시스템 설계 ····· 13
3.1 빅데이터 수집 및 분석 시스템 구성 ····· 13
3.1.1 구글 애널리틱스 아키텍처 설정 ····· 14
3.1.2 웹 로그 데이터 수집 사이트 구성 ····· 16
3.1.3 웹로그 수집 및 구글 애널리틱스 연동 ····· 17
3.2 구글 애널리틱스 API 연동 R 프로그래밍 ····· 17
3.2.1 R 프로그램 ····· 17
3.2.2 R 프로그램과 구글 애널리틱스 API 연동 알고리즘 ····· 18
3.2.3 R 프로그램을 구글 애널리틱스 API 연동 ····· 19
3.2.4 구글 애널리틱스 API 사용 ····· 20
3.2.5 웹로그 데이터 분석 ····· 21

dCollection @ dau

3.2.6 Google Analytics 디멘션과 메트릭스 ····· 22
3.2.7 Google Analytics API 연동 R 프로그래밍 ····· 23
3.2.8 R 프로그램 데이터 분석 시각화 프로그래밍 ····· 24

Ⅳ. 실험 및 결과 ····· 26
4.1 구글 애널리틱스 API 연동 R 프로그래밍 ····· 26
4.2 구글 애널리틱스 API 연동 데이터 추출 ····· 27
4.3 R 프로그램밍 시각화 ····· 29

Ⅴ. 결론 ····· 30

참고문헌 ····· 31

dCollection @ dau

그림 목차

<그림 1> 빅데이터의 수집 특성
<그림 2> 빅데이터의 분석기술의 활용서비스
<그림 3> 빅데이터의 수집 및 분석 개요도
<그림 4> 구글 애널리틱스 계정 생성
<그림 5> 구글 애널리틱스 생성 트래킹 코드
<그림 6> 분석대상 웹사이트에 트래킹 코드소스 삽입
<그림 7> 웹로그 수집 웹사이트
<그림 8> 구글 애널리틱스 연동 확인
<그림 9> 구글 애널리틱스 API 연동 알고리즘
<그림 10> 구글 애널리틱스 API 연동 Sorce Code
<그림 11> ga_profile 데이터
<그림 12> 구글 애널리틱스 데이터 추출
<그림 13> 구글 애널리틱스 추출 데이터
<그림 14> 구글 애널리틱스 디멘전, 메트릭스
<그림 15> 구글 애널리틱스 Overview
<그림 16> 구글 애널리틱스 데이터 결과 값
<그림 17> 웹사이트 각 페이지별 방문자 현황 R 프로그램 소스
<그림 18> 웹사이트 각 페이지별 방문자 현황 결과 값
<그림 19> 시각화 프로그래밍
<그림 20> 웹로그 데이터 수집 및 분석
<그림 21> 구글 애널리틱스 API 연동 R 프로그래밍
<그림 22> 구글 애널리틱스 GA 계정 데이터
<그림 23> 구글 애널리틱스 API 연동 데이터 추출 R 프로그래밍
<그림 24> 웹사이트 사용기기별 접속 현황
<그림 25> 방문페이지별 접속 현황
<그림 26> 시각화 프로그래밍
<그림 27> 사용자 접속현황 시각화 도표

dCollection @ dau

표 목차

<표 1> 빅데이터 분석 인프라 기술
<표 2> 빅데이터 분석 기법
<표 3> 국내 빅데이터 수집 및 분석 기술 현황
<표 4> 국외 빅데이터 수집 및 분석 기술 현황
<표 5> 산업별 빅데이터 활용 선도과제
<표 6> 구글 애널리틱스 API

dCollection @ dau

Ⅰ. 서 론

최근 ICT기술의 발전 및 트위터, 페이스북, 유튜브 등 소셜 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[1]. 그만큼 빅데이터 활용 및 분석의 중요도는 현대사회의 정치문제, 경제문제, 사회문제 등 다방면의 주요 이슈를 해결하는 수단으로 부상되면서 새로운 가치와 수익을 창출할 것으로 기대가 크지만, 빅데이터의 활용 및 분석을 담당할 전문 인력이 아직까지는 부족하고 빅데이터의 분석 결과를 사용자들이 쉽게 이해할 수 있도록 시각적으로 표현하는 시각화 전달 기능이 아직 보편화 되지 않고 있으며, 현재까지 많은 연구가 진행되고 있는 상황이다[2].

해외 선진국에서는 빅데이터의 수집, 분석, 처리 기술과 연동될 수 있도록 Hadoop, Nosql, Map-reduce, R 프로그램 등 다양한 인프라 기술들이 연구되고 있으며, 분석기법으로는 Text Mining, Opinion Mining, Socail Network Analytics, Cluster Analysis, WebLog Analytics 등이 연구되고 있다.

국내 IT 인프라는 세계적 수준이지만, 빅데이터 기반기술력은 아직까지 취약한 실정이다. 국내 빅데이터 수집 및 분석 산업의 효율적인 발전을 위해 H/W, S/W 관련 인프라 지원이 시급하며, 빅데이터 수집 및 분석을 위한 핵심 원천기술 연구가 필요하다.

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.

dCollection @ dau

Ⅱ. 관련 연구

2.1 빅데이터 수집 및 분석 기술

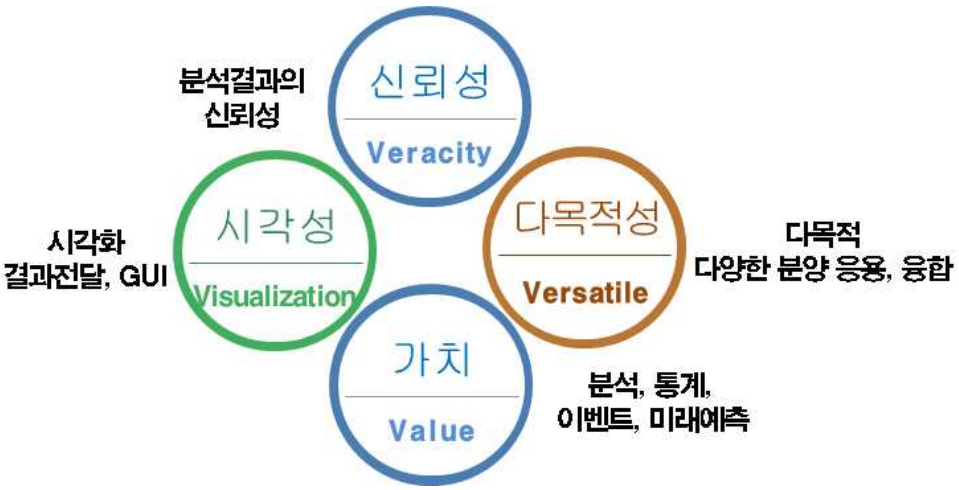
최근 전 세계적으로 정보통신기술 분야의 급속한 발달로 인하여 기존 정형화된 데이터 범위를 넘은 공간정보, 수치지도, 위치정보, 사용자 사용정보, 개인활동 내역 등 다양한 형태의 데이터들이 인터넷이나 소셜미디어를 통해 수집되고, 데이터의 크기도 과거와는 비교될 수 없을 정도로 방대해 지고 있다. 이러한 데이터들의 수집 및 축적도 중요하지만 이러한 빅데이터를 어떻게 심층 분석과 해석을 통해 산업분야에서 활용하는 것이 중요한 이슈로 부각되었다. 빅데이터 수집 및 분석 기술은 대용량의 데이터를 수집 및 심층 분석을 통해 해석 가능한 정보를 추출하고 이러한 분석을 바탕으로 미래 예측과 향후 변화에 능동적으로 대응하기 위한 정보화 분석 처리기술을 의미 한다.

<그림 1>과 같이 빅데이터의 수집 특성은 거대한 크기(Volume), 다양한 형태(Variety), 빠른 속도(Velocity)의 속성을 가지고 있어 3V라고 특성을 정의하고 있다[1].

<그림 1> 빅데이터의 수집 특성

dCollection @ dau

<그림 2>와 같이 빅데이터의 분석기술은 활용 서비스 측면에서 Veracity (신뢰성)의 확보 즉 수집된 빅데이터를 이용한 분석결과를 활용한 분석 신뢰성을 의미하며, Visualization(시각성)은 분석결과를 어떻게 이용자들에게 쉽게 판독할 수 있도록 전달하는 시각화 표현이며, Versatile(다목적성)은 분석한 결과를 다양한 분야에 활용할 수 있는 것을 말하며, Value(가치)는 수집되는 데이터에 새로운 가치를 부여하는 주요한 특성으로 예측된 분석, 가설 등을 말한다[2].



<그림 2> 빅데이터의 분석기술의 활용서비스

2.1.1 빅데이터 처리 인프라 기술

세계 각국의 국가기관 및 기업에서 빅데이터 수집 기술의 확보를 통해 기존 운영되고 있는 산업뿐만 아니라 미래 4차 산업분야 차별화된 경쟁력 향상을 위하여 지속적으로 많은 연구를 진행하고 있으며, 현재 구축된 빅데이터 수집 및 분석 기술은 분석 인프라기술과 분석기법으로 나눌 수 있다. 빅데이터 분석 인프라 기술로는 <표 1>과 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

dCollection @ dau

것이다[3].

<표 1> 빅데이터 분석 인프라 기술

분석 인프라 기술	내 용
Hadoop	· 정형/비정형 빅데이터 수집 분석 솔루션 · Hadoop Distributed File System을 활용한 분산 자원 관리 · 분산컬럼 기반의 HBase, MapReduce를 연계한 분산 컴퓨팅 지원 프레임워크
NoSql	· Not-Only Sql, No SQL · 대표적인 솔루션 Cassandara, Hbase, MongDB · 테이블 스키마(Table Schema) 고정되어 있지 않음 · 테이블간 조인연산을 지원하지 않아 수평적 확장이 용이함
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 · 통계기법, 데이터 분석 모델링, 최신 데이터 마이닝까지 구현 가능 · Java, C, Python 등 타 프로그래밍 언어와 연결 용이 · 분석 기법의 확장이 용이하여 기업에서 주로 사용
구글 BigQuery	· 페타 바이트급의 데이타 저장 및 분석을 위한 크라우드 서비스 · 기존 SQL 언어 사용 · 구글, Youtube 등에서 수집된 방대한 빅데이터의 분석 가능

dCollection @ dau

2.1.2 빅데이터 분석 기법

기존에 연구된 빅데이터 분석 기법으로는 <표 2>와 같이 Text Mining, Social Network Analytics, Opinion Mining, WebLog Analytics, Cluster Analysis 등이 있고, 빅데이터 분석 인프라기술을 활용한 다양한 빅데이터 분석의 혁신적인 분석기법이 연구되고 있으며, 이를 검증하기 위한 연구도 활발히 진행되고 있다[4-5].

<표 2> 빅데이터 분석 기법

분석 기법	내 용
Text Mining	· 정형·비정형 텍스트 데이터에서 유용한 정보 추출을 위하여 자연어 처리 기술 적용 가능 · 방대한 텍스트에서 의미 있는 정보 추출 · 추출된 정보와의 연계성 파악 용이하며 단순한 정보 검색 그 이상의 결과를 추출할 수 있음
Opinion Mining	· 소셜미디어의 긍정, 중립, 부정 등의 정형·비정형 텍스트 선호도 판별 · 특정 상품 및 서비스에 대한 소비자 반응, 시장규모 예측, 입소문 분석에 활용
Social Network Analytics	· 소셜미디어 및 네트워크의 연결강도 및 연결구조를 기반으로 해당 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 · 웹사이트의 실시간 사용자, 방문 형태별, 사용자 이탈률 등 데이터 분석 용이

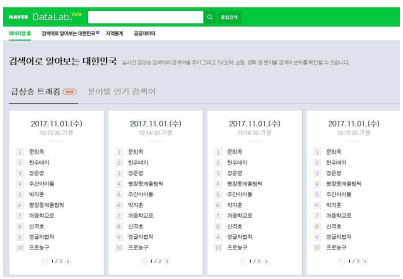
dCollection @ dau

2.2 빅데이터 수집 및 분석 기술의 국내·외 개발 현황

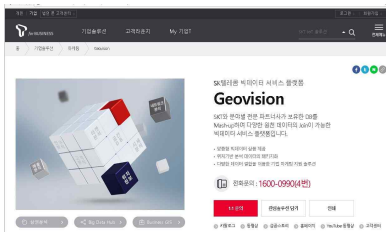
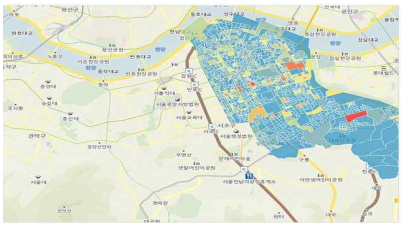
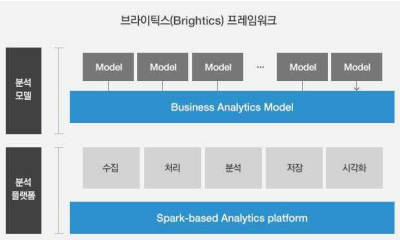
2.2.1 빅데이터 기술의 국내 개발 현황

국내 빅데이터 수집 및 분석 기술은 해외기업과 비교하면 연구와 사업화진행이 부족한 실정이다. <표 3>포털사이트인 네이버와 다음, SK, KT, 삼성 SDS 등 스마트기기가 발생하는 로그 데이터를 하둡을 활용해 처리하는 기술이 보편화 되어 있고, KT NexR(넥스알)은 클라우드 기술 및 하둡 기술을 기반으로 다양한 빅데이터 사업을 추진하고 있으며 국내 연구소와 공동으로 국내 오픈 소스 커뮤니티를 지원하면서 하둡, R프로그램, 메라듀스 등의 오픈소를 기반으로 적극적인 연구를 진행하고 있다. 또한 국가적으로 한국정보화진흥원에서 "빅데이터센터(http://kbig.kr)"을 운영하여 국내 빅데이터 수집 및 분석 활성화를 위한 다양한 정책을 지원하여 빅데이터 생태계 조성 및 산업 육성 기반 확대에 힘쓰고 있다[6-8].

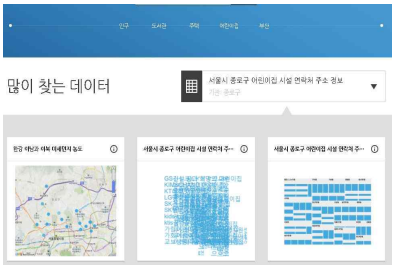
<표 3> 국내 빅데이터 수집 및 분석 기술 현황

회사	내 용
네이버	· 포털사이트 네이버에 축적된 로그 데이터를 활용한 빅데이터 분석 (http://datalab.naver.com) · Naver Analytics API와 활용하여 네이버에 축적된 로그데이터 연동가능 · 국내 공공데이터포털 연동을 통한 사용자 편의성 증대 · 관련 자료 

dCollection @ dau

다음 소프트	· 자연어 처리와 텍스트 마이닝 분석 기술을 바탕으로 빅데이터 분석 (http://www.socialmetrics.co.kr) · 블로그, 트위터 문서 분석 모니터링 제공 · Social Network Analytics 분석 강점 · 관련 자료
SK 텔레콤	· 국내 최대 이동통신 사업자로 보유 데이터의 강점을 활용한 유동인구, 지리정보, 소비업종, 상품판매 현황을 종합적으로 분석하는 지오비전(Geovision) 개발 · 위치기반 분석 데이터 및 기업 마케팅 지원솔루션 특화 · 관련 자료  
삼성 SDS	· 제조, 물류분야 효율적인 운영을 위한 데이터분석 솔루션 “브라이틱스(Brightics) 출시 · 실시간 생산시설분석, 물류리스크 모니터링 등 생산성 향상 지원 · 관련 자료 

dCollection @ dau

신한 카드	· 고객의 카드사용 현황 빅데이터 분석을 바탕으로 소비패턴 등 생활방식 도출하는 “코드나인” 출시 · “코드나인”을 활용한 세부적인 맞춤형 카드를 앞세워 포화 상태에 이른 국내 카드시장 활로 모색 · 관련 자료
빅데이터 센터	· 지속적으로 발전 가능한 빅데이터 생태계 조성 및 산업 육성 기반 확대 (http://kbig.kr) · 빅데이터 분석 테스트베드 제공 등 사업화 지원 인프라 지원 · 관련 자료 

<출처 각 회사 홈페이지>

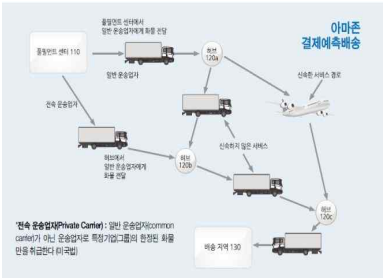
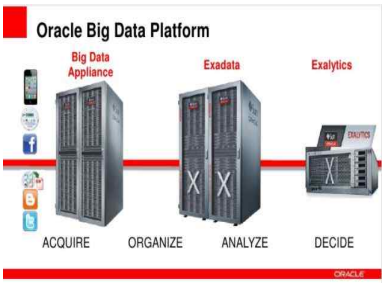
2.2.2 빅데이터 수집 및 분석 기술의 국외 개발 현황

빅데이터 분석이라는 이슈가 설립되기 이전부터 빅데이터 분석에 가까운 개념의 서비스를 가장 먼저 제공했던 회사인 아마존과 데이터베이스의 강자인 오라클 등 빅데이터 분석의 국외 대표 주자들은 지속적으로 노하우를 축적하고 고도화하면서 높은 성과를 달성하고 있는 가운데, 이들 기업 뿐 아니라 다양한 분야의 기업들이 빅데이터 수집 및 분석을 위한 기술을 개발하고 보

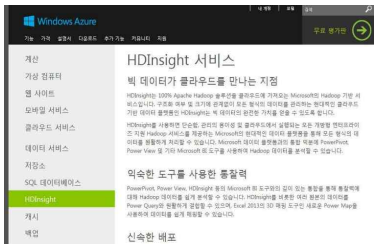
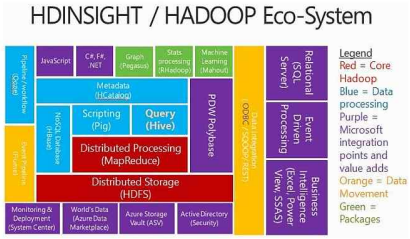
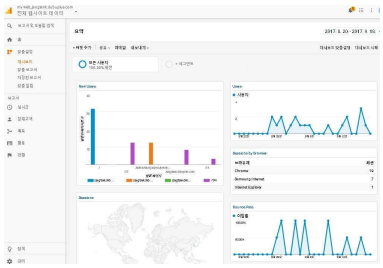
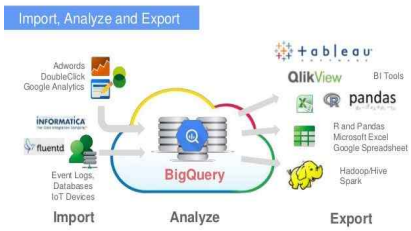
dCollection @ dau

급하고 있으며, 빅데이터 분석을 활용하여 일선 산업분야의 연계와 미래 산업에서 발생될 수익의 정확한 예측을 통해 빅데이터 수집과 분석 시장의 우위를 선점하고 있고. 국외 개발현황은 <표 4>와 같다[9-12].

<표 4> 국외 빅데이터 수집 및 분석 기술 현황

회사	내 용
아마존	· 아마존 예측 배송 · 빅데이터 분석을 이용한 고객 구매 이용 패턴을 분석하여 구매 유도 · 고객 수요 예측을 통한 주문 전 상품 배송 시스템 · 예측되는 장소에 선 배송 서비스 · 관련 자료 
오라클	· 하이페이온을 인수하여 오라클 빅데이터 어플라이언스 출시 (Oracle Big Data Appliance) · 기존 오라클 제품 및 하둡, NoSql, MySql, 오라클 빅데이터 커넥터 등을 포함하여 빅데이터 분석 솔루션 제공 · 관련 자료 

dCollection @ dau

마이크로소프트	· 하둡 기술 전문 업체인 호튼웍스(Hortonworks)와 협력하여 윈도우 기반 HDInsight 서비스 출시 · 클러스터 확장, .NET, Java와 같은 다양한 개발 환경 제공 · 관련 자료  
구글	· Google BigQuery Analytics 시스템 구축 · Google, Google 애드워즈, Youtube 등에서 취득된 다양한 형태의 빅데이터 분석 솔루션 제공 · Google Analytics API, Google 애드워즈 등을 활용한 로그 데이터 활용에 용의 · 관련 자료  

<출처 각 회사 홈페이지>

2.3 국내 빅데이터 분석 활성화를 위한 과제

국내외 시장분석기관들은 앞으로 빅데이터가 주도할 거대 시장규모를 예측하여 향후 전망에 대하여 발표하고 있으며, 대중매체에서는 급성장하는 빅데이터 분석 현황 및 국내외 발전전망을 발표하고 국내의 미흡한 연구 성과에

dCollection @ dau

대한 개선책과 대중의 적극적인 관심과 투자의 필요성을 강조하고 있다.
또한 정부에서는 “정부 3.0” 추진 기본계획을 바탕으로 국내 빅데이터 수집, 분석, 활용 산업의 활성화를 위한 <표 5>와 같은 6대 주요산업별 선도과제를 선정하여 지원하고 있다[13-15].

<표 5> 산업별 빅데이터 활용 선도과제

분 야	내 용
의료 및 건강	· ICT 힐링 플랫폼 구축을 통한 개인건강정보의 축적 및 의료기관 등과 공유 및 활용
과학기술	· BT, ICT, NT 등 R&D 성과물을 기반으로 실시간 과학기술 빅데이터 공유 플랫폼 구축
정보보안	· 빅데이터 분석을 통한 정보보안 강화, 해킹 및 보안사고 등의 대비를 위한 대응시스템 구축
제조 및 공정	· 제품 품질향상을 위해 중소기업 및 대기업의 공정, 납품, 운송 등 연동
소비 및 거래	· 소비 트렌드 예측, 판매 시뮬레이션 등 구매패턴 포트폴리오 구성 지원 및 트렌드 분석 지원
교통 및 물류	· 교통, 물류 수요예측 및 제어 시스템 도입을 통한 유통체계 최적화

국내 빅데이터 수집 및 분석 관련 정부지원 및 연구는 지속되고 있지만 해외시장처럼 시장을 주도할만한 성과는 도출되지 않고 있으며, 국내시장의 성과는 아직까지 없는 실정이다. 빅데이터 산업은 앞으로 다가올 4차 산업혁명의 기초적인 기반산업이며, 육성해야 할 산업이다. 현재 국내의 빅데이터 도입에 대한 사회적인 인식 부족 및 관련법규의 모호한 규제인 개인정보법 제약조건, 부족한 전문 인력 등이 빅데이터 수집 및 분석 산업의 활성화에 걸림돌로 작용하고 있지만, 향후 국내 빅데이터 산업의 발전을 위해 각 주체별로

dCollection @ dau

적절한 역할 분담이 체계적으로 이루어져야 할 것이다. 정부와 국내 산업체는 성공사례 전파를 통한 빅데이터 분석 기업의 사업 참여를 유도하고, 개인정보 관련법 정비, 빅데이터 분석 전문 인력 양성에 지속적인 노력을 기울여야 한다. 산업 업계는 빅데이터 산업 분야의 R&D 개발을 통한 고도화된 기술사업화 노력이 수반되어야 하며, 금융업계는 적절한 금융 상품 지원을 통하여 아직 시장초기에 있는 빅데이터 활용산업의 생태계 조성을 뒷받침하여야 할 것이다[16-18].

2.4 기존 연구의 차별성

본 논문에서 연구된 빅데이터 수집 및 분석과 관련된 연구로는 빅데이터 수집 기술과 Google Analytics 연동 기술, 분석 시각화 기술로 구분할 수 있다. 일선 산업체에서 초기 인프라 투자비용 없이 빅데이터 수집 및 분석을 바로 적용할 수 있도록 빅데이터 수집 기술은 Google Analytics를 활용하였고, 빅데이터의 분석을 위한 R 프로그램을 활용하여 빅테이터를 분석하고 분석된 빅데이터의 사용자 가독성을 위한 시각화 표현 기법을 제시한다[19-21].

기존 빅데이터 수집과 분석을 위해 많은 인프라 구축비용이 필요하지만 Google Analytics API 및 R 프로그램을 활용하면 추가 비용 없이 빅데이터 수집 및 분석이 가능하다[22-24].

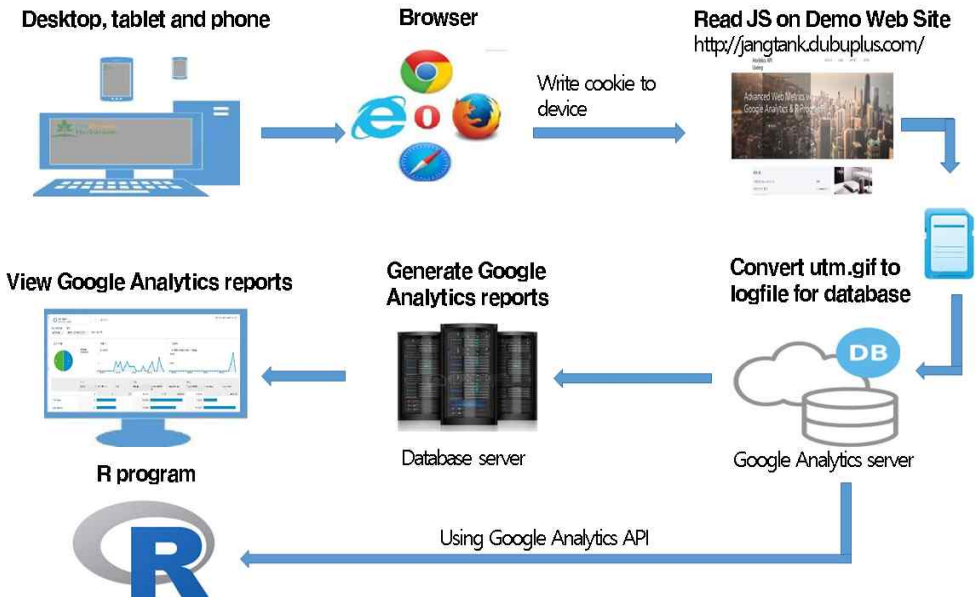
이에 따라 일선 산업체에 바로 적용할 수 있는 빅데이터 수집 및 시각화 분석을 위한 기능을 구현하였다.

dCollection @ dau

Ⅲ. Google Analytics API를 연동한 R 프로그래밍 데이터 시각화 시스템 설계

본 장에서는 웹로그 빅데이터를 수집하기 위해 웹로그 데이터 수집용 웹사이트 구축하였고, 웹로그 데이터의 수집을 위한 Google Analytics 환경 설정, 빅데이터 분석을 위한 R 프로그램과 Google Analytics API 연동, Google BigQuery, 사용자 가독성 향상을 위하여 데이터 시각화 분석 관련된 알고리즘과 프로그래밍에 대해 기술한다.

3.1 빅데이터 수집 및 분석 시스템 구성



<그림 3> 빅데이터의 수집 및 분석 개요도

본 논문에서는 구글 애널리틱스 웹로그 데이터의 추출 및 분석을 위해 웹로그 분석 기법을 활용하였다. 웹로그 분석은 웹사이트 방문자의 웹 사이트 접

dCollection @ dau

속시 생성되는 로그파일(Log file)을 분석하는 것이며, 웹서비스 방문경로, 지역적 위치, 유입경로, 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 타겟팅에 특화된 상품을 연동할 경우 사이트 방문자의 연령이나, 성별, 주요 관심분야 등 데이터를 웹 로그데이터를 수집하기 위해 <그림 3> 과 같이 데모사이트를 구축하여 구글 애널리틱스 데이터베이스에 저장할 수 있도록 구글 애널리틱스 서비스에서 트래킹 코드를 생성하여 웹사이트에 트래킹 코드 소스를 삽입하여 구글 애널리틱스 서버와 연동하였다.

3.1.1 구글 애널리틱스 아키텍처 설정

웹로그 분석을 위한 데이터베이스 구축은 구글 애널리틱스 계정이 있어야 가능하며 <그림 4> 와 같이 구글 애널리틱스 사이트에 접속하여 계정을 생성해야 한다.

<그림 4> 구글 애널리틱스 계정 생성

구글 개발자 콘솔(Google Develoers console)에 사용자 등록 후 분석 대상 웹서버의 URL을 구글 애널리틱스에서 제공하는 configuration 과정을 진행

dCollection @ dau

하면, 구글 애널리틱스 서버는 실질적인 트래킹이 가능한 트래킹 코드를 <그림 5>와 같이 생성한다. 웹로그 분석 대상의 데이터 수집을 위한 트래킹 코드를 분석 대상 웹사이트 소스에 <그림 6> 과 같이 추가하면 일반유저가 분석 대상 웹사이트(http://jangtank.dubuplus.com) 접속시부터 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 구글 애널리틱스 서버에 웹비콘 이미지(UTM.gif)를 파싱한 후 구글 애널리틱스 서버에 실시간 웹로그 데이터를 저장하게 된다.

<그림 5> 구글 애널리틱스 생성 트래킹 코드

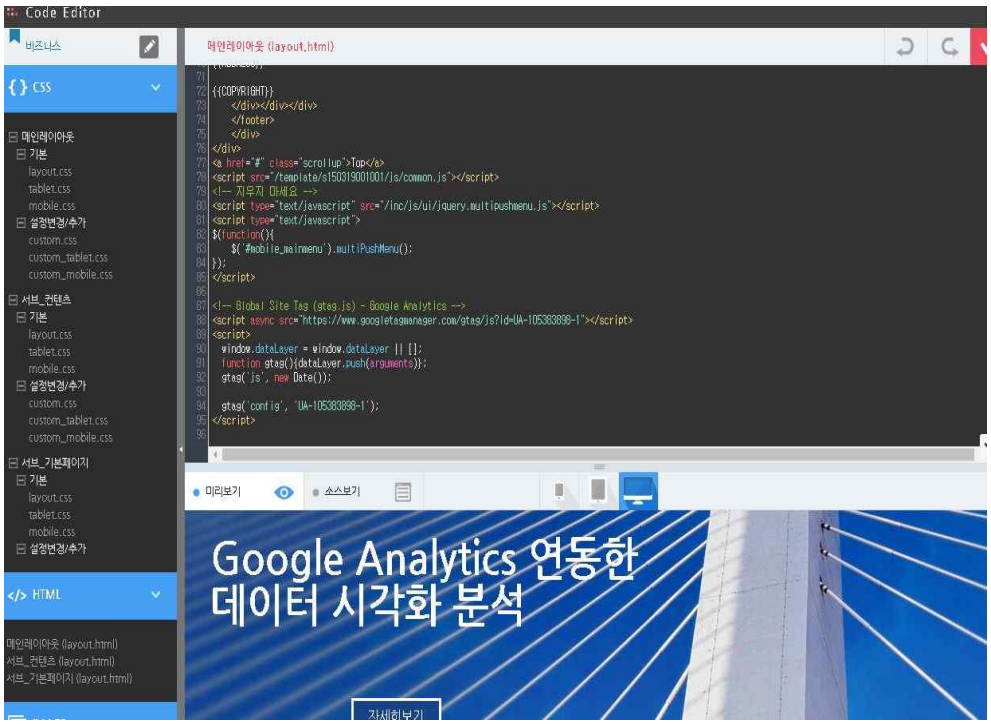
<그림 6> 분석대상 웹사이트에 트래킹 코드소스 삽입

dCollection @ dau

3.1.2 웹 로그 데이터 수집 사이트 구성

웹로그 수집을 위해서는 웹사이트가 필요하다. 웹사이트 구축을 위해 국내 무료 웹사이트 구축 서비스인 두부플러스를 이용하였다. 두부플러스 웹 서비스를 활용하여 스마트폰, 테블릿, PC에서 접속할 수 있는 웹사이트(http://jangtank.dubuplus.com)를 구축하였다. <그림 7>와 같이 기본 도메인은 jangtank.dubuplus.com이며, 웹사이트 소스는 PHP를 사용하였고, DB는 Mysql로 구축하였다.

구글 애널리틱스 API을 연동하기 위해 메인프레임 소스에 트래킹 코드 소스를 삽입하여 웹로그 데이터의 실시간 연동이 가능하도록 구축하였고, 웹사이트 각 메뉴별 구글 애널리틱스 데이터베이스 실시간 연동을 위해 메인프레임에 삽입된 트래킹 코드를 재사용 하여 범용성을 증대하였다.

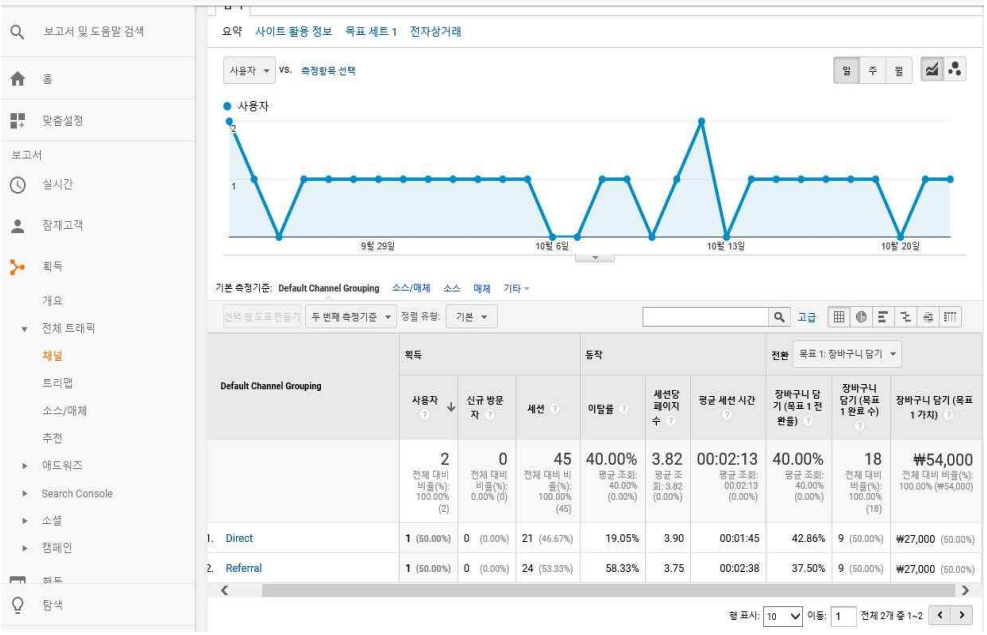


<그림 7> 웹로그 수집 웹사이트

dCollection @ dau

3.1.3 웹로그 수집 및 구글 애널리틱스 연동

웹로그 수집을 위한 사이트와 구글 애널리틱스 데이터베이스 연동 확인을 위해 웹사이트(http://jangtank.dubuplus.com)에 사용자가 접속을 하면 실시간으로 구글 애널리틱스 데이터베이스에 <그림 8> 과 같이 자료가 축적된다.



<그림 8> 구글 애널리틱스 연동 확인

3.2 구글 애널리틱스 API 연동 R 프로그래밍

3.2.1 R 프로그램

R 프로그램은 무료로 사용할 수 있는 전세계적으로 가장 많이 활용되는 통계 분석 소프트웨어 플랫폼이며, 최근에는 정형·비정형 빅데이터 분석, 머신러닝 등에 많이 활용이 되고 있다. 또한 빅데이터 시각화 분석을 위한 패키지 지원이 강력하여 다양한 데이터 시각화 기능을 제공하기도 한다.

본 장에서는 구글 애널리틱스에 수집된 웹로그 분석 데이터를 R 프로그램 라이브러리를 사용하여 Google Analytics API 연동 후 빅데이터 분석 및 시각화 방안을 구현하였다. 구글 애널리틱스의 기본적인 분석 리포트로도 어느

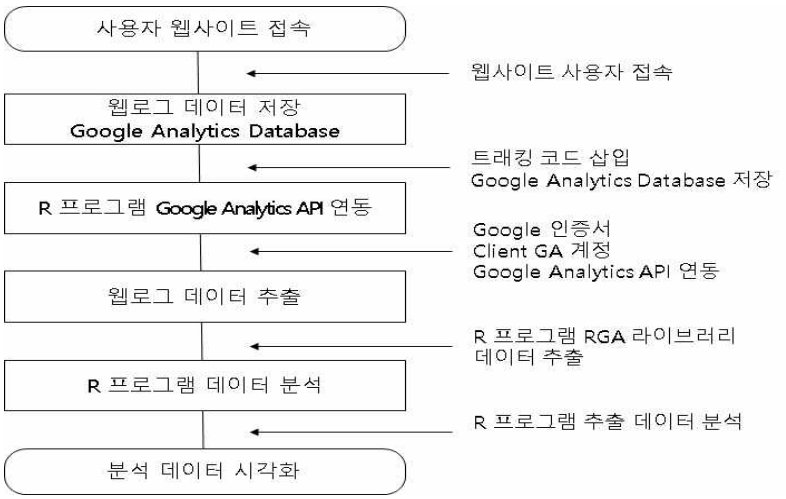
dCollection @ dau

정도의 분석이 가능하지만, 구글 애널리틱스 서비스는 수집된 데이터의 상관관계 분석, 회귀분석 등 미래예측을 위한 통계적인 분석 기능이 제한적이다.

구글 애널리틱스 API 연동 R 프로그래밍을 통해 구글 애널리틱스 데이터베이스에 저장된 데이터를 불러와서 빅데이터 분석을 진행할 수 있으면, 추가적인 통계적 분석을 진행할 수 있다.

3.2.2 구글 애널리틱스 API 연동 R 프로그래밍 알고리즘

본 논문에서 웹로그 데이터 수집 및 분석은 구글 애널리틱스 데이터 베이스 및 R 프로그램을 사용한다. <그림 9>와 같이 사용자의 웹사이트 접속 정보는 웹로그 수집 웹사이트에서 수집되며 이 수집된 웹로그 데이터는 구글 애널리틱스 데이터베이스에 실시간으로 저장된다. 웹로그 데이터의 수집을 위해서는 웹사이트 소스에 구글 애널리틱스 트래킹 코드를 삽입하여 사용하고, R 프로그램을 활용한 분석을 위해서는 구글 애널리틱스 Client GA 계정을 이용하여 구글 애널리틱스 API을 연동하여야 한다. R 프로그램에서 구글 애널리틱스 API을 이용하기 위한 RGA 라이브러리를 사용하였다.



<그림 9> 구글 애널리틱스 API 연동 알고리즘

dCollection @ dau

3.2.3 구글 애널리틱스 API 연동 R 프로그래밍

소프트웨어 구성은 R Studio를 통해 개발하였고 웹로그 분석을 강화하기 위해 구글 애널리틱스 API를 연동하였다. R 프로그램에서 구글 애널리틱스에 데이터를 추출해오는 라이브러리는 RGoogleAnalytics, RGA, ganalytics, GAR 등이 있으며, 본 논문에서는 RGA 라이브러리를 사용하여 개발하였다. <그림 10>과 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 구글 애널리틱스 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.

```
## 1. system config
##==========================================================================================

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

## 인증서 저장 위치 설정
setwd("C:/data/work")

# GA 계정인 등록
client.id <- "245367242808-t6ch0ngee26nahpi79a7t0gsf8nljopk.apps.googleusercontent.com"
client.secret <- "QuTOpEBWhXDI8dW8a1DzAnfD"
ga_token <- authorize(client.id = client.id, client.secret = client.secret)

# 토큰 정보 저자
save(ga_token, file="./ga_token")
ga_profile <- list_profiles(token = ga_token)
load("ga_token")
validateToken(token)
```

<그림 10> 구글 애널리틱스 API 연동 Sorce Code

dCollection @ dau

<그림 11>와 같이 API 연동 확인 및 데이터 검증을 위해 ga_profile에 저장된 구글 애널리틱스 사용자 ID 및 사용자 설정정보를 추출한 데이터를 확인할 수 있다.

Filter

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

<그림 11> ga_profile 데이터

3.2.4 구글 애널리틱스 API 사용

웹사이트 추적코드를 삽입한 데모사이트에 사용자가 접속하면, 사용자 추적정보가 구글 애널리틱스 서버에 저장된다. 구글 애널리틱스 서버는 웹사이트 트래픽 추적 분석 및 수많은 웹로그 데이터 소스를 통합하여 하나의 빅데이터 분석 플랫폼을 구축할 수 있다. <표 6>과 같이 구글 애널리틱스 API 및 GA를 활용하면 웹 추적코드에서 발생한 데이터를 통합해 하나의 보고서로 만들 수 있고, 구글 애드워즈, 웹마스터 도구(Search console), 애드센스, YouTube, 이메일 뿐 아니라, 구글 설문지 온라인 데이터도 통합이 가능하며 구글 외의 제품도 데이터를 통합할 수 있다.

<표 6> 구글 애널리틱스 API

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

dCollection @ dau

3.2.5 구글 애널리틱스 데이터베이스 데이터 추출

<그림 12>과 같이 구글 애널리틱스 API를 연동하여 웹사이트에 접속한 로그데이터를 추출한다. 추출데이터 항목은 접속일자, 접속기기, 접속자 수, Sessions, 페이지뷰 이다. 추적코드 및 로그데이터는 구글 애널리틱스 서버에 데이터베이스로 저장되어 있으며 R 프로그램을 이용하여 관련 데이터베이스를 Query을 통해 추출할 수가 있다.

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.

추출데이터의 쿼리항목은 변수 profileId에 구글 에서 발급받은 API ids 값을 할당하고, 추출할 데이터의 시작일, 종료일을 지정한다. 측정값인 메트리스에는 사용자, Sessions, 페이지뷰 항목을 기재하고, 측정항목인 디멘전에는 접속 날자 및 접속기기 정보를 추출하도록 하였다.

```
##==============================================================================
## 2. 데이터 추출
##==============================================================================

ga_profile <- list_profiles(token = ga_token)
load("ga_token")

source("ga_conf.R") # id <- "ga:1xxxxxxxxxx"

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)
```

<그림 12> 구글 애널리틱스 데이터 추출

dCollection @ dau

	date	devicecategory	users	sessions	pageviews
1	2017-09-12	desktop	1	2	46
2	2017-09-13	desktop	2	2	13
3	2017-09-16	desktop	2	7	190
4	2017-09-16	mobile	1	1	1
5	2017-09-17	desktop	1	1	16
6	2017-09-19	desktop	11	12	92
7	2017-09-19	mobile	1	1	9
8	2017-09-23	desktop	1	2	17
9	2017-09-23	mobile	1	3	10
10	2017-09-24	desktop	1	2	9
11	2017-09-26	mobile	1	1	8
12	2017-09-27	desktop	1	1	1
13	2017-09-28	mobile	1	1	9
13	2017-09-28	mobile	1	1	9
14	2017-09-29	desktop	1	1	15
15	2017-09-30	mobile	1	1	7
16	2017-10-01	desktop	1	2	15
17	2017-10-02	desktop	1	1	1
18	2017-10-03	mobile	1	1	1
19	2017-10-04	mobile	1	1	1
20	2017-10-05	mobile	1	2	4
21	2017-10-08	mobile	1	1	1
22	2017-10-09	desktop	1	5	22
23	2017-10-11	mobile	1	2	9
24	2017-10-12	desktop	1	1	1
25	2017-10-12	mobile	1	4	16
26	2017-10-14	desktop	1	2	2

<그림 13> 구글 애널리틱스 추출 데이터

3.2.6 Google Analytics 디멘션과 메트릭스

구글 애널리틱스의 데이터베이스는 데이터의 분석을 용이하게 하기 위하여 데이터를 디멘전과 메트릭스로 구분하여 표현한다. <그림 14>와 같이 디멘전은 접속 지역정보, 유입경로, 페이지 타이틀 등과 같이 개별 고객(Users, Visitors)의 특징, 세션 (Sessions, Visits) 관련 정보 또는 상호작용 정보를 정의한다. 메트릭스는 방문자수, 페이지뷰, 컨버전 등과 같은 숫자로 표기 되는 데이터를 의미한다.

Dimensions Metrics

	페이지	페이지뷰 수	순 페이지뷰 수	평균 페이지에 머문 시간	방문수	이탈률	종료율(%)	페이지 값
		73 전체 대비 비율(%): 100.00%(73)	46 전체 대비 비율(%): 100.00%(46)	00:00:46 평균 조회: 00:00:46 (0.00%)	27 전체 대비 비율(%): 100.00%(27)	51.85% 평균 조회: 51.85%(0.00%)	36.99% 평균 조회: 36.99%(0.00%)	₩8,565 전체 대비 비율(%): 100.00%(₩8,565)
1.	/	23 (31.51%)	19 (41.30%)	00:02:05	17 (62.96%)	70.59%	65.22%	₩2,368 (27.65%)
2.	/15	16 (21.92%)	7 (15.22%)	00:00:02	6 (22.22%)	0.00%	37.50%	₩36,429(425.31%)
3.	/13	10 (13.70%)	4 (8.70%)	00:00:24	0 (0.00%)	0.00%	10.00%	₩10,500(122.59%)
4.	/14	8 (10.96%)	3 (6.52%)	00:00:04	0 (0.00%)	0.00%	0.00%	₩15,000(175.13%)
5.	/admin/design/proc/preview.php	6 (8.22%)	3 (6.52%)	00:00:11	1 (3.70%)	100.00%	50.00%	₩667 (7.79%)
6.	/안사람	2 (2.74%)	2 (4.35%)	00:00:03	0 (0.00%)	0.00%	50.00%	₩0 (0.00%)
7.	/admin/design/design	2 (2.74%)	2 (4.35%)	00:03:52	2 (7.41%)	0.00%	0.00%	₩1,000 (11.68%)
8.	/공지사항/1345708	1 (1.37%)	1 (2.17%)	00:00:00	1 (3.70%)	100.00%	100.00%	₩0 (0.00%)
9.	/오시는길	1 (1.37%)	1 (2.17%)	00:00:03	0 (0.00%)	0.00%	0.00%	₩0 (0.00%)
10.	/admin/design/design?tp=s15031900 1001	1 (1.37%)	1 (2.17%)	00:00:08	0 (0.00%)	0.00%	0.00%	₩1,000 (11.68%)
11.	/admin/manage/apps/DUBU_Sourcec ode/sourcecode/proc	1 (1.37%)	1 (2.17%)	00:00:06	0 (0.00%)	0.00%	0.00%	₩0 (0.00%)
12.	/login	1 (1.37%)	1 (2.17%)	00:03:26	0 (0.00%)	0.00%	0.00%	₩1,000 (11.68%)
13.	/login/login_info	1 (1.37%)	1 (2.17%)	00:02:33	0 (0.00%)	0.00%	0.00%	₩1,000 (11.68%)

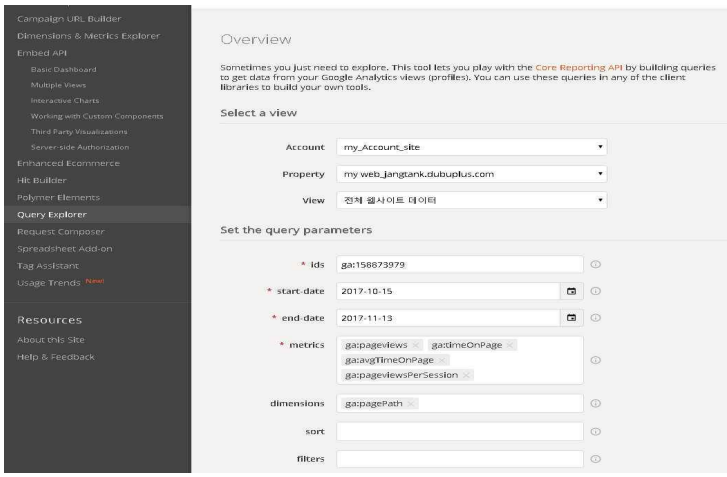
행 표시: 25 이동: 1 전체 13개 중 1-13

<그림 14> 구글 애널리틱스 디멘전, 메트릭스

dCollection @ dau

3.2.7 Google Analytics API 연동 R 프로그래밍

분석 웹사이트의 웹로그 데이터의 분석을 위해 구글에서는 <그림 15> 와 같이 Query View를 제공한다. 분석 웹사이트의 페이지별 사용자 접속현황 분석을 위해 디멘전은 분석 웹사이트의 각 페이지로 설정하고 메트릭스는 접속사용자, 신규사용자, 페이지 체류 시각으로 정의하여 쿼리를 실행하여 데이터를 추출하였다. <그림 16>은 쿼리 실행 데이터 결과 값이다.



<그림 15> 구글 애널리틱스 Overview

my.web_jangtank.dubuplus.com (전체 웹사이트 데이터)

Results showing: 13 | Total results found: 13 | Contains sampled data: No | ↓ Skip to bottom

Page	Pageviews	Pages / Session	Unique Pageviews	Time on Page	Avg. Time on Page
/	23	1.353	19	996	124.5
/13	10	0	4	215	23.889
/14	8	0	3	32	4
/15	16	2.667	7	18	1.8
/admin/design/design	2	1	2	464	232
/admin/design/design?tp=s150319001001	1	0	1	8	8
/admin/design/proc/preview.php	6	6	3	34	11.333
/admin/manage/apps/DUBU_Sourcecode/sourcecode/proc	1	0	1	6	6
/login	1	0	1	206	206
/login/login_info	1	0	1	153	153
/공지사항/1345708	1	1	1	0	0
/오시는길	1	0	1	3	3
/인사말	2	0	2	3	3

<그림 16> 구글 애널리틱스 데이터 결과 값

dCollection @ dau

위의 결과값을 도대로 R 프로그램에서 구글 애널리틱스 <그림 17>과 같은 소스를 개발하여 실행하면 위 결과값과 같은 결과 값을 도출할 수 있다.

```
ga.PageView <- get_ga(profileId = id,
                      start.date = "2017-10-15", end.date = "2017-11-13",
                      metrics = c("ga:pageviews", "ga:uniquePageviews"," ga:timeOnPage",
                                  "ga:pageviewsPerSession"),
                      dimensions = c("ga:pagePath"), sort = NULL, filters = NULL,
                      segment = "gaid::-1", samplingLevel = NULL, start.index = NULL,
                      max.results = 50, include.empty.rows = NULL, fetch.by = NULL, ga_token)
```

<그림 17> 웹사이트 각 페이지별 방문자 현황 R 프로그램 소스

	pagePath	pageviews	uniquePageviews	timeOnPage	pageviewsPerSession
1	/	23	19	996	1.352941
2	/13	10	4	215	0.000000
3	/14	8	3	32	0.000000
4	/15	16	7	18	2.666667
5	/admin/design/design	2	2	464	1.000000
6	/admin/design/design?tp=s1503190001001	1	1	8	0.000000
7	/admin/design/proc/preview.php	6	3	34	6.000000
8	/admin/manage/apps/DUBU_Sourcecode/sourcecode/proc	1	1	6	0.000000
9	/login	1	1	206	0.000000
10	/login/login_info	1	1	153	0.000000
11	/공지사항/1345708	1	1	0	1.000000
12	/오시는길	1	1	3	0.000000
13	/인사말	2	2	3	0.000000

<그림 18> 웹사이트 각 페이지별 방문자 현황 결과 값

3.2.8 R 프로그램 데이터 분석 시각화 프로그래밍

분석 웹사이트의 사용자 접속 현황 웹로그 데이터를 추출하여 분석하기 위해 <그림 19>와 같이 ggplot 함수를 사용하여 시각화하였다. 추출된 웹로그 파일을 gd.df 테이블로 할당하여 시각화를 위한 그래프 형태를 선언하고 시각화 그래프 x축은 접속일자, y측은 웹사이트 접속현황을 시각화 도표에 할당하여 해당 값을 표현하게 프로그래밍 하였다. 해당 시각화 표현은 분석 웹사이트 접속현황을 시각화 하였고, 사용자의 웹사이트 접속일시, 접근기기별, 접속현항을 시각화를 통해 사용자 가독성을 향상시켰다.

dCollection @ dau

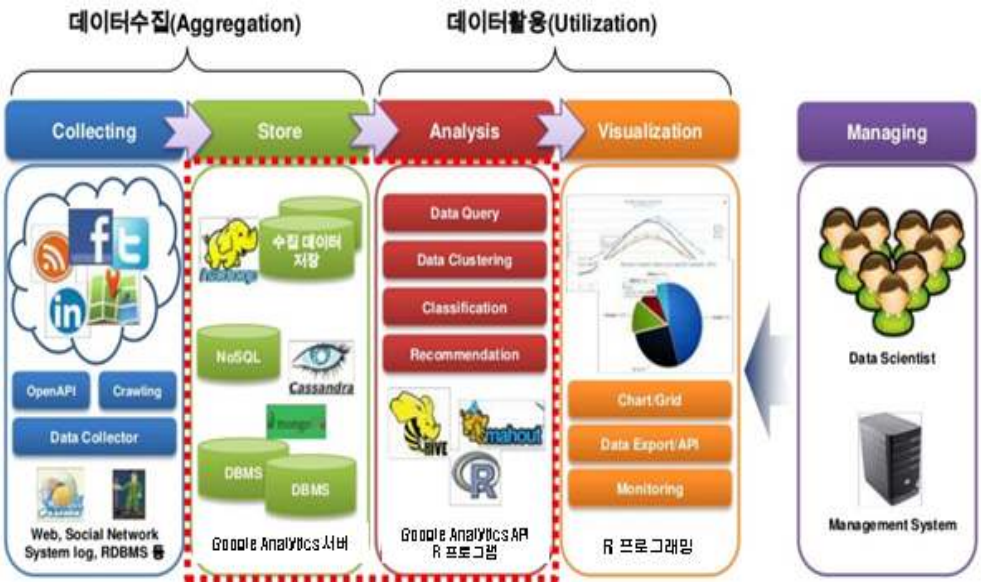
```
ggplot(data = ga.df, mapping = aes(x = date, y = sessions, fill=devicecategory,
      colour = devicecategory)) + geom_bar(stat="identity") + scale_fill_hue(l=80) +
  geom_line() + geom_point(size=3, colour="#CC0000") +
  labs(title = '사용자 접속 현황',
       x = '접속일자',
       y = '웹사이트 접속수') +
  theme_bw()

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 = '접속일자',
       y = '웹사이트 접속수') +
  theme_bw()
```

<그림 19> 시각화 프로그래밍

dCollection @ dau

Ⅳ. 실험 및 결론

구글 애널리틱스 데이터베이스에 저장된 웹로그 데이터의 추출 및 분석을 위해 <그림 20>과 같이 실험환경을 구축하고, 사용자의 웹사이트 접근 시 웹로그 저장은 구글 애널리틱스 서버에 저장되며, R 프로그래밍을 이용하여 Google Analytics API 연동하여 웹로그 데이터를 추출하여 데이터 분석을 위한 시각화를 구현하였다.



<그림 20> 웹로그 데이터 수집 및 분석

4.1 구글 애널리틱스 API 연동 R 프로그래밍

R 프로그램을 활용하여 구글 애널리틱스 API 연동을 위해 <그림 21> 과 같이 구글 GA 계정을 연동하는 프로그래밍을 구축하였고 <그림 22> 와 같이 구글 GA 계정 데이터를 R 프로그램으로 확인할 수 있다.

dCollection @ dau

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

# GA 계정인 등록
client.id <- "245367242808-t6ch0ngee26nahpi79a7t0gsf8nljopk.apps.googleusercontent.com"
client.secret <- "QuTOpEBWhXDI8dW8a1DzAnfD"
ga_token <- authorize(client.id = client.id, client.secret = client.secret)

# 토큰 정보 저자
save(ga_token, file="./ga_token")
```

<그림 21> 구글 애널리틱스 API 연동 R 프로그래밍

Filter

	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

siteSearchQueryParameters	stripSiteSearchQueryParameters	type	created	updated	eCommerceTracking	botFilteringEnabled
jangtank	FALSE	WEB	2017-08-28 10:52:02	2017-08-28 11:15:00	FALSE	FALSE
NA	NA	WEB	2017-08-27 15:14:26	2017-09-16 11:01:25	FALSE	FALSE
jangtank	FALSE	WEB	2017-08-28 01:04:26	2017-08-28 11:15:25	FALSE	FALSE

<그림 22> 구글 애널리틱스 GA 계정 데이터

4.2 구글 애널리틱스 API 연동 데이터 추출

웹로그 데이터의 사용을 위해 R 프로그램의 <그림 23>과 같이 get_ga 함수를 사용하여 구글 애널리틱스 데이터베이스에서 사용자별 접속현황 및 방문한 페이지 현황 웹로그를 추출하였고, <그림 24>와 같이 gd.df는 사용기기별 접속현황 결과를 얻었고, <그림 25>와 같이 ga.PageView 테이블에서 방문페이지별 접속현황 결과를 얻었다.

```
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)


ga.PageView <- get_ga(profileId = id,
                start.date = "2017-10-15", end.date = "2017-11-13",
                metrics = c("ga:pageviews", "ga:uniquePageviews"," ga:timeOnPage",
                            "ga:pageviewsPerSession"),
                dimensions = c("ga:pagePath"), sort = NULL, filters = NULL,
                segment = "gaid::-1", samplingLevel = NULL, start.index = NULL,
                max.results = 50, include.empty.rows = NULL, fetch.by = NULL, ga_token)
```

<그림 23> 구글 애널리틱스 API 연동 데이터 추출 R 프로그래밍

dCollection @ dau

	date	devicecategory	users	sessions	pageviews
1	2017-09-12	desktop	1	2	46
2	2017-09-13	desktop	2	2	13
3	2017-09-16	desktop	2	7	190
4	2017-09-16	mobile	1	1	1
5	2017-09-17	desktop	1	1	16
6	2017-09-19	desktop	11	12	92
7	2017-09-19	mobile	1	1	9

<그림 24> 웹사이트 사용기기별 접속 현황

	pagePath	pageviews	uniquePageviews	timeOnPage	pageviewsPerSession
1	/	23	19	996	1.352941
2	/13	10	4	215	0.000000
3	/14	8	3	32	0.000000
4	/15	16	7	18	2.666667
5	/admin/design/design	2	2	464	1.000000
6	/admin/design/design?tp=s1503190001001	1	1	8	0.000000
7	/admin/design/proc/preview.php	6	3	34	6.000000
8	/admin/manage/apps/DUBU_Sourcecode/sourcecode/proc	1	1	6	0.000000
9	/login	1	1	206	0.000000
10	/login/login_info	1	1	153	0.000000
11	/공지사항/1345708	1	1	0	1.000000
12	/오시는길	1	1	3	0.000000
13	/인사말	2	2	3	0.000000

<그림 25> 방문 페이지별 접속 현황

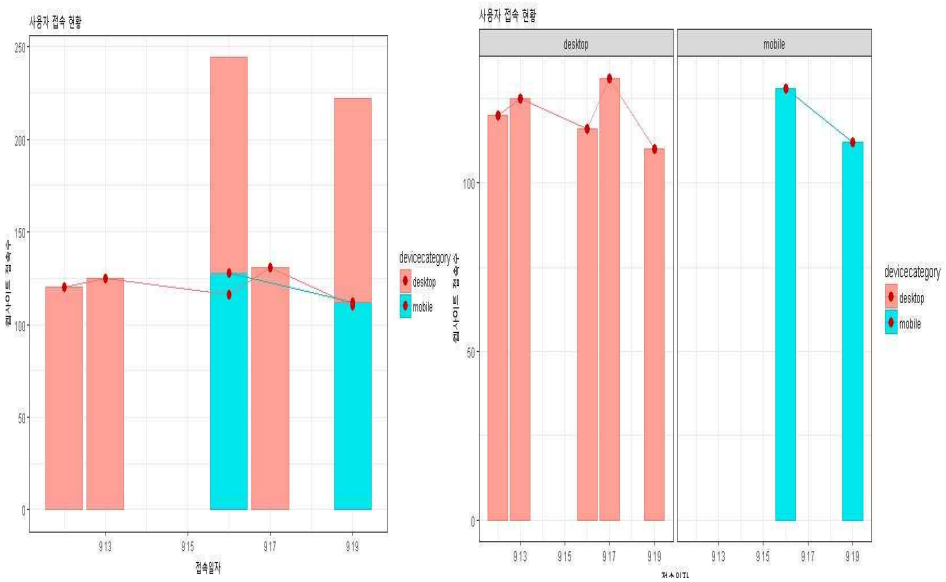
dCollection @ dau

4.3 R 프로그램밍 시각화

추출된 데이터의 시각화 분석을 위해 <그림 25>과 같이 ggplot 함수를 사용하여 시각화 도표를 표시할 수 있도록 하였고 <그림 26>와 같이 웹사이트 기기별 접속현황를 시각화로 표현하였다.

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

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 = '접속일자',
       y = '웹사이트 접속수') +
  theme_bw()
```

<그림 25> 시각화 프로그래밍



<그림 26> 사용자 접속현황 시각화 도표

dCollection @ dau

V. 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 웹로그 데이터를 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.

빅데이터의 정보 표현 및 해석에 있어서 R 프로그램의 시각화 기법을 활용하면 전통적인 방식의 분석그래프, 표 형식에서 벗어난 현대감각에 맞는 데이터 시각화 기법의 많은 표현을 구현할 수 있다. 이와 같은 시각화 기법을 통해 빅데이터 분석 시 새로운 형태의 빅데이터 분석 정보를 볼 수 있어 유용한 해석이 가능하며, 방대한 데이터에 대한 이해 및 의사결정을 가능하게 한다. 이를 위해 본 논문에서는 별도의 비용이 발생하지 않는 R 프로그램을 이용하여 데이터 수집 및 분석 프로그래밍 방향을 제시하였고 효율적인 데이터 분석 시각화 표현을 위한 프래그래밍을 제시하였다.

향후 4차 산업 혁명이라는 산업계의 변화 중심에는 빅데이터의 수집 및 분석이 가장 중요한 이슈가 될 것이다. 이같은 변화의 중심에서 정부 및 민간에서 보유한 빅데이터의 추출 및 분석을 활발히 연구하고 관련 연관사업의 네트워크 인프라 고도화 정책을 반영하여 빅데이터 산업이 보다 효율적이고 활발하게 연구되어 산업체에 반영되어야 할 것이다. 그간 활발한 IT 개발 및 활용에 있어 IT 선진국인 우리나라가 앞으로의 정보화 사회에서 빅데이터를 활용한 신기술 개발에 주도적인 나라로 부상하길 바란다.

dCollection @ dau

참 고 문 헌

[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)

dCollection @ dau

[9] 방승열, “공공개방데이터 활용을 위한 빅데이터 기반의 소프트웨어 아키텍처 설계에 관한 연구” 숭실대학교 소프트웨어 특성화대학원 석사학위논문 p13~p20 (2015)

[10] 정형진, “R 과 D3.js를 이용한 SNS 데이터 분석과 시각화” 서울과학기술대학교 산업대학원 석사학위논문 p16~p25 (2015)

[11] 이미선, “발전된 보안 시각화 효과성 결정 모델” 고려대학교 정보보호대학원 석사학위논문 p8~p16 (2017)

[12] 장유희, “대상과 전개를고려한 데이터 시각화 기법” 숭실대학교 소프트웨어전공 석사학위논문 p12~p17 (2016)

[13] Jordan Tigani, Siddartha Naidu, “Google BigQuery Analytics”, 에이콘, p230~240 (2016)

[14] 유충현, 홍석학, “R을 활용한 데이터 시각화”, 인사이드, p60~p240 (2015)

[15] 노만 매트로프 “빅데이터 분석 도구 R 프로그래밍”, 에이콘, p100~p300 (2012)

[16] 송태민, 송주영 “R을 활용한 소셜 빅데이터 연구방법론”, 한나래, p30~250 (2016)

[17] 나종화, “R 데이터마이닝”, 자유아카데미, p50~300 (2017)

[18] 백영민, “R를 이용한 텍스트 마이닝”, 한울, p40~220 (2017)

dCollection @ dau

[19] https://developers.google.com/analytics/devguides

[20] https://developers.google.com/analytics/devguides/reporting/core/dimsmets

[21] https://ga-dev-tools.appspot.com/query-explorer/

[22] https://developers.google.com/analytics/solutions/articles/hello-analytics-api?hl=ko

[23] https://code.google.com/p/google-api-javascript-client

[24] https://developers.google.com/analytics/

dCollection @ dau

Building and Visualizing Big Data using Google Analytics API

Jang Keun Ahn

Dept. of IT Convergence, Graduate School,
Dong-Eui University

Abstract

Internet data traffic has explosively grown due to the increasing number of smart phone and mass media users around the globe in recent years. The unprecedented surge of mobile data traffic has drawn attention to Big Data, which means huge stockpiles of data, and a considerable number of enterprises and governments have placed their focus on the collection and analysis of Big Data. Big data refers to a wide range of the enormous data sets collected from various sources rather than the creation of a new database. It is difficult for entrepreneurs and researchers to process big data using traditional processing tools. However, in many cases they cannot afford to build new infrastructure for big data processing and storage, such as IoT sensor devices to collect both structured and unstructured data, server equipments to

dCollection @ dau

save collected data, application software to collect and analyze big data.

This paper suggests a novel data analytic approach which does not require the procurement and deployment of new Big Data infrastructure. It is the data analysis method which creates data visualizations utilizing Google Analytics data in the R language. The aim of this study is to analyze data on access records, operations, and event occurrences of websites by creating data visualizations in R and examine different approaches to develop web contents which can apply to various types of websites.

dCollection @ dau

감사의 글

먼저 본 논문 관련 연구가 성공적으로 완료될 수 있도록 격려와 아낌없는 지도를 해주신 장시웅 교수님께 진심으로 감사드립니다. 그리고 바쁘신 와중에도 논문 심사와, 보다 충실한 완성도 높은 논문이 되기 위해 충고와 조언을 해주신 안귀임 교수님, 정덕길 교수님 진심으로 감사드립니다.

대학원 생활 2년 동안 회사업무와 대학원학업을 병행한 주경야독 생활을 하며 정신없는 2년을 보낸 것 같습니다. 늦게 남아 대학원 공부를 마무리 할 수 있어 많은 분들게 고마음을 전하고 싶습니다.

연구실에서 자신의 일을 착실하게 해내는 정동훈 후배님과 정희찬 후배님 등 스마트 연구소에 있는 사람들에게도 감사드립니다.

그리고 항상 나의 힘이 되어준 사랑하는 아내 및 아들·딸에게도 감사드립니다. 또 항상 걱정해주시고 챙겨주신 부모님에게도 감사드립니다. 마지막으로 본 연구에 많은 격려와 연구지도를 해주신 장시웅 교수님께 감사의 인사를 드립니다.

2017년 12월

안장근 올림

dCollection @ dau