빅데이터 분석을 위한 자료 수집 방안 비교

김성국[1] · 오창헌[2*]

[1]두원공과대학교 · [2]한국기술교육대학교

Conparison of Data Collection Methods for Big Data Analysis

Sung-kook Kim[1] · Chang-heon Oh[2*]

[1]Doowon Technical University · [2]Korea University of Technology and Education

E-mail : sayes73@koreatech.ac.kr

요 약

최근 빅데이터 분석에 대한 관심이 높아지고 자료의 수집 방법에 대한 방법도 다양하게 개발되어 지고 있으나 연구자가 이러한 대규모 데이터를 수집·이용하기는 여전히 쉽지 않은 실정이다. 본 논문에서는 연구자가 여러 가지 방법을 활용하여 빅데이터를 수집하는 방안을 비교·분석하여 제시하고자 한다. 본인의 연구 목적에 부합하는 수집 방법을 잘 선택하여 활용한다면 원하는 연구결과를 제공 받을 수 있을 것으로 기대한다.

ABSTRACT

Recently there has been growing interest in big data analysis and methods for collecting data have been developed diversely but researchers are still not easy to collect and use these large scale data . In this paper, researchers try to compare and analyze the method of collecting big data by using several methods and present it. I hope that you can provide the results of your research if you select and use methods that match your research objectives.

키워드

Big Data, Formal Data, Informal Data, Social Data, Crawling, ETL

Ⅰ. 서 론

본 논문에서는 빅데이터를 어떻게 수집하는지, 분석 가능한 데이터에는 무슨 데이터가 있는지 그리고 과거하고 달리 비정형 데이터들을 포함한 아주 다양한 형태의 데이터들을 어떻게 수집하는지 등에 대해 간단히 살펴보고 이를 효과적으로 활용할 수 있는 방안을 제시할 예정이다.

빅데이터의 수집 기술은 조직의 내외부에 있는 다양한 시스템으로부터 로우데이터(raw data)를 효과적으로 수집하는 기술이다. 빅데이터 수집에는 기존의 수집 시스템보다 더 크고 다양한 형식의 데이터를 빠르게 처리해야 하는 기능이 필요한데, 그래서 확장이 가능하고 분산 처리가 가능한 형태로 구성해야 한다. 빅데이터 수집기는 로우(raw) 시스템의 다양한 인터페이스 유형(DB, 파일, API, 메시지 등) 과 연결되어 정형, 반정형, 비정형 데이터를 대용량으로 수집한다 [1-5].

Ⅱ. 데이터의 종류

빅데이터를 형태별로 분류하면 정형 데이터, 반정형 데이터, 비정형 데이터로 분류할 수있다. 정형 데이터는 형태(고정 필드 존재)가 있으며, 연산이 가능하고 예를 들어 관계형 데이터베이스와 스프레드시트, CSV 등이 있다. 반정형 데이터는 역시 형태(스키마, 메타데이타)가 있는 대신에 연산이 불가능한 데이터로서 XML, HTML, JSON, 로그 등이 있다.

마지막으로, 비정형 데이터는 형태가 없으며, 연산도 불가능한 형태로 소셜데이터(트위터, 페이스북), 영상, 이미지, 음성, 텍스트 등이 여기에 포함

* Corresponding author

된다 [6-8].

표 1. 빅데이터의 종류 및 활용

데이터 형태	저장 형태	예시	활용	처리 난이도
정형 데이터	NoSQL File	Excel	관계형 데이터베이스의 데이터처럼 Excel 형식 등의 형식으로 저장됨	하
반정형 데이터	File DBMS NoSQL	XML HTML	XML, HTML 파일과 같은 형식으로 일반적으로 파일 형식으로 저장됨	중
비정형 데이터	File NoSQL	동영상, 이미지, SNS, TXT	언어 분석이 가능한 기사, SNS 등 텍스트 데이터 또는 이미지, 동영상 등으로 저장	상

III. 데이터 수집

데이터 수집이란 조직 내부 또는 외부에 분산되어 존재하는 여러 데이터 소스로 부터 필요로 하는 데이터를 검색해서 수동 또는 자동으로 수집하는 단계를 말한다. 또한, 수집한 데이터를 저장하거나 분석하기 위해서 데이터를 변환하거나 통합하는 단계도 넓은 의미로 데이터 수집이라고 할 수도 있다.

빅데이터의 수집을 위한 데이터의 구분은 데이터 소스의 위치에 따라서 내부 데이터 또는 외부 데이터로 나눌 수 있다.

첫째, 내부 데이터는 자체적으로 보유한 내부의 파일 시스템이나 데이터베이스 관리 시스템과 자체적으로 갖고 있는 센서 등에서 발생한 데이터로써 내부에 있는 데이터의 근원으로부터 자료를 모으는 데이터를 말하며 전통적인 분석을 위해 필요한 자료라 할 수 있다.

두 번째는 외부 데이터로써 인터넷으로 연결된 외부의 자료를 수집한 데이터를 말한다. 예를 들면 우리 회사에서 만들고 있는 상품에 대해서 사람들이 어떻게 생각하는가? 즉, SNS나 웹상에 우리 상품과 관련해서는 어떤 단어들이 많이 나오고 연관해서 늘 나타나는 단어는 무엇이고 우리 상품에 대해서 감정은 어떤지, 긍정적인지 부정적인지 등을 분석해 볼 수 있다.

내부 데이터를 수집하는 방법으로 ETL (Extraction, Transform, Loading)이 대표적이며 원천 시스템으로부터 필요로 하는 데이터를 추출해서 분석 가능하고 조회 가능한 형태로 변환하여 목표 시스템인 데이터 웨어하우스 등에 전송.적재하는 것을 말한다.

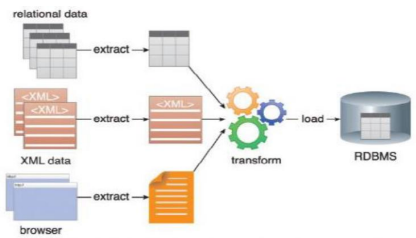

그림 1. ETL 작동 방식

외부 데이터를 수집하는 방법으로 크롤링이 대표적이며 구글 트랜드, 네이버 키워드, 다음 Social Metrics 등 검색어 분석 방식 등이 있다.

먼저, 크롤링은 웹 페이지의 내용 전체를 수집하고 저장하고자 하는 수집 대상을 추출하여 데이터화 하는 것으로 C, JAVA, R 등 프로그래밍 언어를 이용하는 방법과 Textom, KONAN 등 크롤링을 위해 개발된 패키지 형태의 어플리케이션을 이용하는 방법이 있다 [9].

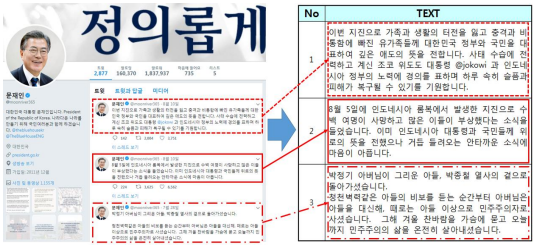

그림 2. 트위터 내 텍스트 데이터 수집 예시

표 2. 검색어 분석 방식 비교

구분	특징	장점	단점
다음 CLIX	·해당 키워드가 어느시점에 얼마나 조회되고 관련 키워드에 대한 정보 제공	·시계열 변화 확인을 통해 사업 운영 전략의 선택과 집중을 위한 우선순위 판단 가능	·네이버나 구글에 비해 검색수가 상대적으로 낮아 거시적 관점의 분석으로는 부적합
다음 Social Metrics	·트위터, 카페, 블로그 등의 소셜 빅데이터 분석 데이터 제공	·SNS나 블로그 검색어 트랜드 확인하여 감성 키워드 분석 등 가능	·정량화된 정보를 제공받지 못해 데이터화를 진행하는데 한계가 있음
네이버 DataLab	·네이버 빅데이터 포털, 공공데이터 등 검색어 트랜드 지표 제공	·분야별 검색어 분석이 가능하여 큰 방향의 트랜드 분석 가능	·거시적 관점의 분석을 제공하여 현재 트랜드를 반영하는데 한계 있음
네이버 키워드	·해당 키워드에 대한 1년간 네이버 검색수 등을 PC와 모바일을 구분하여 제공	·많은 방문 수로 현재 트랜드를 잘 반영하고 있음 ·성별/연령별 분석을 제공하며 연관 키워드에 대한 분석도 제공	·2~3년 전 데이터 등 추세를 볼 수 있는 데이터를 제공하고 있지 않음
구글 트랜드	·검색어에 대해 전 세계를 대상으로 검색/뉴스 추이 그래프와 도시, 국가, 언어, 관련 article 정보 등을 제공	·국가별, 관련 주제별, 관련 검색어별 광범위한 데이터 제공 ·사회적 이슈가 되고 있는 주제에 대한 사람들의 관심도 추적 가능	·지역적 특성에 기반한 분석에는 취약 ·실시간 트랜드는 반영되지 못하고 있으며 세대별, 성별, 분야별 구분에 대한 세부 정보를 반영하지 못함

다음으로, 검색어 분석 방식은 국내외 포털사이트에서 제공하는 데이터로 사용하기 쉽고 무료라는 장점이 있는 반면 데이터가 일정한 형식으로 제공되고 있어 사용자가 원하는 데이터 형식으로 가공이 어렵다는 단점을 가지고 있다 [10].

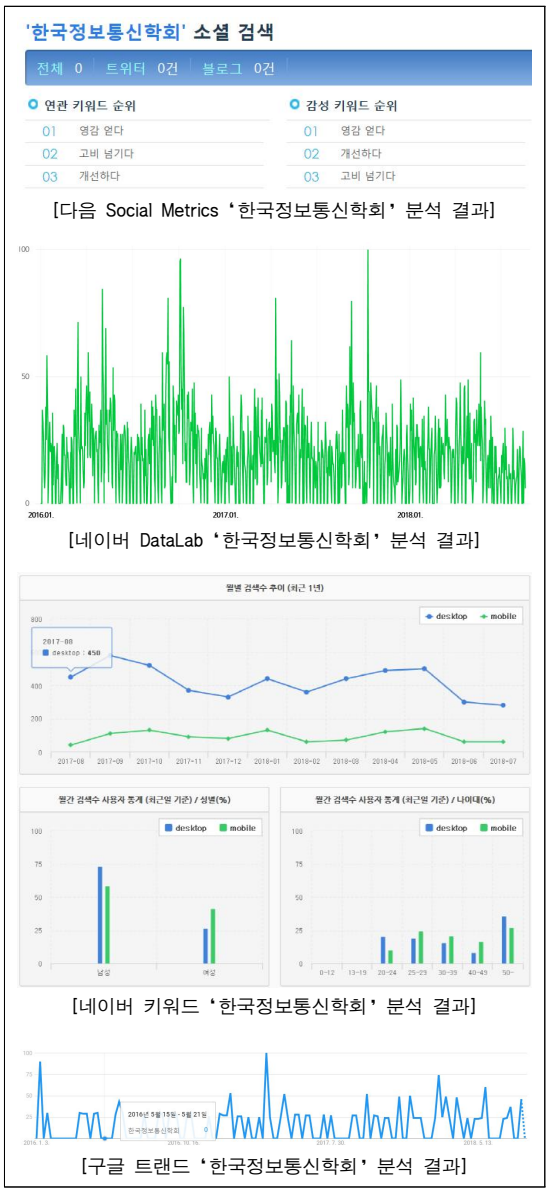

그림 3. 검색어 분석 결과 예시

IV. 결 론

데이터 수집은 조직 내외부에 분산된 여러 데이터 소스로부터 필요한 데이터를 검색하여 수동 또는 자동으로 데이터를 수집하여 저장하거나 분석을 위해 변환 및 통합하는 작업 등 일련의 작업을 뜻하며, 이 과정에 문제가 있을 경우 추후 잘못된 데이터 분석 결과 및 잘못된 의사 결정을 초래하게 되므로 매우 중요한 과정이라 할 수 있다. 따라서 데이터 원천의 물리적 특성뿐만 아니라 데이터 분석 목적을 고려하여 데이터 수집을 수행해야 향후 발생할 수 있는 문제를 예방할 수 있다.

References

[1] W. S. Cho, J. E. Lee, and C. H. Choi, "Refresh cycle optimization for Web Crawlers," *Journal of the Korea Contents Association*, Vol. 13, No. 6, pp. 30-39, 2013.

[2] J. L. Kim, and H. K. Bahn, "An efficient log data management architecture for Big Data processing in Cloud Computing environments," *Journal of the Institute of Internet, Broadcasting and Communication*, Vol. 13, No. 2, pp. 1-7, 2013.

[3] Y. J. Jang, and S. K. Cho, "A comparative Analysis of data gathering and sampling methods for social data," *Journal of Social Science*, Vol. 25, No. 2, pp. 3-25, 2014. 04.

[4] Y. M. Ji, J. J. Yoo, and H. Seo, "Introduce of the large distributed cluster based system for the collecting and processing of big-data of time-series," in *Proceeding of the Korea Institute of Information & Telecommunication Facilities Engineering Conference*, pp. 232-233, 2016. 09.

[5] J. H. Lee, and H. G. Lee, "Development of emotional word collection system for emotional set," in *Proceeding of the Korea Society of Management Information Systems Conference*, pp. 584-590, 2018. 05.

[6] J. H. Yoon, and Y. T. Shin, "Preventing internal information leakage using Big Data-based ETL model," *Journal of the Korea Society of Information Technology Policy & Management*, Vol. 10 No. 1, pp. 675-681, 2018. 02.

[7] Bernard Marr, *BIG DATA: Using smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance,* John Wiley & Sons, 2015.

[8] binglee, Hadoop ecosystem: Big Data collection. [Internet]. Available: http://dabingk.tistory.com/10/.

[9] K. S. Shin, Big Data World. [Online]. Available: http://www.kmooc.kr/courses/course-v1:EwhaK+EW11237K+2017-S06/courseware/93c4d79cbf854cbaa88400e51575ea2b/d8ea77b895d54924ab66f81ba2dfb497/.

[10] M. G. Kang, Big Data analysis using Naver [Internet]. Available: https://brunch.co.kr/@okwinus/24/.