하둡 기반 빅데이터 수집 및 처리를 위한 플랫폼 설계 및 구현

Design and Implementation of Hadoop-based Platform "Textom" for Processing Big-data

손기준*, 조인호, 김찬우, 전채남
(주)더아이엠씨

Son ki-jun*, Cho in-ho, Kim chan-woo,
Jun chae-nam
The IMC, Inc.

요약

빅데이터 처리를 위한 소프트웨어 시스템을 구축하기 위하여 필요한 대표적인 기술 중 하나가 데이터의 수집 및 분석이다. 데이터 수집은 서비스를 제공하기 위한 분석의 기초 작업으로 분석 인프라를 구축하는 작업에 매우 중요하다. 본 논문은 한국어 기반 빅데이터 처리를 위하여 웹과 SNS상의 데이터 수집 어플리케이션 및 저장과 분석을 위한 플랫폼을 제공한다. 해당 플랫폼은 하둡(Hadoop) 기반으로 동작을 하며 비동기적으로 데이터를 수집하고, 수집된 데이터를 하둡에 저장하게 되며, 저장된 데이터를 분석한 후 분석결과에 대한 시각화 결과를 제공한다. 구현된 빅데이터 플랫폼 텍스톰은 데이터 수집 및 분석가를 위한 유용한 시스템이 될 것으로 기대가 된다. 특히 본 논문에서는 모든 구현을 오픈소스 소프트웨어에 기반하여 수행했으며, 웹 환경에서 데이터 수집 및 분석이 가능하도록 구현하였다.

I. 서론

최근 컴퓨팅 패러다임이 클라우드 환경으로 전환되면서 빅데이터의 처리에 대한 관심이 고조되고 있다. 다양한 종류의 빅데이터 중에서 웹과 SNS 데이터는 소프트웨어 시스템에서 고품질의 서비스를 위해 효율적인 관리가 필수적이라 할 수 있다. 이는 실제 서비스가 제공되고 있을 때 사용자들의 관심이나 행동을 확인할 수 있는 유일한 데이터이다. 하지만 다수의 사용자가 생성한 데이터를 수집 및 저장, 분석하기 위하여 추가적인 자원이 필요하다. 이러한 문제는 데이터를 수집하는 것 뿐만 아니라 데이터를 분석할 수 있는 플랫폼을 필요로 한다.

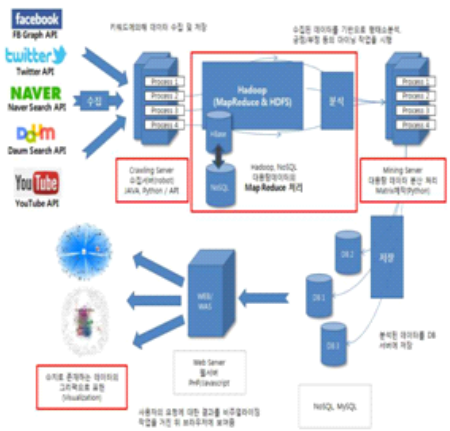
전통적인 내용분석 방법의 경우 연구자가 직접 문서를 읽고 코딩을 한후 분석을 진행해 왔으며[1], 이때 사용하는 대부분의 분석 프로그램이 영어권에 적합하게 제작이 되어 있어서 한국어의 분석에는 적지 않은 한계를 가지고 있다[2][3]. 이런 문제를 해결하기 위하여 대량의 데이터를 안전한 저장 장치에 저장하고 효율적으로 분석할 수 있는 한국어기반의 빅데이터 처리 플랫폼인 '텍스톰'을 제안한다.

본 논문에서는 웹과 SNS상에서 대량으로 생산되는 데이터를 실시간으로 수집하고 효율적으로 저장할 수 있는 빅데이터 기반 데이터 수집 및 분석 플랫폼을 제안한다. 제안하는 방식은 크롤러를 통하여 웹과 SNS상의 데이터를 수집하고, 이를 하둡 파일시스템에 저장한다. 특히 데이터의 안정성 확보와 빠른 저장 및 병렬 처리가 가능하도록 하기 위해 하둡 인프라 환경을 활용한다.

Ⅱ. 빅데이터 기반의 수집 및 분석 플랫폼

1. 빅데이터 수집 및 처리

웹과 SNS상에서 발생된 데이터를 수집하고, 이를 통해 형태소 분석을 위한 자연어처리의 정제 과정, 비정형데이터 내 단어의 공출현 빈도를 계산하여 시각화의 데이터로 활용할 매트릭스를 생산하는 과정, 수집된 데이터를 모두 다 보유할 수 있는 저장 과정을 구조로 갖추고 있다.



▶▶ 그림 1. 빅데이터 수집 및 처리 과정

2. 하둡 분산 환경을 사용한 빅데이터 저장 및 분석

빅데이터라 칭할 수 있는 비정형데이터(텍스트)는 그 저장 및 처리 과정이 효율적이어야 한다. 하둡 분산 환경에서는 대용량 파일을 일정한 블록 크기로 나눈 후 서로 다른 분산 파일시스템의 노드들에 복사본을 분산 저장한다.

본 논문에서 제안하는 아키텍처는 대용량 데이터를 일정 크기의 블록들로 나눈 후 분산 저장함으로써 데이터 처리에서의 효율성을 높인다. 이와 더불어 하나의 데이터에 복수의 사본을 분산 저장하여 오류에 대한 대처를 함께 제공한다.

빅데이터 관리에 있어 또 다른 중요한 요소 중 하나는 저장된 데이터로부터 필요한 정보를 빨리 획득할 수 있어야 한다는 점이다. 하둡 분산 환경을 활용한 인프라는 대용량 파일을 분산 저장하는 것 외에도 빅데이터를 처리할 수 있는 MR(MapReduce) 프로그래밍 모델을 지원한다.

Ⅲ. 구현 및 실험

본 장에서는 제안하는 플랫폼의 구현 및 성능 측정 및 실험에 대해 기술한다. 본 논문에서는 플랫폼 상의 각 계측을 공개 소프트웨어에 기반하여 구현하였으며, 개발 프로토타입 또한 오픈소스 소프트웨어 형태로 누구나 사용할 수 있도록 하였다. 또한 샘플 어플리케이션을 추가로 구현하여 성능 측정 및 확장성 검증을 하였다.

Textom은 빅데이터에서 가장 기본이 되는 데이터를 수집할 수 있는 기능이 갖추어져 있다. 데이터를 수집하는데 있어서 자체 개발된 크롤러를 이용하여 수집이 가능하고, 이를 통해 연구자가 분석하고자 하는 데이터를 수집할 수 있으며, 정확한 데이터를 수집하기 위해서 분석 단어의 '단어묶기'기능을 제공해주며, 분석데이터의 현재 수집 현황에 대한 확인이 가능하다. 또한, 특정 기간의 데이터를 수집할 수 있는 기간 설정 기능을 포함하고 있다.

▶▶그림 2. 빅데이터 수집

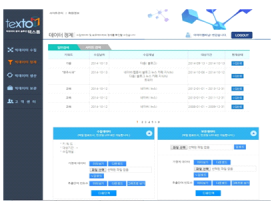


그림 3. 빅데이터 분석

Ⅳ. 결론

제안하는 플랫폼은 다양한 오픈 소스 소프트웨어 모듈로 개발되었으며, 특히 하둡 인프라 환경을 통해 대용량의 웹 및 SNS 데이터를 효율적으로 저장하고 처리 및 분석할 수 있는 빅데이터 기반의 수집 및 분석 플랫폼을 구축하였다.

"이 논문은 2014년도 미래창조과학부의 재원으로 'SW융합기술고도화사업'의 지원을 받아 수행된 연구임" (S1004-14-1010).

■ 참 고 문 헌 ■

[1] Krippendorff, K.,Content Analysis: An Introduction to Its Methodology (2nd Edition). Sage Publication, Thousand Oaks, CA., 2004

[2] 이창환, 윤애선, '한국어 분석 프로그램 (K-LIWC)의 특성과 개발과정', 한국심리학회 연차학술대회논문집, 2004.

[3] 박한우, Leydesdorff, 한국어의 내용분석을 위한 KrKwic 프로그램의 이해와 적용: Daum.net에서 제공된 지역혁신에 관한 뉴스를 대상으로, Journal of the Korean Data Analysis Socity, vol. 6, No. 5, Oct. 2004, pp. 1377-1387, 2004.