클라우드 컴퓨팅의 웹 스크래핑을 이용한 텍스트 데이터 분석에 대한 연구

A Study on the Analysis of Text Data Using Web Scraping of Cloud Computing

저자
(Authors)

김영선, 서춘원
Young-Sun Kim, Choon-Weon Seo

출처
(Source)

대한전자공학회 학술대회 , 2018.6, 1445-1447(3 pages)

발행처
(Publisher)

대한전자공학회
The Institute of Electronics and Information Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07516075

APA Style

김영선, 서춘원 (2018). 클라우드 컴퓨팅의 웹 스크래핑을 이용한 텍스트 데이터 분석에 대한 연구. 대한전자공학회 학술대회, 1445-1447

클라우드 컴퓨팅의 웹 스크래핑을 이용한 텍스트 데이터 분석에 대한 연구

*김영선,**서춘원
*대림대학교 , **김포대학교
e-mail : *yskim306@daelim.ac.kr, **cwseo@kimpo.ac.kr

A Study on the Analysis of Text Data Using Web Scraping of Cloud Computing

*Young-Sun Kim. **Choon-Weon Seo
Daelim University, Kimpo University

Abstract

To collect the text to be used as a text data learning material, it is necessary to measure the accuracy of the concept tagging ability to provide appropriate multiple classification when the classifier learns the text by the web scraping technique will be. In this paper, cloud computing service helps to expand contents and many cloud computing services automatically adjust the number of servers and server load to minimize the cost while maintaining responsiveness.

.

I. 서론

가전제품과 단말기 등을 통하여 유무선 통신으로 구축된 환경에서 정보를 생산하고, 다른 사물 또는 사람에게 전달하는 것으로 수요를 파악하거나 예측 등으로 일정 수준의 자동화된 결정을 주거생활에 적용하는 시스템을 의미한다. 클라우드 플랫폼 환경은 사물인터넷 센서로 부터 수집된 정보를 센터로 데이터를 전송하고, 전송된 데이터들을 클라우드에서 딥러닝 알고리즘으로 각종 센서들의 정보를 사용자 특성에 따라 학습하고, 개인별DB를 구축한다. 구축된 시스템은 각종 센서에서 얻은 정보를 개인별 분석하여 사용자가 동작하지 않아도 이전학습을 기억하고 처리하는 것이다.

본 논문에서는 사용자가 센서들의 상태를 확인하고 제어하는 것이 아닌 스스로 클라우드 플랫폼에서 학습된 정보를 바탕으로 판단하고 사용자에게 알려주는 것이다. 사물을 원격에서 제어하거나 사용자의 스마트폰을 활용하여 센터의 시스템을 모니터링하는 클라우드 플랫폼을 구현하는 것이다.

II. 본론

2.1 클라우드 플랫폼

다양한 형태의 클라우드를 접목한 기술과 서비스들이 등장하고 있다. 통신, 방송, 가전, 건설, 콘텐츠, 로봇 등 다양한 분야로 구성되는 융합 산업의 최대 수요처이다. 네트워크 산업에서 오토 기기 중심의 폐쇄적이고 공급자 중심 환경에서 스마트폰, 스마트TV 확산과 더불어 정보 가전과 백색 가전이 지능화되고 네트워크화되는 스마트 산업으로 빠르게 이동하고 있다.

클라우드 플랫폼의 효과는 간편하고 빠른 IT 자원 운용으로 번거로운 설치, 계약, 운용 과정 없이 서버나 스토리지를 웹서비스로 이용이 가능이다.

2.2 웹 스크래핑 프로그램

웹을 이용하여 웹 사이트 및 블로그에 표시되고 있는 데이터를 추출하는 프로그램으로 유용한 데이터를 복

사하거나 Word 문서에 붙여 넣는 것을 웹 스크래핑 프로그램이라고 한다. 웹 스크래핑 프로그램은 요구사항에 따라 다양한 사이트에서 자동으로 데이터를 로드하고 추출한다.

(1) API와 HTTP

웹에서 데이터 추출을 위한 기본 API와 HTTP 개념을 이해하는 것이 필요하다. 정보를 전달하기 위해서 국제표준화기구(OSI)에서 제시한 OSI 모형 (Open Systems Interconnection Reference Model)3을 사용하고, 이를 기반으로 응용 프로그램을 웹서비스와 데이터 형식에 과거 SOAP와 XML 조합을 많이 사용했다면, 최근에는 RESTful API와 JSON 조합을 주로 사용한다. 웹에서 정보를 얻기 위해서는 서버 API에서 정보를 제공하는 형식에 맞춰 정보를 전달하기만 하면 서버가 제공하는 정보를 받을 수 있고, 이를 이후 데이터과학 작업을 위해 데이터 처리작업을 수행하면 된다.

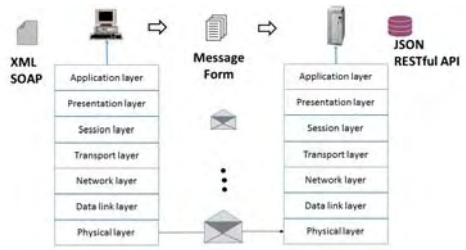

그림 1. API와 HTTP 관계

(2) 웹 스크래핑

웹 스크래핑 라이브러리는 웹 스크래핑 프로그램 중 하나로 비용 부담이 없고 구조화되어 체계적인 파일을 생성한다. 가장 중요한 것은 다른 웹 응용 프로그램의 콘텐츠를 추출하여 다양한 형식과 스프레드 시트, 기타 코드 등으로 변환이 가능하다. 이메일이나 웹 스크래퍼를 사용하면 모든 이메일 주소 및 웹 콘텐츠를 추출 할 수 있으며 해당 데이터를 복사하여 Word 파일에 붙여 넣을 수 있다. 그 외 CSV 파일, PDF 및 XML 파일을 자체 DB에 저장할 수 있어 언제든지 액세스가 가능하다.

Ⅲ. 구현

3.1 텍스트 데이터 추출

웹페이지에서 저자의 위치를 분석해 유형에 따라 자동으로 저자를 추출할 수 있는 시스템 구축을 위한 시스템을 이용해 텍스트의 위치를 추출하도록 학습한다. 많은 양의 정보가 웹을 통해 생산되고 퍼져나가고 저장한다. 텍스트 분석에 필요한 다량의 텍스트 자료를 모으기에 적합하다. 웹으로부터 자료를 자동으로 검색하고 수집하는 웹 스크래핑의 중요성이 갈수록 커지고 있다. 웹 스크래핑은 하이퍼 링크를 따라가며 새로운 웹 문서를 발견하는 기능과 웹 문서에서 사용자가 수집하고 싶은 특정 정보만을 추출하는 기능을 제공한다. 일반적으로 사람들이 웹 문서에서 관심을 두거나 수집하고자 하는 것은 문서의 핵심 정보를 담고 있는 텍스트의 웹 문서 내에서 자동으로 본문을 추출하는 기능을 웹 콘텐츠 추출(Web Content Extraction)한다.

그림 2. 클라우드 시스템 구성도

3.2 텍스트 데이터 추출의 알고리즘

텍스트 데이터 추출은 웹 문서에서 본문 영역이 흔히 가지고 있는 구조나 서식의 특성(feature)을 활용한다. 웹 문서에서 텍스트 데이터 블록이 흔히 가진 특성으로는 텍스트 데이터 블록은 서식을 표시하는 태그(tag)에 비해 많은 양의 텍스트를 포함한다. 웹 문서에서 광고나 메뉴 등은 사이트 내 여러 곳에 나타나지만 텍스트 데이터는 한 번만 나타나는 경향이 있다. 텍스트 데이터 블록은 웹 문서의 DOM 구조에서 일종의 안정(plateau)형태를 갖는다. 텍스트 데이터는 문단을 나타내는 <p> 태그를 많이 포함하고 있다. 텍스트 데이터 추출은 데이터의 특성들을 바탕으로 연구자가 설정한 규칙을 통해 추출하거나 혹은 위의 데이터 특성들을 변수로 활용하여 기계학습을 통해 이루어진다. 머신 러닝 알고리즘은 의사 결정 트리(decision tree)를 이용한다. 의사 결정 트리는 의사 결정 지원 도구로, 우연한 이벤트 결과, 리소스 비용 및 유틸리티를 포함하여 트리와 같은 그래프 또는 의사 결정 모델 및 가능한 결과를 사용한다.

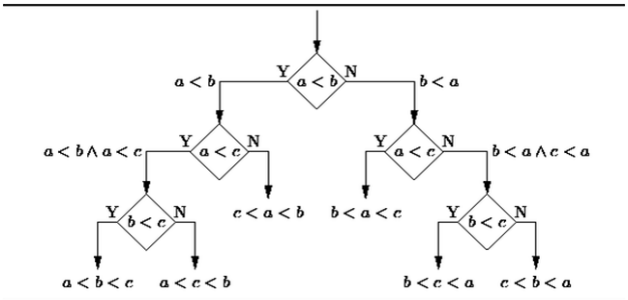

그림 3. 의사 결정 트리 알고리즘

의사 결정 트리는 대부분의 시간에 올바른 결정을 내릴 확률을 평가하기 위해 질문해야 하는 예/아니오 질문의 최소 수이다. 방법으로 논리적인 결론에 도달하기 위해 구조화되고 체계적인 방식으로 문제에 접근할 수 있다.

```
<html>
<body>
  <div class="class1">
    <p>문단1이다.</p>
    <p>문단2</p>
    <p>문단3이다.</p>
</div>
<div class="class2">
  <a href="/home">링크1</a>
  <a href="/board">링크 2</a>
  <div class="sub1">
    문장1</br>
    문장2</br>
  </div>
</div>
</body>
</html>
```

그림 4. 텍스트 데이터 추출 모델

Ⅳ. 결론 및 향후 연구 방향

텍스트 데이터 추출은 하나의 문서 혹은 다수의 문서로부터 문서의 핵심적인 내용을 담고 있는 단어(word) 혹은 구(phrase)를 찾아내는 기술이다. 텍스트 데이터 추출은 주제 분류, 문서 검색, 문서 요약 등에 활용될 수 있는 자연어 처리의 중요한 기반 기술로, 주로 문서 요약 알고리즘을 이용하는 경우가 많다. 텍스트 데이터 추출은 문서에 포함된 단어 혹은 구들이 문서의 핵심적인 내용을 담고 있어야 하기 때문에 키워드 추출 알고리즘으로는 통계치 이용 방법, 언어학적 방법, 기계학습을 이용한 방법이 있으며 이를 복합적으로 이용하는 하이브리드 방법이 있다. 본 연구는 순수 단어 출현 빈도 및 문서 내 단어의 동시 출현 빈도와 형태소 분석 결과에 기반하여 키워드 추출 알고리즘을 개발하는 것으로 다량의 텍스트와 그에 따른 분류를 전산 가능한 모양으로 변환시켜 분류기를 학습시키는 것이 목표로 하고 있다.

참고문헌

[1] 노승민, 최용수, “특허분석을 통한 빅 데이터의 시각화 기술 분석”, 대한전자공학회, 전자공학회논문지 51(7), 2014.7, 149-154.

[2] 김현종, 조정민, 안병구, “데이터마이닝 및 텍스트 마이닝 엑세스 분석결합을 통한 의료 데이터베이스의 효과적인 이용 ”, 대한전자공학회, 대한전자공학회 학술대회, 2012.11, 793-795.

[3] 황명하, 하수욱, 인민교, 이강찬, ‘텍스트 마이닝 기반 클라우드 시스템을 이용한 보안 트렌드 분석’, 보안공학연구논문지 제14권 제5호 (2017년 10월).

[4] 박선주, “SW 교육 뉴스데이터의 감성분석”, 한국정보교육학회, 정보교육학회논문지, 제21권 제1호 (2017. 2),

[5] 김재환, 이재문, “빅데이터 분석을 활용한 워터파크 현황 및 인식 분석”, 한국디지털정책학회, 디지털융복합연구 제15권 제10호 (2017년 10월).

[6] 남길임, 조은경, "한국어 텍스트 감성 분석", 커뮤니케이션북스. 2017.

[7] 조인행,“Word2Vec을 이용한 위키피디아 텍스트 데이터 분석 시스템 구현”,숭실대학교 소프트웨어특성화대학원, 석사논문, 2017.2. p13-21.

[8] 오하영, 박정식,“컴퓨터와 R을 활용한 텍스트 데이터 분석 기초”, Human Science(휴먼싸이언스), 2017. p.23-36.