데이터셋 검색 지원을 위한 메타데이터 자동 추출에 관한 연구

정종진, 김경원, 김구환

전자부품연구원

mozzalt@keti.re.kr, kwkim@keti.re.kr, mdbow81@gmail.com

A Study on Automatic Metadata Extraction to Support Dataset Search

Jung Jong Jin, Kim Kyung Won, Kim Gu Hwan

Korea Electronics Technology Institute

요 약

본 논문은 웹 또는 인터넷망에 무수히 많이 흩어져 있는 데이터 저장소에 존재하는 데이터셋의 특징들을 추출하고 이를 메타데이터로 자동으로 표현하는 방법을 소개한다. 개별 데이터셋이 가진 특징들이 메타데이터로 표현되기 때문에 분석 목적에 맞게 보다 정확히 검색되어 활용도를 높일 수 있다. 또한 데이터셋의 특징을 기술하는 유력 메타데이터 형식들을 상호 호환 가능하여 활용에 있어 보다 확장성을 갖는 방법도 제시한다.

Ⅰ. 서 론

웹에는 수천 개의 데이터 저장소가 있어 통상 데이터들은 데이터 셋 단위로 배포나 공유가 되므로 수백만 개의 데이터 세트에 검색 할 수 있다. 국가 및 지역 정부, 과학 출판사 및 컨소시엄, 상업용 데이터 제공 업체 및 기타 기관은 사회 과학, 생명 과학, 고 에너지 물리학, 기후 과학 등의 분야에 대한 데이터를 제공 중이며 이 데이터 셋 들에 대한 검색을 위한 접근성을 부여하는 것은 분석 결과를 고도화 하거나 데이터 셋의 재생산성 촉진할 수 있도록 하는 데 중요하다. 최근 Google 에서는 Google Data Search를 통하여 데이터 검색 툴을 제공 중인데, 이 툴은 웹에 게시 된 모든 데이터 집합에 대해 검색 기능을 제공하는 툴이다. 이 툴의 핵심은 데이터 셋 소유자와 공급자가 자체 사이트에 의미 있게 향상된 메타 데이터를 함께 게시하여 효과적인 검색이 되도록 하는데 있다[1]. 따라서 본 논문에서는 데이터 셋 제공자가 데이터를 공유 시 제공한 메타데이터를 해석하여 데이터 셋 검색을 할 수 있도록 메타데이터를 변환 보완하는 메타데이터 자동 추출 연구를 소개한다.

Ⅱ. 메타데이터 자동 추출 시스템 설계

1. 데이터 셋과 메타데이터

데이터는 많은 분야의 서비스 개발자, 연구인들의 분석 및 세계를 더 잘 이해하기를 원하는 사람들을 위한 주요 재료이다. 웹 또는 인터넷망에 게시하는 데이터가 많을수록 데이터 검색 문제가 더욱 복잡 해 진다. 웹 초기시절에는, 많은 유저들이 야후 같은 웹 검색도구와 같은 것을 브라우징 함 으로써 필요한 데이터들을 찾았다. 하지만 최근에는 웹이 매우 비대해져서 웹상에서 정보가 어디에 있는지 찾기 매우 어려워 졌다. 웹 검색 엔진은 여러가지 이유 때문에 데이터를 잘 못 찾는다[1]. 이런 문제점을 해결하고자 확보중인 데이터 셋의 주요 특징들을 메타데이터로 표현하고 이를 데이터 셋과 함께 공유가 되면 데이터를 필요로 하는 입장에서는 보다 정확단 데이터 셋을 검색하고 획득 할 수 있다.

2. 메타데이터 자동 추출

메타데이터는 정확한 탐색이 되도록 하는 주요 수단이기 따문에 데이터 셋의 주요 특징을 잘 반영하여야 한다. 데이터 셋의 주요 특징들은 예들들어 데이터 셋을 구성하고 있는 개별 데이터들의 공통 주제, 데이터 셋 발행자, 소유자, 활용 빈도 등등이 어떤 데이터인지에 대한 정보와 해당 데이터들이 분석 수요자들에게 활용 가치가 어떠한지에 대한 주요한 정보가 될 수 있다[2]. 흔히 데이터들은 주로 문서 위주의 Text형태, Image, 비디오 형태로 존재한다. 이런 형태로 존재하는 데이터셋 들은 데이터 셋에 대한 특징을 기술하는 메타데이터가 존재 하는 경우도 있지만, 대부분 그렇지 않은 경우도 많고, 있다 하더라도 정확한 탐색을 위해서는 주요한 정보들이 포함되어 있지 않다. 따라서 본 논문에서는 Text, Image, Video 형태로 존재하는 데이터 셋들을 대상으로 데이터 셋 내 개별 데이터들의 주요 특징들을 추출하고, 이들을 군집화 분석하여 데이터 셋이 가진 특징을 메타데이터화 하는 과정을 그림 1과 같이 설계하였다.

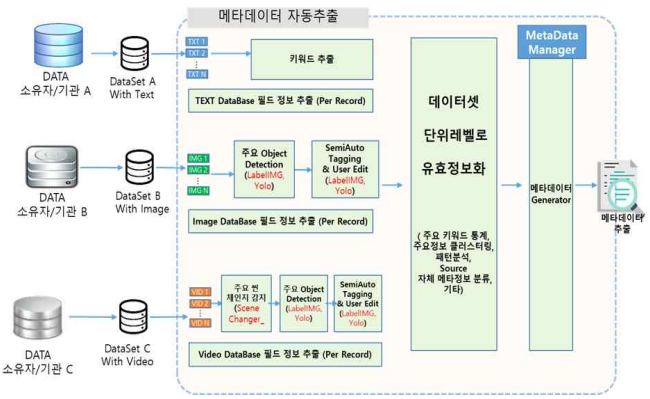

그림 1. 데이터 셋 유형별 특징 추출 및 메타데이터 구성 시스템

그림 1에서 살펴 볼 수 있듯이 Text들로 구성된 문서 위주의 데이터 셋이라면 데이터 셋 내 개별 문서하나하나의 텍스트 들로부터 자연어 처리, 빈

도 분석 및 토픽 모델링 기법을 활용하여 키워드/주제어를 도출한다. 이런 키워드/주제어 도출은 모든 문서들에 대해 동일하게 수행되며, 모아진 개별 문서단위 키워드/주제어들에 대한 군집화 분석이 가능해 진다. 따라서 데이터 셋들이 담고있는 데이터들이 어떤 주제들로 구성되어 있는지에 대해 메타데이터화가 가능해진다. 이미지 데이터들인 경우 개별 이미지내에 포함된 객체들을 인식한 후 이를 라벨링 하여 저장 관리한 뒤 전체 이미지를 대상으로 모두 수행하게 되면 어떤 객체들이 분포되어 있는지에 대한 통계정보나, 주요 객체들이 무엇인지 등이 주요하게 메타데이터화 되어 표현된다. 비디오인 경우 주요 씬(Scene)이 교체되는 지점을 감지하고, 각 씬(Scene)마다 이미지 데이터를 처리하는 과정을 동일하게 반복 수행하게 되면 비디오 데이터 셋에 대한 메타데이터들도 자동 추출되거나 보완이 된다. 한가지 실험으로 써 학술 정보를 다루는 데이터 셋, 즉 Text 위주의 데이터셋을 대상으로 메타데이터 자동 추출을 하는 경우라면 그림 2와 같은 과정을 거쳐 학술정보에 대한 주요 정보들이 메타데이터화 되어 검색에 활용 될 수 있다.

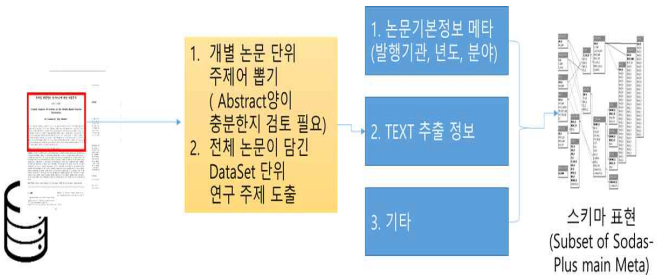

그림 2. 학술정보(논문) 데이터셋 대상 메타데이타 자동 추출 예시

3. 메타데이터 자동 변환

앞서 설명한 바와 같이 데이터 셋에 포함된 메타데이터는 데이터 셋을 표현하는 중요한 수단이자, 검색엔진이 이를 해석하여 정확한 검색을 해 주는데 활용된다. 하지만 웹 상에 존재하는 많은 데이터 셋들은 하나의 형식으로 메타데이터로 표현되어 있지 않기 때문에 제각각 다른 형식으로 표현되어 있는 메타데이터를 해석하여 데이터 셋 검색을 할 수 있도록 해줘야 한다[3]. 물론 세상에 존재하는 모든 메타데이터들을 호환 할 수는 없다. 이렇게 하기 위해서는 메타데이터 변환 모듈이 이 세상에 활용중인 모든 메타데이터 해석기를 가지고 있어야 하며, 있다 하더라고 개별 형식에 대한 업데이트를 지속적으로 추적 관리해야 한다. 따라서 본 논문에서는 현재 데이터 셋을 기술하는 대표적인 DCAT과 Schema.org로 표현된 메타데이터를 우선 대상으로 해석하여 변환하는 연구를 그림 3과 설계하였다[4][5][6].

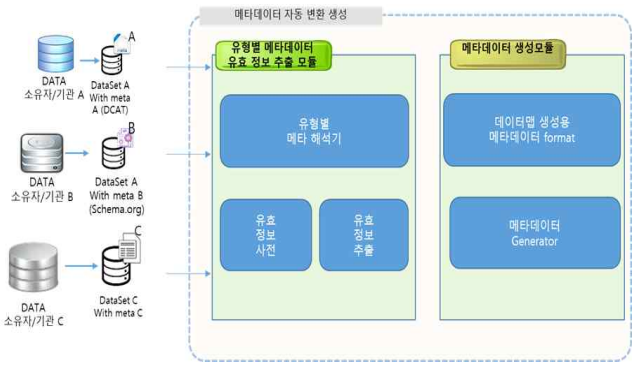

그림 3. 메타데이터 자동 변환 및 생성 설계도

Ⅲ. 결론

본 논문에서는 분석에 필요한 데이터를 보다 쉽게 검색하여 활용할 수 있게 도움을 주는 데이터 탐색을 위한 주요한 첫 단계인 데이터 셋으로부터 메타데이터를 추출하여 생성하는 연구를 소개하였다. 데이터 셋의 특징을 보다 올바르게 분석하여 이를 기술함으로써 웹 상에 무수히 많이 존재하는 데이터 저장소들로부터 정확한 데이터셋을 활용하게 하는 기술임과 동시에 향후 데이터 경제가 활성화 되는 많은 서비스 분야에서 주요하게 활용되리라 기대된다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. 2020-0-00077, 데이터랩 기반 지능형 빅데이터 탐색·활용 핵심 기술 개발)

참 고 문 헌

[1] Natasha Noy, Matther Burgress, Dan Brickley."Google Dataset Search: Building a search engine for datasets in an open Web ecosystem" WebConf '2019, May 2019, San Francisco, CA USA ,pp. 129-132.

[2] Guha, R. V., Brickley, D., and Macbeth, S. Schema.org: evolution of structured data on the web. Communications of the ACM 59, 2 (2016), 44 - 51.

[3] Kaggle datasets. https://www.kaggle.com/datasets.

[4] Data Catalog Vocabulary (DCAT). https:// www.w3.org/ TR/vocab-dcat/.

[5] CKAN. http://ckan.org(http//www.nist.gov/aes).

[6] Open data network. https://www.opendatanetwork.com/.