구글 클라우드 자연어 API를 이용한 DBpedia 웹 검색 애플리케이션

이수형, 김태영, 박선재, 이용주
경북대학교 IT대학 컴퓨터학부
e-mail: poponaya@naver.com, tykim0910@nate.com,
xz3272@naver.com, yongju@knu.ac.kr

DBpedia Web Search Application using Google Cloud Natural Language API

Suhyoung Lee, Taeyoung Kim, Parksun Jae, Yongju Lee
School of Computer Science and Engineering, Kyungpook National University

요 약

본 논문은 링크드 오픈 데이터(Linked Open Data)의 일종인 DBpedia 개체를 자연어 기반으로 검색하는 애플리케이션 개발에 관한 논문이다. Google Cloud Natural Language API를 이용하여 자연어 입력을 분석하고, 이를 바탕으로 RDF(Resource Description Framework) 검색 언어인 스파클(Sparql) 질의 문장을 작성하여 결과를 웹 형식으로 반환해준다. 이를 통해 비문가도 손쉽게 링크드 오픈 데이터에 접근할 수 있는 기회를 제공하며 다양한 응용 가능성을 가진다.

1. 서론

1990년대 월드 와이드 웹이 창시된 이래로, 다양한 웹 기술의 발전으로 인터넷은 일상생활에도 다양하게 활용되며 뗄래야 뗄 수 없는 관계가 되었다. 수많은 이용자로부터 다양한 정보가 생성 및 소비, 유통되고 있다. 반면 정보의 양이 폭발적으로 증가하고 있지만, 기존의 웹은 정보의 재사용이나 컴퓨터가 데이터 안에 있는 의미를 분석, 활용하기에는 어려운 부분이 많았다. 이러한 점을 보완하기 위해 시맨틱 웹이 창시되었고 최근 시맨틱 웹에 대한 개발이 활발히 진행되고 있으며, 이를 활용한 다양한 방법으로 응용하여 사용하는 애플리케이션과 서비스가 등장하고 있다. 시맨틱 웹 기술은 정보간의 연계가 쉽고, 추론이 가능한 장점을 가져서 정보서비스를 뛰어 넘는 지식 서비스가 가능한 장점을 가진다[1].

본 논문은 구글 클라우드 자연어 API를 이용하여 링크드 오픈 데이터를 활용하는 웹 애플리케이션 개발 사례로, 시맨틱 웹을 구현하는 방법인 링크드 오픈 데이터(Linked Open Data) 형식을 가지는 DBpedia를 활용하며, 자연어 입력을 통한 DBpedia 검색 애플리케이션이다. 입력된 자연어를 바탕으로 개체를 찾을 수도 있고, 개체가 가지는 속성에 대한 정보를 찾을 수 있다.

링크드 오픈 데이터의 규모는 빠르게 커져가고 있으며 다양한 도메인에서 활용되고 있다. 본 논문에서 개발한 웹 애플리케이션은 링크드 오픈 데이터의 일종인 DBpedia를 손쉽게 접근할 수 있도록 도와주어, 다양한 상황에서 링크드 오픈 데이터를 활용할 수 있도록 해준다.

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2016R1D1B02008553). 이 논문은 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학사업의 연구결과로 수행되었음(2015-0-00912).

2. 이론적 배경

2.1 시맨틱 웹

전통적인 웹에 존재하는 데이터는 인간이 사용하기 위해 만들어진 데이터로 사람만 이해할 수 있으며 컴퓨터에게는 무의미한 경우가 대부분이다. 시맨틱 웹이란 한글로 표현하면 의미론적 웹을 말하는 것으로, 기존 웹과의 차이점을 간단하게 표현하면 컴퓨터가 소프트웨어 및 프로그램을 이해하고 조작할 수 있는 환경을 뜻한다.

시맨틱 웹에서는 정보들 사이에 관계를 정의하고, 또한 정보들 관계 사이에서 추론 등을 통해 새로운 지식을 생성할 수 있다. 이러한 환경을 통해 시맨틱 웹은 기존 웹과 같이 단어를 식별해서 관련된 사이트나 문서를 찾아줌과 동시에 새롭게 구성된 문서에 사물간의 관계를 명확히 기술하여 정확하고 의미 있는 정보를 제공하는데 그 목표가 있다[2].

2.2 링크드 오픈 데이터

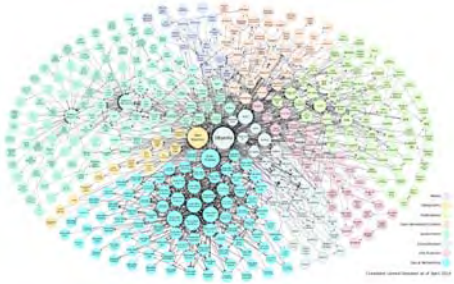
기존의 웹에 존재하는 데이터들은 대부분 정형화되어 있지 않고 다양한 형태를 가져 각 데이터들이 서로 연결되지 않았다. 링크드 오픈 데이터(Linked Open Data) 또는 링크드 데이터(Linked Data)는 시맨틱 웹이 표방하는 데이터 웹(Data Web)을 구체적으로 구현하는 방법으로,

웹으로 접근 가능한 이름(URI)을 붙이고 이를 통해서 RDF(Resource Description Framework) 형태의 시맨틱 데이터를 서로 연결함으로써 데이터를 공개하고, 공유하고, 연결하기 위한 방법이다[3]. 이러한 상호 연결성을 가지기 때문에 내가 직접 만들지 않은 데이터나 다른 곳에 있는 데이터도 연결을 통하여 하나의 데이터처럼 사용이 가능하다.

위와 같은 특성을 바탕으로 링크드 데이터는 기존 문서 위주의 WWW 전달 방식을 페이지가 아닌 데이터간의 연결을 중심으로 전환하여 보다 풍부한 자원의 생산과 효율적인 활용이 가능한 방식으로 웹을 지능화시킨다[4]. 또한 기존의 웹과 별개로서 존재하는 것이 아닌 기존의 HTTP, URI 등 표준 웹 기술을 사용함으로 기존의 웹과 함께 사용 가능하다.

링크드 오픈 데이터는 주로 RDF(Resource Description Framework) 형식으로 구성되어 활용 되고 있으며, 이러한 링크드 오픈 데이터는 현재 미디어, 지리정보, 정부, 의과학, 등 공공/민간의 다양한 부문에서 데이터 구축 및 활용 사례가 등장하고 있다. 링크드 오픈 데이터는 2007년 12개의 데이터 세트를 시작으로 2017년 기준 10년 만에 1,163여 개로 빠르게, 다양한 분야에서의 데이터들이 구축되고 있다. (그림 1)은 하나의 데이터 세트를 원으로 표시하고 연결 관계를 표시한 모습으로 많은 데이터 세트들이 연결되어있는 모습을 볼 수 있다. (그림 1)은 2014년 기준의 모습으로 현재는 약 2배 정도의 숫자를 가지는 데이터 세트가 존재한다.

본 웹 애플리케이션에서 활용하고자 하는 DBpedia는 온라인 백과사전인 Wikipedia의 데이터를 바탕으로 만들어진 링크드 오픈 데이터이다. 때문에 다양한 분야에 걸쳐 많은 데이터를 가지고 있기 때문에 다양한 데이터 세트들의 중심에 위치하고 있는 것이 DBpedia로 (그림 1)과 같이 여러 분야에 많은 연결을 가지고 있는 것을 볼 수 있다.



(그림 1) LOD cloud

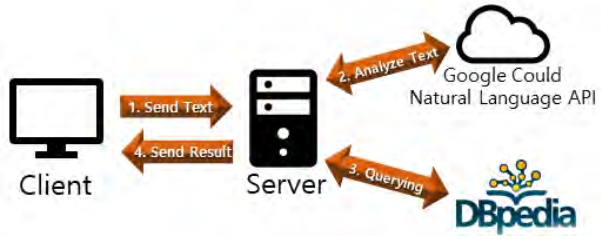
3. 웹 애플리케이션 개발

DBpedia 개체에 접근할 때에는 URI를 이용하기 때문에 특정 개체에 해당 개체의 URI가 필요하다. 이러한 주소를 검색하기 위해서 RDF 질의 언어인 스파클(Sparql)을 이용한다. 스파클을 이용하여 해당 개체가 가지고 있는 다양한 속성을 검색하거나 특정 속성을 가지는 개체를 검색하는 등 다양한 검색이 가능하다. 하지만 스파클을 사용하기 위해선 어느 정도 스파클에 대한 사용법 학습이 필요하다. 그렇기 때문에 스파클에 관한 학습 없이 DBpedia에 대한 데이터에 접근하는 것에 어려움이 따르기 때문에, 각 데이터 세트들을 활용하기 위한 서비스 및 애플리케이션들이 등장하고 있다. 본 논문에서 구현한 애플리케이션은 이러한 접근성을 보완하기 위해 작성된 애플리케이션으로, 사용자로부터 일반적인 문장 형태의 입력을 받고, 이를 해석하여 DBpedia 개체를 검색할 수 있도록 도와준다. 이를 통하여 특정 도메인이나 특이한 상황에 맞는 애플리케이션이 아닌 범용적인 검색을 목적으로 사용할 수 있게 애플리케이션을 구현하였다.

3.1 애플리케이션 구현

애플리케이션은 웹을 기반으로 작동하는 서버-클라이언트의 형식으로 구현했다. 서버-클라이언트 구현에는 Node.js 프레임워크를 사용했다. 먼저 웹 페이지를 통하여 사용자로부터 입력받은 텍스트를 서버로 전송하고, 서버는 전송받은 텍스트를 Google Cloud Natural Language API를 이용하여 구성성분을 분석한다. 분석 결과를 바탕으로 서버에서 주어, 술어와 같은 부분을 파악 및 선택하여 스파클 질의문을 작성한다. 또한 술어부분에 관하여 개체가 가진 속성 이름과 입력된 텍스트의 술어부분이 정확하게 일치 하지 않아도 유사한 단어를 검색 하도록 구현하여 검색 확률을 높였다.

위와 같은 과정을 통해 작성된 스파클 질의 문장을 DBpedia에서 제공하는 스파클 Endpoint에 제출하여 결과를 받아온다. 스파클 질의 결과를 바탕으로 클라이언트 측으로 전송하여 사용자에게 웹 페이지 형태의 결과를 제공한다. (그림 2)는 전체적인 애플리케이션 작동 절차를 도식화하여 나타낸 모습이다.

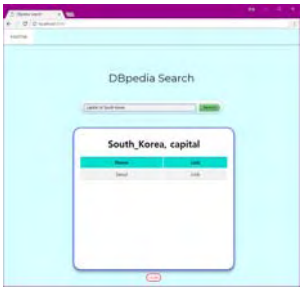


(그림 2) 시스템 구성도

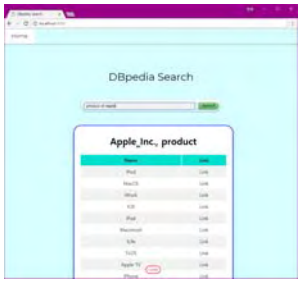
3.2 애플리케이션 구성

첫 페이지에 접속하면 (그림 3)과 같이 텍스트를 입력할 수 있는 상자와 검색 버튼이 존재한다. 텍스트 상자를 통해 사용자로부터 문장으로 이루어진 입력을 받고, 검색 버

튼을 눌러 텍스트를 서버로 전송하며, 전송된 텍스트를 자연어 처리를 한 뒤, 결과를 사용자에게 반환한다. 반환 결과로는 개체의 이름과 링크를 제공해준다. (그림 3)과 (그림 4)는 각각 'capital of South Korea', 'product of Apple'를 입력하여 얻은 결과 화면이며 실행 결과 (그림 3)은 'capital of South Korea'의 검색 결과로 Seoul이 표시되어있고, (그림 4)는 'product of Apple'의 검색 결과로 Ipod, MacOS 등이 표시되어 있으며 각 개체들에 접근할 수 있는 링크들을 제공하고 있다. 또한 'Where is the capital of South Korea?'과 같이 입력해도 (그림 3)과 같은 결과를 얻을 수 있다.



(그림 3) 서울의 수도를 검색한 결과



(그림 4) Apple의 제품을 검색한 결과

3.3 애플리케이션의 개발 기법

3.3.1 Node.js

Node.js는 Chrome V8 JavaScript 엔진으로 빌드된 JavaScript 런타임으로, JavaScript로 이루어져 있는 단일 스레드로 동작하는 고성능 비동기 IO(Non-blocking I/O)를 지원하는 플랫폼이다. JavaScript 기반이며 개발 구조가 단순화되어 빠르게 개발이 가능하기 때문에 다른 플랫폼보다 생산성 측면에 유리한 점이 있다. 성능과 생산성 측면에서의 장점을 바탕으로 최근 Node.js를 많은 곳에서 활용하고 있으며, 많은 라이브러리들이 등장함에 따라 더욱 더 개발이 용이하게 되었다. 본 논문의 웹 애플리케이션의 구현도 백엔드와 프론트엔드 모두 Node.js를 활용하여 제작하였다.

3.3.2 Google Cloud Natural Language API

입력 받은 자연어 처리를 위해서 웹을 통해 입력받은 문장을 Google Cloud Natural Language API를 이용하여 처리했다. Google Cloud Natural Language API의 기능중 Analyzing Syntax를 사용했으며, 이를 이용해 문장의 구성 성분을 분석할 수 있다. 이를 이용하여 문장의 주체가 되는 부분과, 찾고자 하는 속성을 나타내는 단어를 찾고. 해당 단어들을 바탕으로 Sparql 질의 문장을 작성한다. 작성된 Sparql 질의 문장을 DBpedia에서 제공하는 Sparql End-point에 보내어 실행된 결과를 얻고 이를 바탕으로 하여 웹 페이지를 구성하여 사용자에게 전송해 결과를 표시해 준다.

4. 결론

본 논문에서는 시맨틱 웹을 활용하는 링크드 오픈 데이터의 일종인 DBpedia를 활용하는 웹 애플리케이션으로서, 링크드 오픈 데이터를 비전문가도 별도의 학습 필요 없이 쉽게 접할 수 있도록 검색을 도와주는 애플리케이션을 개발하였다. 이를 통하여 사용자에게 다양한 시맨틱 웹에서 얻은 다양한 정보를 제공할 수 있으며, 이와 연결된 다양한 데이터를 접할 수 있는 시작점을 제공하게 된다. 또한 본 애플리케이션을 최근 다양하게 등장하고 있는 음성인식 서비스와 결합하여 음성을 텍스트로 변환하여 얻은 텍스트를 본 논문에서 사용한 DBpedia 검색 기능에 대입하여 다양한 응용을 기대해 볼 수 있을 것이다.

참고문헌

[1] 이미경, "연구개발 전략 수립자원을 위한 테크놀로지 인텔리전스 서비스," 정보과학회논문지, Vol. 17, No. 5, pp. 337-341, 2011

[2] 김창수, "시맨틱 웹 기반 사용자 중심 검색시스템에 관한 연구," 한국정보통신학회논문지, Vol. 19, No. 4, pp. 871-876, 2015

[3] Y. H. Noh, "A Study on Configuring dCollection as the Linked Data," Journal of Korean Library and Information Science Society, Vol. 43, No. 2, pp. 247-271, 2012

[4] 이용주, "링크드 오픈 데이터를 활용한 시맨틱 모바일 메쉬업," 한국정보기술학회논문지, Vol. 14, No. 11, pp. 93-100, 2016