링크드 데이터 기반 대구 맛집 차트

정은미, 전은구, 이찬준, 이용주
경북대학교 IT대학 컴퓨터학부
e-mail:jeunmi021@gmail.com, wjsdmsrn93@naver.com,
ckstmd1946@naver.com, yongju@knu.ac.kr

Charts of Famous Restaurants in Daegu based on Linked Data

Eunmi Jung, Jeon Eun Koo, Lee Chan Jun, Youngju Lee
School of Computer Information, Kyungpook National University

요 약

웹의 발달로 많은 양의 데이터를 손쉽게 접할 수 있지만, 이러한 데이터들로 얼마나 의미 있는 정보를 잘 끌어내어 공개하고 얼마나 잘 활용시키느냐가 중요한 이슈가 되었다. 본 연구에서는 각각의 자원들이 연결된 데이터 중심의 웹을 구성하기 위해 대구시에서 제공하는 공공데이터를 이용하여 링크드 데이터를 구축한다. 수집한 데이터에서 제공하는 정보를 바탕으로 맛집에 대한 온톨로지를 구축하여 데이터를 발행하고, SPARQL을 활용한 간단한 웹 어플리케이션을 구현한다.

1. 서론

우리에게 친숙한 웹은 전세계 대학과 연구소에 흩어진 물리학자들 간의 공동연구에 필요한 즉각적 정보교환 방안으로 1989년 팀 버너스리(Tim Berners-Lee)에 의해 시작되었지만 오늘날 웹이 없는 세상은 상상하기 힘들다[1]. 이러한 웹의 발달로 어마어마한 양의 데이터를 손쉽게 접할 수 있게 되었지만, 이제는 이러한 데이터 사일로에서 얼마나 의미 있는 정보를 잘 끌어내어 공개하고 얼마나 잘 활용시키느냐가 중요한 이슈가 되었다[2].

링크드 데이터는 웹을 구성하는 데이터들 간의 연결을 목표로 기존의 문서 중심의 웹이 아닌 각각의 자원을 대상으로 상호 연결된 데이터 중심의 웹을 구성하는 것이다. 링크드 데이터는 사실 데이터를 포함하는 데이터 객체에 URI를 부여하고 이를 웹 프로토콜인 HTTP를 통해 발행하여 누구나 웹상에서 자유롭게 데이터를 활용할 수 있게 하는 기술이라 할 수 있다[3]. 즉 링크드 데이터를 통해 기존 문서 위주의 World Wide Web 전달 방식을 페이지가 아닌 데이터 간 연결 중심으로 전환하여 보다 풍부한 자원의 생산 및 효율적인 활용이 가능한 방식으로 웹을 지능화시키는 것이다[4].

기존의 연구들이 링크드 데이터 구축과 SPARQL 질의 처리 방안에 초점을 두었다면, 본 연구는 대구시 공공 데이터를 시범 데이터로 선정하여 구축한 링크드 데이터와 SPARQL을 활용한 웹 어플리케이션을 구현하고자 한다.

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 온톨로지(Ontology)

온톨로지란 사람들이 세상에 대하여 보고 듣고 느끼고 생각하는 것에 대하여 서로 간의 토론을 통하여 합의를 이룬 바를, 개념적이고 컴퓨터에서 다룰 수 있는 형태로 표현한 모델로, 개념의 타입이나 사용상의 제약조건들을 명시적으로 정의한 기술이다. 온톨로지는 일종의 지식표현(knowledge representation)으로, 컴퓨터는 온톨로지로 표현된 개념을 이해하고 지식처리를 할 수 있게 된다.

온톨로지는 시맨틱 웹을 구현할 수 있는 도구로서, 지식개념을 의미적으로 연결할 수 있는 도구로서 RDF, OWL, SWRL 등의 언어를 이용해 표현한다. 온톨로지의 구성 요소는 클래스(class), 인스턴스(instance), 관계(relation), 속성(property)으로 구분할 수 있다[5].

이러한 온톨로지를 기술하는 언어 중 가장 대표적인 것은 1999년 W3C 권고안으로 채택된 RDF와 2004년 W3C 권고안으로 채택된 RDF Schema, OWL(Web Ontology Language)등이 있다. 그 중 OWL은 다른 마크업 언어들보다 더 많은 의미 표현 수단을 제공하므로 표현력과 추론 능력이 뛰어나다고 평가받는다[6].

2.2 링크드 데이터

위키피디아에 의하면, 링크드 데이터는 웹상에 존재하는 데이터를 개별 URI로 식별하고, 각 URI에 대한 링크 정보를 부여함으로써 상호 연결된 웹을 지향하는 모델로 정의한다. 이러한 링크드 데이터를 통해 사람이 이해하고 활용하는 문서 중심의 웹을 기계 또한 이해하고 자동으로 처리할 수 있는 데이터 중심의 웹을 구축할 수 있다[7].

기술적으로 링크드 데이터의 핵심 아이디어는 HTTP URI의 사용이다. URI는 웹 문서들을 식별하는 것뿐만 아니라 임의의 실세계 객체들을 식별할 수 있다. 링크드 데이터를 구축하기 위해서는 비구조적인 데이터를 구조화하는데 시맨틱 형태로 표현해야 한다. 여기서 자원을 구조화한다는 것은 데이터를 RDF 형태로 표현하는 것이다[8].

RDF(Resource Description framework)[9]는 웹 상의 데이터를 교환하기 위한 표준 모델로서, 주어(subject), 서술어(predicate), 목적어(object)의 트리플(triple) 구조로 정보를 표현한다. RDF는 사람이 쉽게 읽고 이해할 수 있을 뿐만 아니라 기계적인 처리, 즉 응용프로그램들이 웹에 표현된 정보들을 처리하기 용이하여, 다양한 어플리케이션 영역에서 사용될 수 있다[10].

2.3 SPARQL

SPARQL(SPARQL Protocol and RDF Query Language)은 웹표준화기구인 W3C에서 개발한 RDF 질의 언어이며, 웹에 공개되어 있는 RDF 데이터를 검색하기 위해 사용된다. 관계형 데이터베이스 질의 언어인 SQL과 유사한 형태를 지니며, PREFIX, {SELECT, ASK, DESCRIBE, CONSTRUCT}, WHERE로 구성된다. 여기서 PREFIX는 질의를 통해 검색할 데이터 셋을 지정할 때 사용하며, URI를 간단하게 표현할 수 있다. ASK는 질의 결과가 존재하는지 여부를 boolean 형태(true, false)로 나타내며, DESCRIBE는 찾고자 하는 정보에 대해 그 정보와 연결된 모든 트리플을 반환하고, CONSTRUCT는 질의 결과를 사용자가 원하는 트리플 형태로 반환해준다[11].
다음 코드는 간단한 SPARQL 질의의 예이다.

```
PREFIX ex: <http://example.com/>
SELECT *
WHERE {
        ?s ex:name ?o.
}
LIMIT 10
```

SQL과 유사한 형태를 보이는 이 질의는 해당 SPARQL 서버에 있는 데이터(RDF Triple) 중 ?s와 ?o의 관계가 ex:name인 데이터를 조회하여 모두 보여준다. 질의에서 사용되는 변수는 “?변수명”으로 표현한다. 마지막 행의 LIMIT 10은 조회한 모든 데이터 중 처음 10개만 반환해 달라는 의미이다[12].

3. 어플리케이션 구현

3.1 온톨로지 구축

대구시에서 제공하는 공공데이터를 활용하여 링크드 데이터를 구축해 보며, 이를 위해 공공데이터의 온톨로지를 구축한다.

1) 데이터 선정 및 수집

대구시에서 제공하는 여러 공공데이터들 중 활용도가 높은 데이터를 선정하고자 하였으며, 이에 따라 활용도가 높은 기준을 조회수로 두어 대구맛집 정보를 구축 데이터로 선정하였다. 제공하는 데이터 형태는 Sheet(XLS, CSV, TXT), Open API 두 가지이며, 본 연구에서는 Sheet 데이터를 사용하였다.

2) 데이터 모델링

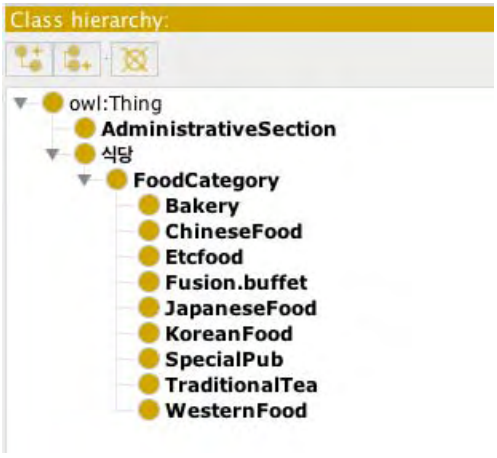
대구 맛집에 대해 제공하는 데이터 요소는 <표 1>과 같으며, 맛집에 대한 정보가 구체적으로 기술되어 있다. 구군 분류와 업소명 외에 구체적인 주소에 대한 정보가 없는 점이 아쉬웠다[13].

<표 1> 대구맛집 정보 데이터 제공 항목

구군분류	음식분류	업소명
전화번호	영업시간	좌석수
주차장	홈페이지	가능 외국어
예약여부	유아시설	좌식유무
후식유무	메뉴	간단설명
지하철	버스	

Protégé라는 오픈소스 온톨로지 제작툴로 OWL(Web Ontology Language)을 이용하여 온톨로지를 구축하였다.

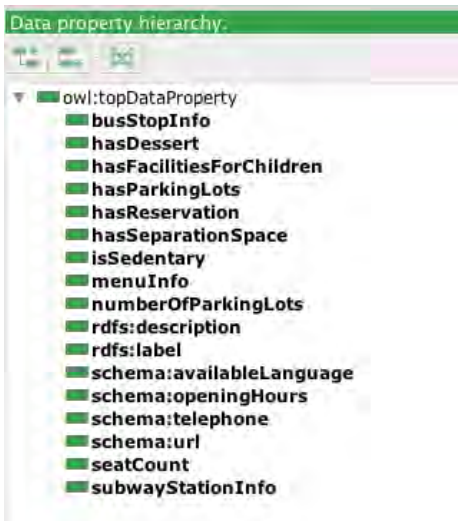
(a) 클래스 정의: 대구 맛집이 각각 고유한 URI를 가지는 하나의 객체이므로, 이러한 객체를 식당이라는 클래스로 정의하였다. 한식, 중식, 양식 등의 음식 분류 항목 또한 고유한 URI를 가지기에 클래스(FoodCategory)로 명명하고, 식당의 하위 클래스로 정의하였다. 마찬가지로 중구, 수성구 등 행정구역을 나타낸 구군 분류 항목도 클래스(AdministrativeSection)로 정의하고 다른 외부 LOD의 정보를 활용한다.



(그림 1) 대구 맛집에 대한 클래스

(b) 속성 정의: 속성은 Object property(클래스간 관계를 나타내는 속성)과 Datatype property(클래스에 대한 특성을 나타내는 속성) 두 가지로 나뉘며, 아래와 같이 정의하였다.

● Object property
– locatedIn: 식당 클래스와 AdministrativeSection 클래스의 관계
– restrauntType: 식당 클래스와 FoodCategory 클래스의 관계
● Datatype property



(그림 2) 정의한 Datatype Property 모습

3.2 링크드 데이터 발행

D2RQ를 이용하여 앞서 구축한 온톨로지 모델을 바탕으로 링크드 데이터를 발행한다. 관계형 데이터베이스의 데이터를 RDF 데이터로 변환하기 위해 D2RQ Mapping Language를 이용하여 맵핑규칙을 작성하고 이를 통해 링크드 데이터를 발행한다.

3.3 SPARQL 질의

웹 어플리케이션 내 대시보드 표현을 위한 SPARQL 질의문을 작성한다. 링크드 데이터 중 차트로 표현할 정보와 질의문 예는 아래와 같다.

1) 대구시의 행정구별 식당수

```
                --PREFIX 생략 --
1  SELECT (?gu as ?행정구역) (count(DISTINCT ?s)
2  as ?식당수)
3  WHERE
4  {
5   { ?s kdp:locatedIn ?gu FILTER
6  regex(str(?gu),"Nam-gu") }
7                    ...중략...
8  UNION
9   {  ?s kdp:locatedIn ?gu FILTER
10 regex(str(?gu),"Dong-gu") }
11  }
12 GROUP BY ?gu
13 ORDER BY asc(?gu)
```

2) 대구시의 음식분류별 식당수(양식, 한식 등)

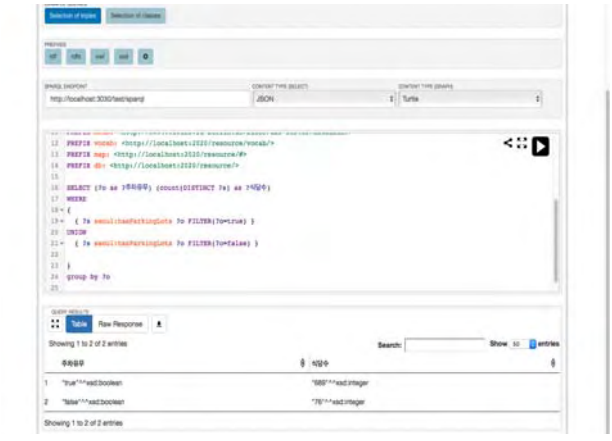
```
                --PREFIX 생략 --
   SELECT (?o as ?음식분류) (count(DISTINCT ?s) as ?식당수)
1  WHERE
2  {
3    { ?s ex:restrauntType ?o FILTER(?o="KoreanFood") }
4                        ...중략...
5    UNION
6    {  ?s ex:restrauntType ?o FILTER(?o="WesternFood") }
7
8  }
   GROUP BY ?o
```

3) 주차 가능 여부에 따른 식당수

```
                --PREFIX 생략 --
1  SELECT (?o as ?주차유무) (count(DISTINCT ?s) as
   ?식당수)
2  WHERE
3   {
4    { ?s seoul:hasParkingLots ?o FILTER(?o=true) }
5    UNION
6    { ?s seoul:hasParkingLots ?o FILTER(?o=false) }
7  }
8  GROUP BY ?o
```

4) 좌석수별 식당수(좌석수가 201~400일 때)

```
                --PREFIX 생략 --
1  SELECT ?s (?o as ?좌석수)
2  WHERE
3  {
4    { ?s kdp:seatCount ?o FILTER(?o>=201 && ?o<=400) }
5  }
6  ORDER BY asc(?o)
```



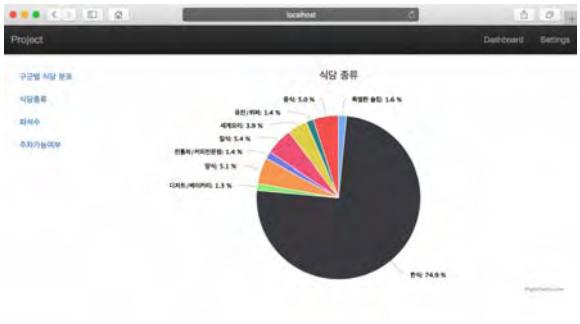
(그림 3) 주차 가능 여부에 따른 식당수 질의 결과 모습

(그림 3)은 SPARQL Endpoint에 대구시의 주차 가능여부에 따른 식당수를 질의한 결과이다.

3.4 구현 결과

앞서 구축한 링크드 데이터 저장소의 Endpoint로 SPARQL 질의를 통해 원하는 데이터를 검색한다. (그림 4)는 행정 구역별로 식당수를 질의하여 대구시 지도 모양으로 그 결과를 나타낸 모습이다. 이외에도 좌석수별 수용가능 좌석수는 세로 차트, 주차가능여부에 따른 식당수와 음식분류별 식당수는 파이 차트로 그 결과를 나타내었다.

(그림 4) 행정구별 식당수 분포 지도 그래프

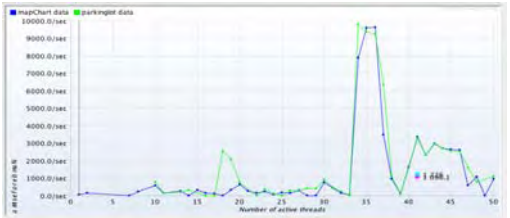


(그림 5) 음식분류별 식당수

3.4 성능 테스트

구현한 대시보드에 대해 Apache Jmeter를 이용하여 부하테스트를 진행하였고, 조건은 아래와 같다.

● 조건 : 5초 간격, 사용자 10명 생성(최대 50명).
상태 100초간 유지 후, 10초 간격으로 10명씩 제거



(그림 6) 사용자 변화에 따른 초당 처리 건수

그림 6과 같이 사용자가 30명 이상일 때, 처리 건수가 급격하게 증가한 것을 알 수 있다. 사용자가 최대 100명일때도 사용자가 60명이상 구간에서 비슷한 결과를 얻었다. 시간상으로는 테스트를 시작하고 30초 이후에 위와 같은 현상이 발생한다. 향후 성능 향상을 위해 질의문과 결과 데이터 처리 기능에 대한 보완이 필요하다.

4. 결론

현재 국내에서는 공공데이터 포탈을 통해 국가서지, 생물 등 여러 공공데이터를 링크드 오픈 데이터로 제공하고 있으나 이러한 링크드 데이터를 활용한 사례는 많지 않다. 본 연구에서는 대구시에서 공개한 공공데이터 중 대구 맛집에 대해 온톨로지를 구축하고 이를 바탕으로 링크드 데이터를 구축하였다. 구축된 데이터의 SPARQL Endpoint에 SPARQL로 질의하고, 그 결과를 여러 형태의 차트로 나타낸 간단한 웹 어플리케이션을 개발하였다. 이번 연구에서는 맛집 데이터에서 제공하는 정보에 한해 차트로 구현했다면, 이 밖에 활용할 수 있는 다양한 데이터를 수집하고 연계하여 보다 유용한 정보, 예컨대 음식점 주변 에 있는 주차장 정보를 지도에 함께 보여주는 것도 향후 연구 주제로 좋을 것이다.

참고문헌

[1] 이재호·양정진, “시맨틱 웹 : 차세대 지능형 웹 기술,” TTA저널 제81호, 2002, 79-85

[2] 윤소영, “공공데이터 활용을 위한 링크드 데이터 국가 연계체계 구축에 관한 연구,” 정보관리학회지, 2013, 30(1), 259-284

[3] 조명대·오원석·박진호, Linked Data 연구개발보고서: 주제명, 저자명 전거데이타 중심, 국립중앙도서관, 2011

[4] 이용주, “링크드 데이터: 빅데이터 구축의 핵심 플랫폼,” 한국디지털경영학회 2014년 춘계학술발표대회논문집, 2014, 237-244

[5] 위키피디아, https://ko.wikipedia.org/wiki/온톨로지

[6] 허승록·임진희·이해영, “오픈소스 도구를 이용한 기록정보 링크드 오픈 데이터 구축 절차 연구”, 정보관리학회지, 2017, 34(1), 341-371

[7] 이병하 외 4인, 알기 쉬운 Linked Open Data, 한국정보화진흥원, 2015

[8] 이용주, "링크드 데이터 구축 및 검색 기법," 한국정보처리학회 2014년 추계학술발표대회논문집, 21(2), 2014, 1057-1060

[9] https://www.w3.org/RDF/

[10] 황석형·조동헌, “형식개념분석법을 이용한 링크드 오픈 데이터 클라우드의 RDF데이터 분석,” 한국컴퓨터정보학회논문지, 22(6), 2017, 57-68

[11] SPARQL 1.1 Query Language, https://www.w3.org/TR/sparql11-query/

[12] 현은석, “링크드 데이터 관점의 빅데이터와 공공데이터,” 디지틀도서관 73(단일호), 2014, 89-106

[13] 정은미, 이용주, "대구시 공공 링크드 데이터 구축 사례," 한국정보처리학회 2017년 추계학술발표대회 논문집, 24(2), 2017, 978-981