도메인 특화 Knowledge Graph 구축 및 활용 기술 개발 : 경북대학교 정보 검색 시스템

*이인성 **임광환 ***윤대성 ****정수현 *****송다은 ******이동욱 ******백호기

경북대학교

*leeinsung12@gmail.com **ghlim100@naver.com ***윤대성 ****jshinkor00@naver.com

*****ssongda0502@naver.com ******dwlee@datastreams.co.kr *******neloyou@knu.ac.kr,

Development of domain-specific knowledge graph construction and utilization technologies : Kyungpook National University Information Search System

*Lee In Sung, **Lim Gwang Hwan, ***Yoon Dae Sung, ****Jeong Su Hyeon,

*****Song Da Eun, ******Lee Dong Wook *******Baek Ho Ki

Kyungpook National University Department of Computer Science

요약

글로벌 빅테크 기업들의 주도하에, 초거대AI (Hyper-Scale AI)에 대한 경쟁이 이루어지고 있으나, 초거대AI 학습 및 운영은 전 지구적으로 손에 꼽힐만한 기관/기업에서만 가능하다. 여전히 많은 기업/기관은 자신들만의 특화된 영역에서의 AI기술 적용이 필요하나, Transformer 기반으로 학습된 Model은 의미있는 결과 및 정확도를 도출하기에는 어렵다. 본 프로젝트는 각 기업/기관들의 특화된 영역들의 데이터를 기반으로, Knowledge Graph를 구축하는 기술을 연구 및 개발한다. 또한 특정 도메인을 경북대학교로 설정하여 경북대학교 학생에게 필요한 정보들이 있는 사이트를 통한 경로와 세부사항을 질의 및 응답하는 서비스를 구축하고 이를 통해 특정 도메인의 Kneowledge Graph를 통한 결과 및 기대효과를 연구한다.

1. 서론

지식 그래프(Knowledge Graph)는 개별 객체의 데이터를 나타내는 정점과 객체 간의 연관성을 의미하는 간선으로 표현할 수 있는 그래프 형태의 저장된 지식을 의미한다. 지식 그래프는 컨텍스트 내의 링킹 및 의미론적인 메타데이터를 저장하며, 이러한 방식으로 데이터 통합, 통일, 분석, 공유(integration, unification, analytics, sharing)를 위한 프레임 워크를 제공한다. 지식 그래프로 데이터를 표현하는 가장 큰 목적은 그래프 구조가 연관성을 포함한 형태로 지식을 축적하고 전달하는데 가장 유리한 자료구조이기 때문이다.

구글은 지난 2012년 5월 16일 지식 그래프를 구글 영문 검색에 처음 적용하였고, 이를 통해 어떤 토픽에 대해 구조화된 정보와 다른 사이트로의 링크를 제공하여 다양한 소스로부터 축적한 시맨틱 검색 정보를 사용하여 검색 결과를 향상했다. 구글의 지식 그래프는 인터넷상의 각기 다른 출처에서 발견된 이런 사실들을 하나의 개념 지도에 연결된 조각들로 간주한다. 이러한 방식으로 표현된 데이터 네트워크는 훨씬 직관적이다. 뿐만 아니라 주어진 도메인에 있는 각 주체 간의 관계를 명료하게 만들어준다. 이를 통해 사용자는 모든 정보를 맥락에 맞게 접근하고 이해할 수 있다.

지식 그래프는 소비자용 애플리케이션에만 유용한 것이 아니다. 기업은 이 기술을 사용하여 제품, 서비스 및 클라이언트에 대한 데이터를 처리할 수 있다. 심지어 여러 데이터베이스, 스프레드시트, 문서에 데이터가 분산되어 있더라도, 처리가 가능하다.

도메인 특화 지식 그래프는 특정 용도와 응용 프로그램에 사용된다. 지능형 콘텐츠 및 패키지 재사용, 대응 및 상황 인식 콘텐츠 권장, 지식 그래프 구동 약물 검색, 의미 검색, 투자 시장 인텔리전스, 규제 문서의 정보 검색, 고급 약물 안전 분석 등과 같은 데이터 및 정보 집약적인 서비스 등이 있다.

본 연구에서는 특정 도메인을 경북대학교로 설정한다. 경북대학교 학생들은 통합정보시스템 등의 사이트를 통해 '학사행정', '학적', '휴학신청' 등의 탭을 선택하여 원하는 정보를 조회하거나 필요한 것을 신청할 수 있다. 하지만 어떤 정보를 어느 카테고리를 통해 들어가야 얻을 수 있는지 알 수 없는 경우가 많고, 하나의 정보가 여러 군데로 흩어져 있어 원하는 정확한 정보를 찾을 때 많은 시간이 소요된다.

이를 편리하게 이용할 수 있도록 경북대학교 정보검색시스템을 개발한다. 경북대학교 학생들에게 필요한 정보들의 검색 경로와 세부사항을 나타낼 수 있는 지식 그래프를 구축하고 웹을 통해 질의응답이 가능하도록 하여 도메인 특화 Knowledge Graph의 구현 방법과 결과 및 기대효과를 연구하는 것이 목표이다.

2. 본론

2.1. 지식 그래프(Knowledge Graph) 설계

2.1.1 지식 그래프 설계 절차

지식 그래프는 여러 데이터 소스로부터 수집한 데이터를 분석하여 구성한다. 문서와 같은 텍스트 데이터로부터 자연어 분석 기술을 이용하여 텍스트에 포함된 내용을 정점과 간선의 형태로 변환하여 축적한다. 검색을 고도화하고 사용자가 원하는 정보를 관련 있는 정보들과 함께 일목요연한 구조로 전달하기 위해 지식 그래프를 구축하는 것이 필수이다.

지식 그래프를 구축하기 위한 기본 설계 절차는 다음과 같다. 우선 지식 그래프 데이터의 사용에 따라 지식 구조를 설계한다. 그 후, 입력된 학습데이터를 처리해야 하나, 대부분 데이터가 텍스트 형식이므로 AI 기반 자연어 처리와 태깅 작업을 거친다. 텍스트 데이터 이외의 기타 데이터는 지식으로 표현할 수 있도록 전처리를 거치고 데이터에 대한 구조화를 진행한다. 데이터의 구조를 바탕으로 모델링을 거친다. 모델링을 통해 데이터 간 연결 관계를 적재한다. 이렇게 형성된 지식 그래프를 조회하고 서비스하는 시각화 단계를 진행한다.

2.1.2. 지식 구조 설계

지식은 축적된 분야에 따라서 개념과 용도가 달라, 지식 그래프를 구성하는 정점과 간선은 분야에 따라 다른 내용이 포함되어야 한다. 지식 그래프를 활용하여 최종적으로 제공하는 서비스에 따라 활용할 정점과 간선의 종류를 구체적으로 결정하는 과정을 지식 설계라고 한다. 정점과 간선은 노드와 에지로도 표현할 수 있다.

경북대학교 정보검색시스템을 위한 지식 그래프를 구축하기 위해 노드와 에지의 종류를 결정한다. 노드와 에지는 다음과 같이 분류한다.

종류	설명	node /edge	예
경북대	모든 노드의 중심 노드	node	경북대학교
사이트	경북대학교의 정보가 있는 사이트 노드	node	통합정보시스템, 수강신청사이트 등
앱	경북대학교의 정보가 있는 앱 노드	node	KNUPIA, 경북대학교도서관 등
구분	사이트나 앱을 통해 접속하여 얻을 수 있는 정보의 카테고리 노드	node	학사행정, 학적, 학적변동관리, 휴학신청 등
세부사항	직접 사이트에 들어가지 않아도 간단하게 알려줄 수 있는 정보들이 입력된 노드	node	
상세정보	여러 곳에 흩어져 있는 정보들의 차이를 알 수 있도록 구체적인 구분 정보들이 입력된 노드	node	
다음 경로	모든 노드들의 관계	edge	

경북대 노드를 중심으로 경북대 학생에게 필요한 정보가 주로 있는 '사이트'와 '앱' 두 가지 종류의 노드로 나누어 생성했다. 각 사이트와 앱을 통해 얻을 수 있는 정보가 있는 탭을 '구분' 노드로 두고, 세부적인 정보가 입력되어 있는 노드를 '세부사항'으로, 여러 곳에 흩어져 있는 데이터들을 비교하여 차이점을 알려주는 노드는 '상세정보'로 분류하였다. 경북대학교 정보검색시스템의 지식 그래프는 경로를 중점적으로 구축하기 때문에 노드들 간의 관계인 에지는 '다음 경로'로 통일한다.

2.1.3 데이터 입력

지식 그래프는 <h,r,t>의 트리플렛 형태로 표현되며 h,t는 객체이고 r은 둘의 관계를 나타낸다. 트리플로 이루어진 지식 그래프에서 요소 간의 관계를 relation을 통해 간단하고 직관적이게 알 수 있다. 특정 도메인의 Knowledge Graph를 구축하기 위해서는 문서 형태로 들어온 해당 기업의 데이터를 트리플렛 형태로 수집하여 AI기반 자연어 처리와 태깅 작업을 거쳐서 구조화를 진행해야 하지만, 연구 기간이 짧고 구현의 어려움이 있으므로 활용하고자 하는 도메인의 데이터를 직접 입력한다.

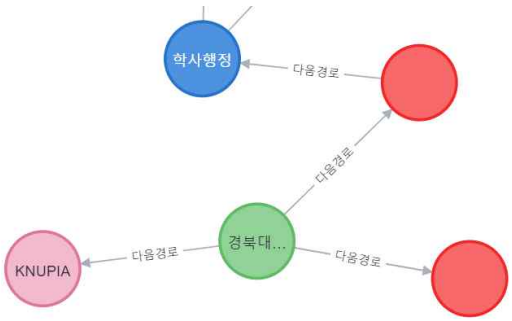
2.1.4. 그래프 데이터 모델링

그래프 데이터베이스는 관계형 데이터베이스의 한계를 해결하기 위해 만들어진 NoSQL 데이터베이스의 한 유형으로, 노드와 엣지의 형태로 표현할 수 있는 데이터를 저장하기 위해 특화된 데이터베이스이다. 그래프 데이터베이스의 질의 결과는 그래프로 추출 될 수 있기 때문에 사용자에게 결과의 추출 과정이나 결과를 시각적인 방법으로 제공할 수 있다. 그래프 데이터베이스 시스템인 neo4j를 통해 경북대학교에 대한 정보와 경로를 알려주는 데이터를 입력하여 지식 그래프를 구축한다. neo4j를 통해 구축하고자 하는 도메인의 데이터를 cypher 언어로 직접 입력한다. 입력 형태는 다음과 같다.

```
CREATE (knu:경북대 {Name : "경북대학교"})
CREATE (knupia:앱 {Name : "KNUPIA"})
CREATE (knuin:사이트 {Name : "통합정보시스템"})
CREATE (sugang:사이트 {Name : "수강신청사이트"})
CREATE (학사행정:구분 {Name : "학사행정"})
CREATE (knu)-[:다음경로]->(knuin)
CREATE (knu)-[:다음경로]->(sugang)
CREATE (knu)-[:다음경로]->(knupia)
CREATE (knuin)-[:다음경로]->(학사행정)
```

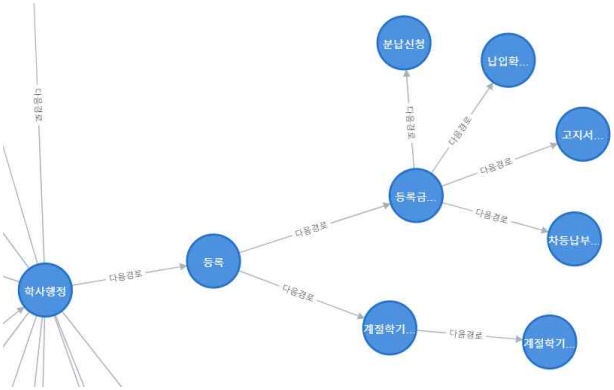
위의 쿼리문은 Name이 경북대학교, KNUPIA, 통합정보시스템, 수강신청사이트, 학사행정인 노드를 만들고 노드들 간의 관계를 설정한 것이다. 경북대학교의 다음 경로 관계에 수강신청사이트, 통합정보시스템, KNUPIA 노드가 있고, 통합정보시스템의 다음 경로 관계에 학사행정 노드가 있다.

다음은 neo4j를 활용하여 구축한 지식 그래프로 위 쿼리문을 통해 만들어진 지식 그래프의 일부이다.



<그림 1. 쿼리문 결과>

아래는 위와 같은 쿼리문을 통해서 만든 지식 그래프의 일부이다.



<그림 2. 분납신청 지식 그래프>

통합정보시스템 사이트에 있는 데이터를 쿼리문으로 작성하여 위와 같은 지식 그래프를 구축하였다. 통합정보시스템 사이트->등록->등록금관리->분납신청의 경로를 통해 등록금 분납신청이 가능하다는 것을 알 수 있다.

2.2. 연구 설계 절차

본 연구는 다음의 설계 절차를 따라 진행하도록 한다. 우선 neo4j를 통해 cypher 언어로 쿼리문을 작성해 지식 그래프를 구축한다. 지식 그래프의 구축이 완료된다면, 도메인에 특화된 서비스를 제공한다. 웹 서버는 입력된 텍스트를 지식 그래프 엔진 및 서비스 엔진에 전달한다. 지식 그래프 엔진에서는 Graph DB인 neo4j와 연동하여 전달받은 데이터로 지식 그래프로의 query 등을 처리한다. 웹 프론트엔드에서는 입력받은 데이터의 정보가 담긴 지식 그래프를 보여준다.

2.3. 개발환경

사용할 툴은 neo4j이며 사용 언어는 Java이다. 웹 프론트엔드 개발 도구는 html, css, javascript, react, Node.js, Vue.js 이다. 웹 백엔드 개발 도구는 spring boot, dJango, Flask이다.

3. 결론

본 논문에서는 도메인 특화 지식 그래프를 구축하는 방법을 조사하고 이를 통해 얻을 수 있는 기대효과를 알아보기 위해 경북대학교를 특정 도메인으로 설정하고 경북대학교 정보검색시스템 구현 방법에 대해 연구하였다. 경북대학교 정보 검색 시스템을 통해 경북대학교 학생으로서 필요한 정보들을 검색했을 때 어느 곳에서 어떤 정보를 얻을 수 있는지를 쉽게 파악할 수 있으며, 간단한 정보들은 검색만으로도 편리하게 찾을 수 있다. 통합정보시스템 사이트, KNUPIA 앱 외에도 학생들이 필요로 하는 다양한 정보들의 경로를 입력하여 지식 그래프의 확장이 가능하다.

지식 그래프는 기존의 테이블 기반 데이터베이스가 명시적으로 정의되지 않은 관계들에 대해 연관성을 찾기 어려운 것과는 달리, 입력된 노드들의 연결을 활용해 직접 정의하지 않은 관계에 대해서도 추론 가능하며 이를 통해 새로운 지식의 지속적인 확장이 가능하며 더 자세하고 정확한 정보를 제공해 줄 수 있다. 또한, 지식 그래프를 이용하면 사용자의 명령에 대해서 AI가 '의도'를 파악할 수 있어 보다 정확하게 사용자가 원하는 결과물을 생성해낼 수 있다.

이러한 점들로 인해, 보다 적은 데이터로도 유의미한 결과와 정확도를 보여주는 AI 개발을 기대할 수 있으며 특화된 영역의 데이터로 구성된 지식 그래프는 각각의 기업/기관들의 자신들만의 특화된 영역에서의 AI 기술을 개발할 수 있을 것이다.

기존의 관계형 데이터베이스가 아닌 영역 특화된 지식 그래프를 기반으로 사용하는 것만으로도 지식 그래프가 Data Access 부분과 데이터의 추가/삭제/변경 부분에 있어서도 보다 높은 성능을 보이기에 방대한 정보를 활용해야 하는 서비스를 제공하기에도 알맞을 것이다. 영역 특화된 지식 그래프는 각각의 영역에서 서비스를 구축할 때, 보다 전문적이고 효과적인 서비스를 만드는 밑바탕이 될 것이다.

입력되는 학습데이터를 분석 및 처리하여 지식 그래프를 생성하는 알고리즘도 개발한다면 앞서 설명한 지식 그래프의 장점을 살려 도메인특화 그래프를 다양한 분야에서 활용하고 사용자 입장에서 데이터를 편리하게 이용할 수 있게 될 것으로 기대된다.

ACKNOWLEDGMENT

Put sponsor acknowledgments.

참 고 문 헌

[1] Davies R. W." The Data Encryption standard in perspective,"Computer Security and the Data Encryption Standard, pp. 129-132.

[2] Miles E. Smid, "From DES to AES," 2000, (http//www.nist.gov/aes).

[3] Shamir, A. "On the security of DES," Advances in Cryptology, Proc.Crypto '85, pp. 280-285, Aug. 1985.

[4] NIST, "Announcing the Advanced Encryption Standard(AES),"FIPS PUB ZZZ, 2001, (http://www.nist.gov/aes).

[5] Daemen, J., and Rijmen, V. "AES Proposal: Rijndael, Version2.," Submission to NIST, March 1999.