Journal of The Korea Society of Computer and Information
Vol. 22 No. 2, pp. 97-104, February 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.02.97

SPARQL Query Automatic Transformation Method based on Keyword History Ontology for Semantic Information Retrieval

Dae Woong Jo*, Myung Ho Kim**

Abstract

In semantic information retrieval, we first need to build domain ontology and second, we need to convert the users' search keywords into a standard query such as SPARQL.

In this paper, we propose a method that can automatically convert the users' search keywords into the SPARQL queries. Furthermore, our method can ensure effective performance in a specific domain such as law. Our method constructs the keyword history ontology by associating each keyword with a series of information when there are multiple keywords. The constructed ontology will convert keyword history ontology into SPARQL query. The automatic transformation method of SPARQL query proposed in the paper is converted into the query statement that is deemed the most appropriate by the user's intended keywords. Our study is based on the existing legal ontology constructions that supplement and reconstruct schema and use it as experiment. In addition, design and implementation of a semantic search tool based on legal domain and conduct experiments. Based on the method proposed in this paper, the semantic information retrieval based on the keyword is made possible in a legal domain. And, such a method can be applied to the other domains.

▸ Keyword : Semantic web, Semantic information retrieval, Keyword search, SPARQL Query Transformation, Legal information retrieval

I. Introduction

정보가 많아지고, 종류가 다양해지면서 정보검색에 관한 상업적, 학문적 이슈가 많다. 기존의 정보검색 분야는 대량의 정보를 빠르게 검색하는 연구에서 사용자 중심의 개인화 검색 및 시맨틱 검색의 형태로 연구 분야가 발전하고 있다. 하지만 법령과 같은 전문지식 분야는 여전히 법률을 모르는 사람이 정보검색을 하고, 지식을 얻기가 힘든 영역이다. 최근에는 이러한 법령정보검색에 시맨틱 웹 기술을 이용하여 기존의 한계를 극복하고자 하는 연구가 진행되고 있다[1]. 시맨틱 웹 기술은 차세대 웹을 위해 발전할 수 있는 분야로 꼽히고 있다. 또한 웹 분야뿐만 아니라 일반 데이터를 저장, 처리, 통합하기 위한 측면

• First Author: Dae Woong Jo, Corresponding Author: Myung Ho Kim
*Dae Woong Jo (jodw@ssu.ac.kr), School of Software, Soongsil University
**Myung Ho Kim (kmh@ssu.ac.kr), School of Software, Soongsil University
• Received: 2016. 11. 11, Revised: 2016. 12. 25, Accepted: 2017. 02. 16.
• This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2014R1A1A2058695).

에서도 응용 연구가 가능하다.

시맨틱 웹 기반의 검색방식은 기계가 이해하고, 처리할 수 있는 검색 환경을 뜻한다[2]. 이러한 환경에서는 OWL과 같은 온톨로지로 된 언어를 기반으로 데이터를 처리하고, 변환하여 결과를 나타낸다. OWL(Web Ontology Language) 온톨로지는 기계가 이해할 수 있는 수준의 어휘로 표준화 되어 있기 때문에 기존의 정보검색방식보다 처리 엔진에서 수행할 수 있는 일이 많아져서 의도한 결과를 도출하기에 기존의 방식에서 하지 못하는 장점을 가지고 있는 것으로 평가받고 있다[3].

사용자는 정보검색을 위한 방법으로 특정 키워드를 중심으로 한 검색 방법을 주로 하고 있다. 키워드는 해당 정보를 대표하는 주제어 일 확률이 높기 때문에, 알고 있는 키워드를 검색에 활용 하는 형태이다. 그 외는 사람들이 자주 찾는 키워드와 연관 검색어를 이용한 방법, 혹은 GUI 기반의 트리 및 그래프 기반의 확장 검색 방식과 자연어를 이용한 방식으로 검색을 수행한다[4]. 자연어를 이용한 방식은 문장 형태의 자연어를 기계에 명확하게 전달해야만 원하는 결과를 얻을 수가 있으며, 문장의 형태에 따라서 복잡한 구조를 가지게 되면 검색의 정확도를 보장 받을 수 없다.

본 논문은 시맨틱 웹 기술을 활용한 정보 검색 방법을 제안한다. 제안하는 형태는 법령과 같은 전문지식 영역에서 사용자가 입력한 키워드를 SPARQL[5]과 같은 시맨틱 웹 표준 질의어로 자동 변환하고, 사용자가 의도한 것과 가장 일치되는 매칭 결과를 볼 수 있도록 한다. 매칭 결과의 정확도를 높이기 위해 본 논문은 사용자 키워드 히스토리 온톨로지(KH-Ontology, Keyword History Ontology)를 구축하여 시맨틱 검색을 수행한다. 키워드 히스토리 온톨로지는 사용자가 입력한 키워드를 관리하고 원하는 결과를 찾을 수 있도록 도와준다. 예를 들면, 사용자는 키워드를 입력하고 원하는 결과가 나오지 않을 때 키워드를 하나 더 추가하여 검색을 수행한다. 그래도 나오지 않으면 키워드를 변경 혹은 관련 키워드를 추가하여 기계가 이해할 수 있는 정보들을 많이 줌으로써 검색의 정확도를 높이고자 노력한다. 따라서 이와 같은 행위를 온톨로지로 구축하고, 연관시키는 작업을 거치게 되면 다른 사용자가 이와 유사한 키워드 검색을 할 때 좀 더 좋은 검색 결과를 볼 수 있다.

본 논문에서 제안하는 키워드 기반의 시맨틱 검색을 위한 방법은 다음의 세 가지를 기반으로 적절한 질의문이 작성되도록 한다. 첫째, 키워드와 관련한 온톨로지 그래프의 간선수를 고려한 질의문 작성과 둘째, 키워드와 관련한 클래스의 인스턴스 수를 고려한 질의문 작성 그리고 셋째, 키워드의 개수가 여러 개일 때 각각의 키워드를 일련의 정보로 저장하고 연관성이 높은 정보로 유추하여 처리하는 단계를 적용한다. 이와 같은 방식을 본 논문에서는 키워드 히스토리 온톨로지를 적용한 것으로 정의한다. 각 단계를 거쳐 사용자가 의도한 결과와 가장 유사한 형태의 정보를 순위화해서 나타낸다. 제안하는 방법에 사용된 첫째, 둘째는 기존의 키워드 기반의 시맨틱 정보검색[6]에서 제안된 방법을 기반으로 법령 분야에 적용하며, 셋째 키워드 히스토리 온톨로지 방법을 추가적으로 적용한다.

본 논문에서는 법령 기반의 시맨틱 정보 검색을 위해 기존의 법령 온톨로지 구축 연구[7]를 기반으로 온톨로지의 스키마 및 구조를 참고하고, 재구축하여 실험 데이터로 활용한다. 본 논문에서 제안하는 방법을 기반으로 법령과 같은 전문정보 영역에서 키워드 기반의 시맨틱 정보 검색을 가능하게 하며, 다른 영역에서도 이와 같은 방법을 활용할 수 있다.

논문의 구성은 다음과 같다. 2장은 시맨틱 정보 검색을 위한 기존의 방법론에 대해 설명하고, 3장은 본 논문에 적용 가능한 키워드 히스토리 온톨로지에 대한 구축 방법 및 질의 변환 형태의 방법에 대해 설명한다. 4장은 논문에서 제안한 방법을 이용한 구현 및 실험 결과에 대해 설명한다. 마지막 5장에서 결론을 한다.

II. Preliminaries

1. Type of Information Retrieval

기존의 정보검색을 위한 방법으로 N-GRAM 검색은 1음절, 2음절과 같이 각 음절을 색인어로 생성해서 검색 키워드와 매칭 시켜 검색을 한다. 음절단위로 나누기 때문에 높은 재현율(Recall)은 보장하지만 의미가 없는 형태이기 때문에 정확도(Precision)는 떨어진다[8].

형태소 검색은 형태소 단위로 색인어를 생성해서 검색어와 매칭 시켜 검색을 한다. 형태소는 의미를 가지고 있는 최소 단위이기 때문에 N-GRAM 방식보다는 정확도가 높다. 하지만 단순 키워드를 기반으로 추출한 방식이기 때문에 어휘 간의 의미를 파악하기가 힘들다는 단점이 존재한다[8].

시맨틱 검색은 문장 단위로 의미를 분석해서 주어, 술어, 목적어 형태로 주제어를 추출하고, 색인하여 검색어와 매칭 시켜 검색을 하는 방식이다[2]. 찾고자 하는 의미가 명확하다면 기존의 검색 방식보다는 정확도가 높게 나올 수 있다. 기존의 정보검색 방식과 시맨틱 검색 방식을 키워드의 개수를 달리해서 실험한 결과에서 키워드가 2개일 때, 시맨틱 검색 방식의 평균 정확도는 67%, 재현율은 69%였고, 3개의 키워드를 이용했을 때는 정확도는 81.5%, 재현율은 64%로 나타났다. 반면 일반 검색은 정확도 52%, 재현율 59%였고, 3개일 때는 정확도 59%, 재현율은 32%로 키워드가 늘어날 때 현격히 떨어졌다[9]. 실험에서 시맨틱 검색 방식은 키워드가 늘어날수록 좋은 결과를 나타낸다.

2. Information Retrieval based on Semantic Web Technology

2.1 Semantic Information Retrieval

시맨틱 웹 기술을 기반으로 한 정보검색 연구는 정보검색 시

질의어에 포함된 검색 키워드와 온톨로지에 포함된 키워드가 상이한 경우에 찾지 못하는 문제를 해결하기 위해, 어휘 간 상이성 해소를 위한 연구가 있다[10]. 어휘 간 상이성 해결을 위해 온톨로지 리소스를 나타내는 어휘 중 검색 키워드와 의미가 가장 가까운 어휘를 찾아내어 교체해줌으로써 검색의 실패를 줄이고 적절한 검색 결과를 제시한다. 또한, 시맨틱 웹 기술을 이해하는 사람들에 의해 검색을 할 수 있도록 시맨틱 질의 포맷을 기반으로 SPARQL로 변환하기 위한 연구가 있다[11]. 하지만 시맨틱 웹 기술 및 SPARQL 질의문을 이해하지 못하는 일반 사용자는 이용할 수 없다는 단점이 존재한다.

또한, 자연어를 대상으로 형태소 및 구문 분석을 통해 파스 트리를 만들고, SPARQL 질의문으로 변환하기 위한 연구가 있다[12]. 자연어를 이용한 구문분석 방법은 형태소 분석기 및 구문분석기의 성능에 따라 결과의 차이가 생기며, 아직 발전해야할 부분이 많다. 본 논문은 의미를 가지고 있는 형태소를 기반으로 키워드 단위의 검색 방법에 있어서 효과적으로 시맨틱 검색을 위한 질의문으로 변환하기 위한 방법에 관해 제안한다.

2.2 Keyword based Semantic Information Retrieval

시맨틱 정보검색에서 키워드를 기반으로 SPARQL 질의문으로 변환하는 기존의 연구는 길이 기반 방법과 빈도수 기반 방법을 이용해서 랭킹을 매기고, 형식화된 구조 질의로 변환하는 연구가 있다[6]. 길이 기반 랭킹 방법(path length based method)은 질의 그래프의 간선수를 기준으로 순위를 정한다. 질의 키워드를 사용해서 생성한 형식화된 구조 질의 중에서 간선수가 적은 질의가 높은 순위가 된다. 빈도수 기반 랭킹 방법(Popularity based method)은 질의 그래프가 더 많은 양의 정보를 대표할 수 있다면, 가치 있는 질의로 판단한다. 즉, 해당 질의를 구성하는 클래스의 인스턴스 수를 기준으로 사용한다[13]. 예를 들어 질의 {‘project1’, ‘kim’} 키워드를 이용해서 질의한 경우 Table 1과 같이 두 개의 질의 그래프가 그려질 수 있으며 길이 기반 랭킹 방법에서는 두 길이가 같기 때문에 랭킹이 같고, 빈도수 기반 랭킹 방법에서 ‘kim’의 클래스인 ‘Researcher’와 ‘Supervisor’ 중에서 ‘Researcher’ 클래스의 인스턴스가 ‘Supervisor’ 클래스의 인스턴스 수보다 많다면, #1 그래프가 #2의 그래프 질의보다 높은 순위로 책정 받고, #1 기반으로 질의 그래프를 생성하는 방식이다.

Table 1. Example of Query Graph

#1	project1 type Project kim type Researcher
#2	project1 type Project kim type Supervisor

본 논문은 기존에 제시된 길이 기반 랭킹 기법과 빈도수 기반 랭킹 기법 연구를 법령과 같은 도메인에 적용하고 추가적으로 키워드가 여러 개 일 때 각각의 키워드를 정보로 기록하고, 기록된 부분을 온톨로지로 작성하여 도메인 온톨로지와 상호작용할 수 있도록 한다.

III. The Proposed Scheme

1. Process Architecture

1.1 Query Transformation Issues

시맨틱 웹 기술 기반의 정보검색에서는 검색 키워드가 여러 개일 때, 기존의 정보 검색 방식보다 원하는 정보를 찾기가 더 수월하다. 기계가 이해할 수 있는 정보를 많이 줄수록 시맨틱 웹 기반의 엔진에서는 처리 가능한 데이터가 많아져서 각각을 정보로 인식하고 원하는 검색 결과를 도출할 수 있다.

시맨틱 검색 기반의 데이터는 RDF(Resource Description Framework) 방식의 트리플 형태의 주어(Subject, S), 술어(Predicate, P), 목적어(Object, O)에 해당하는 S-P-O 구조를 기반으로 온톨로지 형태의 OWL과 같은 어휘로 기술되어 있다. 기본적인 바탕은 트리플 방식으로 데이터를 표현하고, SPARQL 질의 문 역시 트리플을 구조적으로 찾기 위한 그래프 패턴을 기반으로 작성한다. 그래프 패턴은 BGP(Basic Graph Patterns) 매칭 방법을 기반으로 GGP(Group Graph Patterns), OPG(Optional Graph patterns), AGP(Alternative Graph Patterns), PNG(Patterns on Names Graphs)와 같은 질의 패턴[14]을 만들고 온톨로지에 질의 하는 방식이다. 사용자가 시맨틱 질의 패턴을 이해하고 S-P-O 형식으로 질의문을 작성해서 검색을 할 수 있으면 좋겠지만, 시맨틱 웹 환경을 이해하지 못하는 사용자들도 시맨틱 검색을 할 수 있는 방법이 필요하다.

사용자가 원하는 정보가 예를 들어 “project1을 수행하고 있는 kim이 속한 기관의 이름”을 알고 싶다고 했을 때, 시맨틱 검색으로 이와 같은 정보를 찾기 위해서는 SPARQL 질의문의 형식으로 변환이 되어야 한다. 찾고자 하는 정보를 질의 그래프로 표현 했을 때, Fig. 1과 같은 모습의 형태가 가능하다.

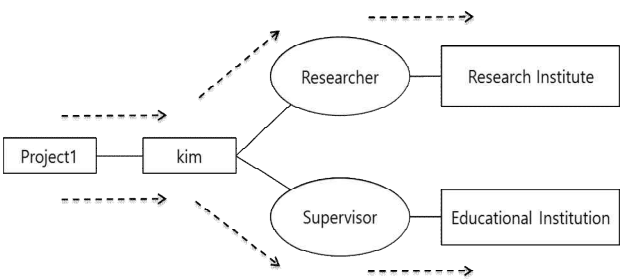

Fig. 1. Flow of Query Transformation

Fig. 1의 질의 처리 구조에 따라 크게 두 흐름의 그래프 생성이 가능하고, “project1”을 수행하고 있는 “kim” 이라는 사

람이 연구자(Researcher) 인 경우의 흐름과 책임자(Supervisor) 인 경우의 흐름에 따라 양 갈래로 나누어지고, 질의 그래프의 방향 또한 바뀌게 된다. 이러한 경우에 그림에서는 위의 화살표 방향으로 갔을 때는 “kim”의 소속 기관의 이름은 “Research Institute”의 결과로 나올 수가 있고, 아래 화살표 방향의 흐름으로는 “Educational Institution”으로 나올 수가 있다. 이때, 기존의 제시된 길이 기반 랭킹 방법과 빈도수 기반 랭킹 방법에 따라 랭킹을 통해 질의 결과가 반영된다고 했을 때, Researcher 클래스의 인스턴스 수가 많다면, 그림에서 위의 화살표 방향의 결과를 보게 된다. 하지만 사용자가 원하는 결과는 “책임자 kim이 속한 기관의 이름” 이였다면 원하는 결과를 만족할 수 없다. 사용자는 원하는 결과를 얻기 위해 “책임자” 라는 키워드를 추가해서 질의를 이어나갈 수 있다. “책임자” 키워드를 기반으로 그림의 흐름은 아래 방향의 그래프 흐름으로 진행되고, 원하는 결과를 얻을 수 있다.

1.2 Flow of processing structure

본 논문이 제안하는 시맨틱 검색 질의 변환을 위한 기본적인 방법은 OWL과 같은 온톨로지 형태의 데이터에 사용자 키워드가 입력으로 들어오고, 입력된 키워드를 기반으로 그래프를 생성한다. 생성된 그래프는 길이 기반 랭킹 방법과 빈도수 기반 랭킹 방법을 이용해서 순위를 정하고, 입력된 키워드를 키워드 히스토리 온톨로지(KH-Ontology)의 인스턴스로 변환, 구축한다. 그리고 다시 키워드 입력이 추가되어 들어오면 키워드 히스토리 온톨로지의 연관클래스의 인스턴스로 변환, 구축해서 입력된 키워드 정보를 관리한다. Fig. 2는 기본적인 처리 구조를 나타낸다.

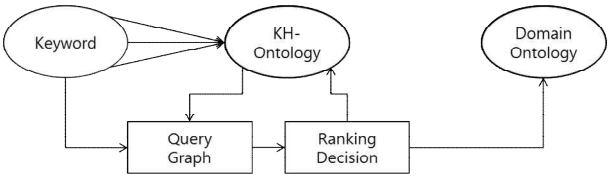

Fig. 2. Processing structure

사용자 키워드는 입력 수에 따라 각각을 분기하고, 키워드 히스토리 온톨로지의 정의된 스키마 룰에 따라 인스턴스 각각이 관계를 맺는다. 입력된 키워드는 이전에 입력된 키워드와의 관계를 기반으로 질의 그래프(Query Graph) 단계에서 질의문이 생성된다. 그다음 순위 결정(Ranking Decision) 과정을 통해 도출된 순위를 기반으로 질의 그래프를 완성하고, 키워드 히스토리 온톨로지에서 정의된 관계와 정보를 반영하고, 최종적인 질의문으로 도메인 온톨로지(Domain Ontology)에 질의한다.

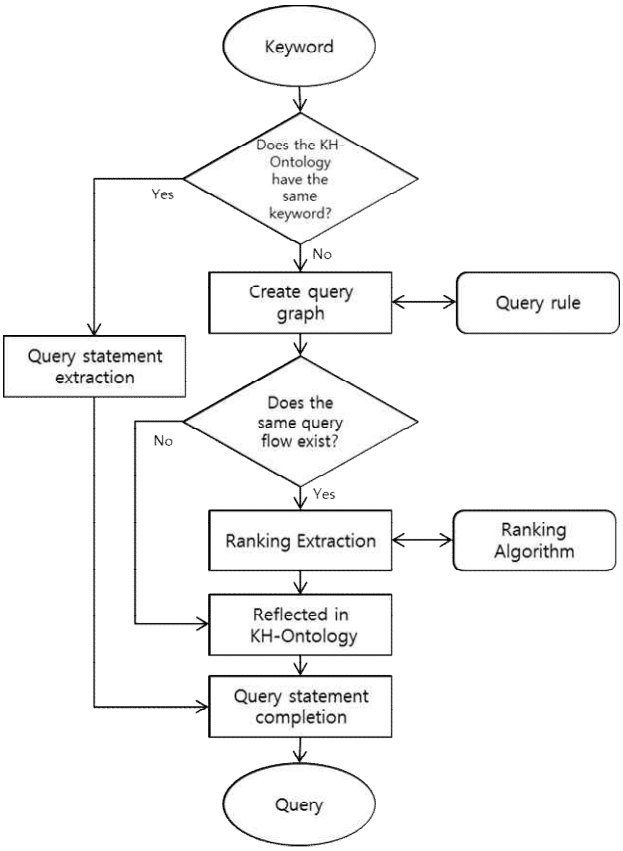

Fig. 3. Query Statement Construction Flowchart

Fig. 3은 처리구조를 기반으로 한 질의 그래프 생성을 위한 구체적인 흐름도를 나타낸다. 흐름도는 입력된 키워드를 기반으로 키워드 히스토리 온톨로지에 동일한 키워드가 있는지 검사한다. 있으면, 키워드 히스토리 온톨로지에 정의된 질의문을 추출하고, 완성된 질의문을 기반으로 도메인 온톨로지에 질의한다. 없다면, 키워드를 기반으로 질의 그래프를 생성하는데 질의 규칙을 이용해서 한다. 질의문 만드는 규칙은 첫째, 키워드와 관련한 온톨로지 정보들을 추출, 둘째 키워드가 속한 클래스 정보 추출을 기반으로 질의문을 작성한다. 작성된 질의 그래프에서 동일한 질의의 흐름이 존재하는 경우는 같은 키워드의 인스턴스 명을 가진 경우에 한해서다. 동일한 인스턴스 명에 따라 흐름이 나누어진다. 이런 경우에는 랭킹을 추출해서 하나의 질의 그래프로 선택하게 된다. 랭킹 추출은 이전에 제시된 길이기반, 빈도수 기반을 토대로 한다. 질의문은 키워드와 함께 키워드 히스토리 온톨로지에 반영하고, 질의문을 완성해서 질의하는 흐름으로 이어진다.

2. 키워드 히스토리 온톨로지

키워드 히스토리 온톨로지는 사용자가 입력한 키워드의 이력을 관리하는 온톨로지다. 키워드는 입력될 때마다 Fig. 4와 같이 K1 클래스부터 Kn 클래스 까지 클래스가 생성 되고, 각 키워드는 클래스의 인스턴스로 연결되어, 관련성 있는 정보로 표시한다. 인스턴스는 해당 키워드에 대한 검색 히스토리를 나

타내게 된다. 검색을 마치는 시점의 키워드를 기준으로 검색된 질의 그래프를 Query 클래스의 인스턴스로 변환하여 이력을 관리한다.

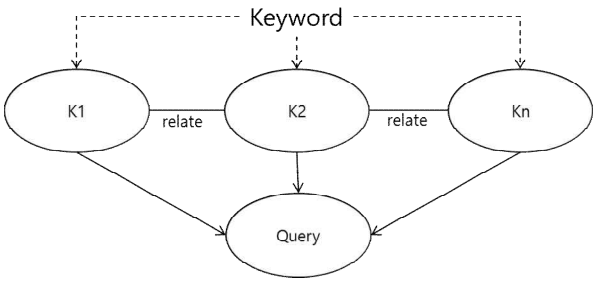

Fig. 4. KH-Ontology

본 논문에서 제안하는 질의 그래프 생성 규칙은 입력된 키워드와 관련한 클래스를 생성하고, 클래스와 관련된 정보들을 도출한다. 그리고 입력된 키워드와 연관된 DatatypeProperty, ObjectProperty[15] 및 인스턴스와 연관된 클래스 정보를 도출하여 다른 인스턴스와의 관계 그래프를 그린다. 첫 번째 키워드를 입력 하게 되면, Fig. 2의 처리구조를 기반으로 Fig. 5와 같이 키워드와 관련한 질의 그래프를 생성한다. 입력된 키워드 ?key1과 관계된 op(ObjectProperty)와 dp(DatatypeProperty)를 추출하고, type관계로 정의된 클래스 ?y를 도출한다. c1, c2는 ?key1과 관계된 op 관계의 클래스들을 의미한다. 해당 키워드 ?key1과 연관된 op 관계가 많거나 ?y에 속한 클래스 관계가 복잡하게 되면 하나의 키워드를 입력하고 나서 나올 수 있는 관련 정보가 많아지게 된다. Fig. 6은 하나의 키워드를 입력 후 도출 가능한 SPARQL 질의 정보를 나타내고 있다. “project 1”을 첫 번째 키워드로 입력 후, 생성 가능한 SPARQL 질의 문이다. “project1”은 ?key1 변수로 하여 관련된 트리플 정보들을 추출하고 있다. 이런식으로 추출된 ?key1과 관련한 모든 정보들은 사용자로 하여금 원하는 정보가 어떤 것인지 명확하게 알려줄 수가 없다.

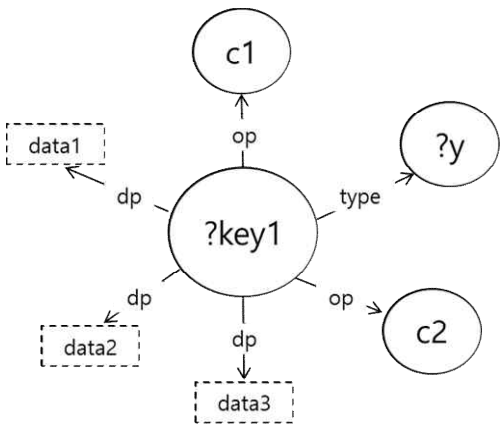

Fig. 5. Keyword Graph

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?key1 ?y ?predicate ?value ?extentionValue
WHERE{?key1 rdf:type ?y.
        ?key1 ?predicate ?value.
        ?y ?relation ?extentionValue.
FILTER (regex(str(?key1), 'project1')).
}LIMIT 50
```

Fig. 6. SPARQL Query 1

따라서 사용자는 원하는 검색 결과를 위해 검색 키워드를 추가하여 검색을 지속한다. 이때, 입력된 키워드는 키워드 히스토리 온톨로지에 K1클래스의 인스턴스로 되고, 작성된 SPARQL 질의는 Query 클래스의 인스턴스로 연관된다. 다시 키워드가 추가되면 기존의 질의에서 추가하여 그래프를 확장해간다. Fig. 7은 관련키워드를 하나 더 추가했을 때, 생성 가능한 질의이다. 처음에 입력된 키워드와 관련성이 있는 것인지 먼저 ASK 구문을 이용해서 검사를 하고, 관련된 키워드면 다음 질의문을 작성하게 된다. 만약, 두 키워드가 직접적인 관련이 없는 트리플 셋의 형태면 UNION을 써서 두 키워드가 들어간 트리플 모두를 질의 해서 결과를 반환한다. Fig. 6에서 작성한 질의문을 바탕으로 Fig. 8과 같이 key2에 대한 질의문을 중심으로 트리플을 확장한다. Fig.9는 ASK 구문과 키워드 2개 일 때 SPARQL 질의문을 생성하는 알고리즘이다.

```
PREFIX rdf: <http//www.w3.org/1992/02/22-rdf-syntax-ns#>
ASK{?key1 ?predicate ?key2.
FILTER (regex(str(?key1), 'project1') && regex(str(?key2), 'kim')).
}
```

Fig. 7. SPARQL Query 2

```
PREFIX rdf: <http//www.w3.org/1992/02/22-rdf-syntax-ns#>
SELECT ?rPredicate ?y ?predicate ?value ?extentionValue
WHERE{?key1 ?rPredicate ?key2.
        ?key2 rdf:type ?y.
        ?key2 ?predicate ?value.
        ?y ?relation ?extentionValue.
FILTER (regex(str(?key1), 'project1') && regex(str(?key2), 'kim')).
}LIMIT 50
```

Fig. 8. SPARQL Query 3

```
Inputs: user keyword n
Outputs : SPARQL query statement

If(ASK statement(keyword 1, keyword 2))
        select predicate(between keyword 1, keyword 2) in triple
        select class of keyword 2
        select object(keyword 2) in triple
        select object(class of keyword 2)
else
        select object(keyword1 union keyword 2) in triple
```

Fig. 9. Algorithm of SPARQL Query Statement Construction

도출된 질의 그래프를 기반으로 분기와 관련한 상태정보가 있을 때는 길이기반, 빈도수 기반 랭킹방법을 통해 하나의 질의문을 완성한다. 완성된 질의문을 해당 키워드의 질의 정보로 규정하고, 키워드 히스토리 온톨로지에서 Query 클래스의 인스턴스로 해당 키워드와 연관시킨다. 이와 같이 키워드가 들어올

때마다 논문에서 정의한 질의 생성 규칙을 통해 관련 정보 그래프를 생성하고, 랭킹을 정해서 질의문 결과의 정확도를 높이는 과정을 반복한다. 본 논문에서는 이와 같이 검색을 할 때 추가 했던 키워드를 중심으로 키워드 이력 그래프를 온톨로지 형태로 변환하여 구축한다. 최종적으로 입력된 키워드를 중심으로 완성된 그래프를 반영하고, 검색 시에 활용할 수 있도록 한다.

IV. Experiment

1. Construction of Domain Ontology

도메인 온톨로지는 OLW 2로 구축하여, 계산 완전성과 시스템에 제약 없는 형태의 추론을 할 수 있도록 한다. 온톨로지 구축을 위한 툴은 Protege[16]를 사용해서 구축하였으며, 법률은 “여권법”을 기반으로 구축한다. 법률에서 명사들을 추출하여 법률 클래스의 인스턴스 형태로 관계를 맺는다. Fig. 10은 Protege를 통해 구축된 클래스의 스키마 및 인스턴스의 일부를 보여주고 있다.

2. Implementation of Semantic Search Tool

본 논문에서는 제안하는 키워드 히스토리 온톨로지 기반의 시맨틱 검색 방법을 위해 시맨틱 검색 툴을 설계 및 구현한다. 키워드 히스토리 온톨로지 기반의 시맨틱 검색은 키워드 기반으로 법령과 같은 도메인 온톨로지에 효과적으로 질의 하고 결과를 반영하는지를 확인할 수 있다. 시맨틱 검색 툴은 Table 2와 같은 기술을 기반으로 구현된다.

Table 2. Spec of Semantic Search Tool

Element	Version
Programming Language	Java 1.8
Ontology	OWL 2
Query Language	SPARQL-DL
Semantic Web Framework	OWL API 3.5

시맨틱 검색 툴은 OWL API[17] 기반의 콘솔 형태에서 동작하고, 전체적인 구조는 Fig. 11과 같다. 툴은 크게 키워드 관련 키워드 히스토리 온톨로지 구축 엔진과 질의처리 엔진으로 구분된다. 검색 인터페이스로부터 검색 키워드 입력을 받고, 입력된 키워드는 키워드 히스토리 온톨로지 구축 엔진에서 키워드를 클래스의 인스턴스화를 위해 NamedInvidual로 추가하는 작업과 추후 입력되는 연관 키워드 검사 및 이전 키워드와의 ObjectProperty 연결 작업등을 하고, 입력되는 키워드를 키워드 히스토리 온톨로지에 이미 있는 키워드 인지에 대한 작업들을 처리한다.

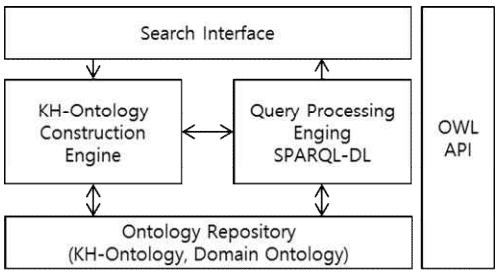

Fig. 11. Structure of Semantic Search Tool

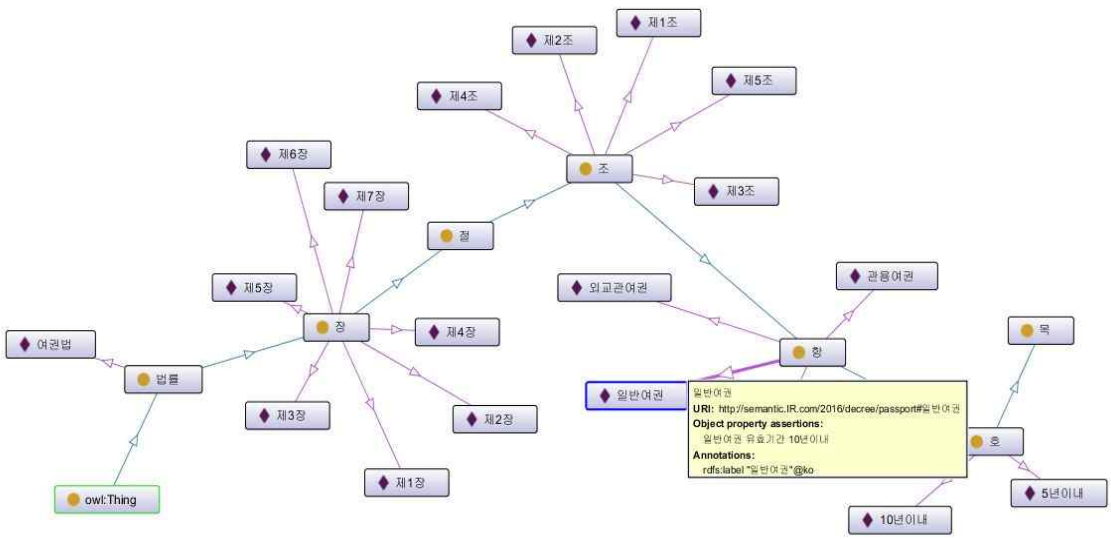

Fig. 10. Domain Ontology Schema

질의처리 엔진에서는 SPARQL-DL 기반으로 Ontology 저장소에 있는 Domain Ontology로부터 입력된 키워드와 관련한 프로퍼티 리스트(dp, op)를 추출하는 작업을 수행하고, 키워드 질의 결과를 키워드 히스토리 온톨로지 구축 엔진에 보내주고, 입력 키워드와 연관성 작업을 맺는다. 그리고 결과는 검색 인터페이스를 통해 확인한다.

3. Query comparison experiment and result

질의 비교 실험은 본 논문에서 구현한 키워드 히스토리 온톨로지 기반의 시맨틱 검색 툴과 법제처의 법령정보검색시스템[18]의 키워드를 기반으로 한 법률 검색을 진행한다. 본 논문에서는 법령정보검색시스템을 a, 본 논문에서 구축한 시맨틱 검색 툴을 b로 명명한다. 법률은 도메인 온토롤지로 구축된 '여권법'으로 Table 3과 같은 질의 목적을 가지고, 키워드를 1개, 2개, 3개씩 늘려가며 의도한 정보를 검색한다.

Table 3. List of Information Retrieval Keyword

#	Search Intent	Search Keyword
Q1	Full passport law	Passport law (여권법)
Q2	Purpose of passport law	Passport law, Purpose (여권법, 목적)
Q3	Article 2 of the Passport law	Passport law, Article2 (여권법, 제2조)
Q4	Type of passport	Passport law, Passport type (여권법, 여권종류)
Q5	Valid period for each type of passport	Passport law, Validity, general passport (여권법, 유효기간, 일반여권)

처음 1개 키워드인 '여권법'으로 a, b 둘 다 기본적인 사항을 확인가능하다. a에서는 모든 여권법의 법률 내용을 볼 수 있었다. Fig. 12는 Table 3의 Q3 검색 질의에 의한 SPARQL 질의 형태로 변환된 모습이다. '여권법' 키워드와 매칭 되는 값을 찾고, '제2조' 키워드와 매칭 되는 목적어(?value) 값을 찾아서 리턴해주는 형태로 질의문이 변환되어 결과를 확인 가능하다.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?s ?key1 ?key2 ?value
WHERE {
  {
    ?s ?p ?key1.
    ?s ?predicate ?value
    FILTER regex(?key1, '여권법')
  } UNION {
    ?s ?b ?key2.
    ?s ?predicate ?value
    FILTER regex(?key2, '제2조')
  }
}LIMIT 50
```

Fig 12. SPARQL Query statement of Q3

Table 3을 기반으로 한 실험에서 a는 원하는 검색 의도를 위해 키워드를 입력하고, 추가 등의 작업을 하였지만 Q1을 제외한 모든 경우에 있어서 찾을 수가 없었고, 정보를 찾기 위해선 전체 여권법에서 원하는 내용을 사용자 스스로 찾아야 했다. 하지만 b에서는 검색 의도에 합당한 키워드를 늘려가면서 원하는 검색 의도를 찾을 수 있었다. 질의 5개에 대한 정확도는 a는 0.2, b는 1.0의 결과를 얻을 수 있다. a에 비해 b방식을 이용한 검색이 5배가량 좋은 성능이 나타남을 알 수 있다. 이와 같은 결과는 b는 명사형으로 정제된 데이터로 온톨로지로 구축하여, 데이터간의 관계가 명시적으로 표현이 되었기 때문이다. 반면, a는 이와 같은 정제작업 없이, 모든 결과를 표시하는 형태의 검색방식을 취했기 때문이다.

이와 같은 정보 검색 방식의 변화는 사용자가 원하는 정보를 명확하게 찾는데 도움을 줄 수 있고, 특히 전문정보와 같은 분야에서 유용한 형태로 사용될 수 있다.

V. Conclusions

본 논문은 키워드 간의 연관성을 바탕으로 시맨틱 정보 검색을 위해 필요한 SPARQL 질의문으로 자동 변환하기 위한 방법을 제안하였다. 또한, 법령 기반의 시맨틱 검색을 위한 실험을 위해 시맨틱 검색 툴을 설계 및 개발하였다. 특히, 입력된 키워드의 이력을 관리하고, 질의의 효율성을 높이기 위해 키워드 히스토리 온톨로지를 고안하고, 검색 툴을 통해 구현 가능하도록 하였다. 본 논문에서 제안한 방법은 기존의 키워드를 기반으로 한 시맨틱 검색에서 키워드간의 연관성에 관한 정보의 부족으로 생길 수 있는 검색 의도의 부정확함에 관한 해결방법으로 사용될 수 있으며, 키워드 히스토리 온톨로지에 누적된 키워드 관련 정보는 관련 키워드의 검색 정확도를 더욱더 향상 시킬 수 있을 것으로 기대된다.

본 논문에서는 제안된 방법의 실험을 위해 검색 툴을 이용해서 법제처의 법령정보검색 시스템과의 의도된 검색 리스트를 가지고 실험을 진행하였으며, 법제처에서는 해당 법률 조문별로 색인이이루어지지 않아서, 세부 검색을 위한 방법을 제공하지 못하고 있었다. 따라서 제시된 검색 리스트의 검색 결과를 볼 수 없었다. 하지만 본 논문에서 제안한 검색 툴로는 원하는 정보를 탐색 가능할 수 있었으며, 이와 같은 결과는 정제된 온톨로지를 기반으로 한 데이터를 통해 시맨틱 검색을 하였기 때문이다. 따라서 자연어에 대한 정제 및 단어 간의 연관 작업의 성숙도를 기반으로 온톨로지 기반의 데이터의 보편화는 시맨틱 검색을 통한 의도된 정보를 찾는데 도움을 줄 수 있다.

본 논문에서 제안한 SPARQL 질의 자동 변환 방법은 시맨틱 검색을 통해 다양한 영역에서 사용자가 의도한 정보를 찾는데 도움을 줄 수 있을 것으로 기대된다.

REFERENCES

[1] D. W. Jo, M. H. Kim, “A Framework for Legal Information Retrieval based on Ontology,” Journal of the Korea society of computer and information, Vol. 20, No. 9, 2015.

[2] Finin, Tim, et al., "Information retrieval and the semantic web," Proceedings of the 38th annual Hawaii international conference on system sciences. IEEE, 2005.

[3] J. Song, et al., "Ontology-based information retrieval model for the semantic web," 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service. IEEE, pp. 152-155, 2005.

[4] H. Jen-Hwa, M. Pai-Chun, and P. YK. Chau. "Evaluation of user interface designs for information retrieval systems: a computer-based experiment," Decision support systems 27.1, pp. 125 - 143, 1999.

[5] E. Sirin, and P. Bijan, "SPARQL-DL: SPARQL Query for OWL-DL," OWLED, Vol. 258. 2007.

[6] T. Tran, et al., "Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data," 2009 IEEE 25th International Conference on Data Engineering. IEEE, pp. 405-416, 2009.

[7] D. W. Jo, M. H. Kim, “A Study on Legal Ontology Construction,” Journal of the Korea society of computer and information, Vol. 19, No. 11, pp. 105-113, 2014.

[8] J. H. Lee, et al., “An n-Gram-Based Indexing Method for Effective Retrieval of Hangul Texts,” Journal of the Korean Society for Information Management, Vol. 13, No. 1, pp. 47-63, 1996.

[9] B. K. Sun, D. H. We, and K. R. Han, "A Study on Paper Retrieval System based on OWL Ontology," Journal of the Korea Society of Computer and Information vol. 14, No. 2, pp. 169-180, 2009.

[10] T. W. Kim, “Query Translation for Resolving the Difference between User Query Words and Ontology Resources,” Journal of the Society of Korea Industrial and Systems Engineering, vol. 34, No. 1, pp. 35-44, 2011.

[11] M. Arenas, and J. Pérez, "Querying semantic web data with SPARQL," Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2011.

[12] C. Wang, et al., "Panto: A portable natural language interface to ontologies," European Semantic Web Conference. Springer Berlin Heidelberg, 2007.

[13] B. J. Kim, et al., “A Method of Ranking Structured Queries for Keyword Search on Semantic Web Data,” Journal of KISS: Databases, Vol. 39, No. 2, pp. 138-146, 2012.

[14] S. Harris, et al., "SPARQL 1.1 query language," W3C Recommendation 21, 2013.

[15] P. Hitzler, et al. "OWL 2 web ontology language primer," W3C recommendation, 2009.

[16] H. Knublauch, et al. "The Protégé OWL plugin: An open development environment for semantic web applications," International Semantic Web Conference. Springer Berlin Heidelberg, 2004.

[17] M. Horridge, and S. Bechhofer, "The owl api: A java api for owl ontologies," Semantic Web 2.1, pp. 11-21, 2011.

[18] Korea Ministry of Government Legislation, http://www.law.go.kr/LSW/main.html

Authors

Dae Woog Jo received the B.S. in Computer Engineering from Hallym University, Korea, in 2008. M.S. and Ph.D. degrees in Department of Computer Science and Engineering from Soongsil University, Korea, in 2010 and 2015, respectively. He is currently an assistant professor in the School of Software, Soongsil University. He is interested in semantic web, ontology engineering and linked data and distributed system.

Myung Ho Kim received the B.S. in Department of Computer Science and Engineering from Soongsil University, Korea, in 1989. M.S. and Ph.D. degrees in Department of Computer Engineering from Postech University, Korea, in 1991 and 1995, respectively. He is currently a professor in the School of Software, Soongsil University. He is interested in parallel computing, distributed computing and system software and information security.