韓國컴퓨터情報學會 論文誌
第16卷 第3號, 2011. 3.

2011-16-3-1-5

온톨로지 검색에 있어서 사용자 질의어와 온톨로지 리소스와의 상이성 해소를 위한 질의어 변환

김 태 완*

Query Translation for Resolving the Difference between User Query Words and Ontology Resources

Kim, Taewan*

요 약

차세대 웹 기술로 각광 받고 있는 시맨틱 웹에서 웹 리소스는 온톨로지에 기술된 다양한 메타데이타에 의해 가용하게 되며 이러한 온톨로지를 대상으로 한 검색을 위해 SPARQL과 같은 여러 시맨틱 질의 언어가 제안되었다. 그러나 이러한 시맨틱 질의 언어를 사용자가 습득하였다 하더라도 올바른 정보를 찾기 위해서는 사용자가 온톨로지의 구조와 어휘를 잘 파악하고 있어야 한다는 어려움이 있다. 특히 사용자가 질의 언어의 문법과 복잡한 형식 논리 이론에 익숙하더라도 온톨로지 리소스 표기에 사용된 어휘를 모를 경우에는 질의 언어 작성 시 전혀 다른 어휘를 사용할 수 있으며 그에 따라 원하는 정보를 찾지 못하는 어휘 상이성 문제를 해결하여야 한다. 본 논문은 이러한 어휘 상이성 문제의 해소에 주안점을 두고 질의어의 의미와 온톨로지 어휘 사이의 의미 유사도를 이용하여 사용자가 온톨로지에 없는 질의어를 사용하였을 경우 그와 의미가 가장 가까운 온톨로지 어휘로 변환하여 줌으로써 어휘 상이성 문제를 해결하는 방안을 제안한다.

▸ Keyword : 온톨로지 검색, 시맨틱웹, 질의어 변환

Abstract

Ontologies are playing an important role in semantic web which is emerging as a next stage of the web revolution because various kinds of metadata are described in ontologies. Correspondingly, many query languages like SPARQL, RDQL etc. have been proposed for querying these ontologies. But users have to know the structures and resource names of ontologies completely to get search results even if they have expertise on complex formal logic and syntax of the query languages. Especially, casual users do not know the resource names and may use different words from resource names when they write their query language. This vocabulary gap problem have to be solved to raise the success rate. In this paper, an approach for translating user's search words to

• 제1저자 : 김태완
• 투고일 : 2010. 11. 25, 심사일 : 2010. 12. 29, 게재확정일 : 2011. 01. 10.
*인제대학교 컴퓨터공학부(School of Computer Engineering, Inje University)
※이 논문은 2007년도 인제연구장학재단 교수연구년 지원에 의한 연구결과임.

corresponding resource names has been proposed. This approach uses semantic similarity between user created search words and ontology resource names.

▸ Keyword : Ontology Querying, Semantic Web, Query Translation

Ⅰ. 서 론

기존의 텍스트 검색이나 HTML 기반 웹 콘텐트는 사람이 이해하고 사용하는 데에는 큰 어려움이 없으나 기계는 자연언어로 기술된 정보를 단순한 문자의 나열로만 인식할 뿐, 그 의미를 이해할 수 없기 때문에 기존의 검색에서는 다양한 언어 자원과 통계적 기법을 이용하여 질의어 확장, 연관 검색, 유사도 계산 등을 거쳐 가능한 한 질의어와 관련성이 높은 정보를 찾는 것이 목적이었다. 반면 시맨틱 웹은 현재의 인터넷과 같은 분산 환경에서 지식, 그리고 지식 사이의 관계-의미 정보를 기계가 처리할 수 있는 온톨로지 형태로 표현하고, 이를 기계가 처리하도록 하는 프레임워크이자 기술이다. 즉, 종래의 HTML 기반 웹 콘텐트들이 사람은 이해할 수 있지만 기계는 의미가 아닌 구문의 형태를 인식하는 수준에서만 가공하고 처리할 수 있었던데 반하여 기계가 그 의미와 개념을 이해하고 처리할 수 있도록 하는 것이 시맨틱 웹 기술이며, 그 의미와 개념을 이해하고 처리할 수 있도록 하는 메타 데이터들은 온톨로지 형태로 기술된다.

온톨로지란 특정 분야의 지식 체계를 개념을 나타내는 클래스와 관계를 나타내는 프로퍼티로 구성하고 단위 지식들을 인스턴스와 그 프로퍼티 값으로 나타내는 지식 표현 방법이라고 할 수 있으며[1][2], 지식의 공유와 교환에도 사용된다. 또한 온톨로지로부터 원하는 지식을 검색하기 위해서는 RDQL[3], SPARQL[4]과 같은 정형 질의 언어가 사용된다. 하지만 이러한 온톨로지 검색에 있어서 문제가 되는 것은 SPARQL과 같은 질의 언어에 포함된 질의어가 온톨로지 리소스와 정확히 매치되어야 올바른 검색 결과를 얻을 수 있으며 그렇지 않을 경우에는 검색이 실패한다는 것이다.

본 논문은 SPARQL과 같은 온톨로지 검색 질의 언어에 포함되어 있는 질의어가 검색 대상 온톨로지 리소스와 매치하지 않을 경우 온톨로지 리소스를 나타내는 어휘 중에서 가장 의미가 가까운 어휘를 찾아내어 교체하여 줌으로써 검색 실패를 줄이고 적절한 검색 결과를 찾을 수 있도록 하여 주는 질의어 변환 방법을 제시한다. 이를 위해 클래스와 그 구성원인 인스턴스들을 표기하기 위해 사용된 어휘들 하나하나는 각각 독자적인 중의성을 갖고 있으나 해당 어휘들이 온톨로지에서 클래스와 그 인스턴스로 존재한다는 사실은 그것들이 해당 온톨로지에서 동일한 의미로 사용되었다는 온톨로지의 내재적 특성을 주로 이용하고 있다.

Ⅱ. 관련 연구

지식베이스로부터 원하는 정보를 자연언어로 검색하고자 하는 연구는 데이터베이스를 자연언어로 검색하고자 하는 연구에서 비롯하였다[5][6][7]. 현재는 지식베이스가 개념과 개념간의 관계를 나타내는 온톨로지로 대체되었다고 볼 수 있다. 또한 온톨로지 검색을 위한 인터페이스로 NLP-Reduce[8], Qurix[9], Ginseng[10], Semantic Crystal[11]등을 비교 분석한 연구[12]에 따르면 일반 사용자들은 익숙하지 않은 GUI 스타일의 온톨로지 검색 인터페이스보다는 익숙한 자연언어 검색을 선호하는 것으로 나타났다. 따라서 자연언어를 이용한 온톨로지 검색 방법을 취하는 것이 올바른 방향이며 이에 따라 사용자가 자유롭게 질의어를 선택하여도 온톨로지 검색이 가능하도록 하는 것이 중요하다고 판단된다.

온톨로지를 대상으로 한 시맨틱 검색에 있어서 사용자가 온톨로지의 구조와 리소스명을 알지 못하여 올바른 검색을 하지 못하는 문제를 해결하기 위한 연구로는 사용자가 익숙한 기존의 키워드 입력 방식을 채택하고 입력된 키워드로부터 query graph를 만들고 변환 규칙을 적용하여 SPARQL 질의언어를 만드는 방법[13], 단순 키워드가 아닌 문단위의 자연언어 질의문을 구문 분석하여 파스트리를 만들고 그로부터 query triple을 생성한 후 온톨로지 트리플과 매핑을 통해 SPARQL을 작성하는 방법[14]이 제안되었으나 이 방법들은 본질적으로 본 논문에서 제안하는 어휘 상이성 해결을 목적으로 하는 연구가 아니며 또한 어휘 상이성 문제에 대처하기 위해 단지 유의어 사전에만 의존하는 한계를 보이고 있다. 본 논문에서는 사용자가 선택한 질의어와 온톨로지 리소스명이 서로 다를 때 사용자가 선택한 질의어를 유의어 사전과 의미체계를 이용하여 그와 가장 의미가 유사한 온톨로지 리소스명으로 변환하여 줌으로써 검색 성공률을 향상시키는 방법을 제안한다. 유의어 사전과 의미체계를 이용한 질의어 생성 방법은 대량의 문서 또는 정보를 검색 대상으로 한 기존의 정보검색, 교차언어 정보검색[15][16][17] 등에서도 이용되고 있는 것으로 본 논문에서는 이를 리소스 어휘들이 한정되어 있으며 일정한 구조를 갖춘 온톨로지 검색을 위한 질의어 변환에 사용하였다.

Ⅲ. 어휘 상이성 해소를 위한 질의어 변환

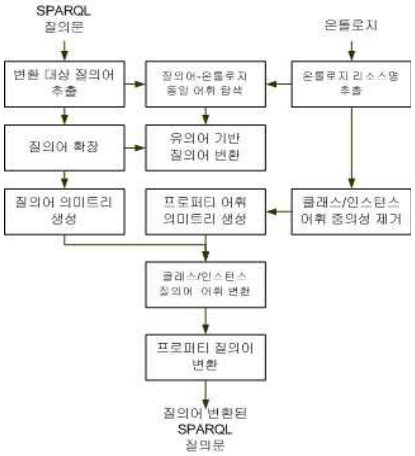

그림 1. 구성도

그림 1은 사용자가 질의어로 사용한 어휘와 온톨로지에서 리소스를 나타내기 위해 사용된 어휘가 서로 다른 경우 사용자가 사용한 질의어에 의미적으로 가장 가까운 온톨로지 리소스 어휘로 질의어를 바꾸어 줌으로써 사용자가 올바른 정보를 찾을 수 있도록 하여 주는 시스템의 구성도이다. 우선 질의 정형 언어 SPARQL 질의문으로부터 변환 대상 질의어를 추출한 후 온톨로지 리소스명과 일치하는 것을 찾아 질의어 변환 대상에서 제외한다. 온톨로지 리소스명과 일치하지 않는 질의어들에 대해서는 유의어 사전을 이용하여 질의어 확장을 수행하고 확장된 유의어들 중에서 온톨로지 리소스명과 일치하는 것이 있을 경우 질의어를 해당 온톨로지 리소스명으로 변환한다. 일치하는 온톨로지 리소스명이 없는 질의어들에 대해서는 어휘 의미들의 상하위 관계를 기술한 한국어 의미 체계 사전을 이용하여 의미 트리를 생성한다. 한국어 의미 체계 사전은 울산대학교의 U-WIN[18]을 사용한다. 온톨로지에 대해서는 우선 온톨로지로부터 온톨로지 리소스명의 표기에 사용된 어휘들을 추출하고 클래스/인스턴스명의 표기에 사용된 어휘들에 대해서는 한국어 의미 체계사전을 사용하여 온톨로지 클래스와 인스턴스의 표기에 사용된 어휘들이 갖는 공통 의미를 결정하여 중의성을 제거하고, 온톨로지 프로퍼티의 표현에 사용된 어휘들에 대해서는 언어 의미체계를 이용하여 의미 트리를 생성한다. 그 후 앞에서 얻어진 질의어들의 의미트리와 클래스/인스턴스들의 공통의미로 선택된 의미, 온톨로지 프로퍼티 어휘들의 의미트리를 이용하여 각 질의어를 그에 대응하는 가장 의미 유사도가 높은 온톨로지 어휘로 변환하여 질의문을 재구성한 후 검색을 수행한다.

1. 변환 대상 질의어 추출

온톨로지 검색용 정형 질의 언어인 SPARQL로부터 질의어를 추출한다. 이하 SPARQL문을 질의문, SPARQL 안에 포함되어 있는, 온톨로지 리소스와 매핑이 필요한 어휘를 질의어라고 정의한다. 시맨틱웹의 근간이 되는 온톨로지는 기본적으로 임의의 공통적 특성을 가지는 개념이나 개체들의 집합인 클래스와 그 구성원인 인스턴스 그리고 프로퍼티로 구성된다. 프로퍼티는 또 클래스와 클래스의 관계를 나타내는 오브젝트 프로퍼티와 클래스가 갖는 속성을 나타내면서 integer, float, boolean, string과 같은 데이터타입 값을 갖는 데이터타입 프로퍼티로 구성된다. 예를 들어 그림 2에서 보인 온톨로지 스키마의 예에서와 같이 '도', '시', '군', '자치구', '산', '하천', '도로'는 클래스 리소스를 나타내는 어휘, '높이', '인구', '면적', '길이' 등은 데이터타입 프로퍼티를 나타내는 어휘, '경유지', '소재지', '시점', '종점'은 오브젝트 프로퍼티를 나타내는 어휘이며 이 스키마를 기반으로 구축한 온톨로지에서 '시' 클래스의 구성원으로 올 수 있는 '대전', '인천', '부산', '광주' 등이 인스턴스가 된다. 또 이러한 온톨로지 지식은 기본적으로 (subject predicate object)라고 하는 트리플(triple) 형태로 기술된 논리항의 집합체로 구성된다. 여기서 subject에는 클래스, 인스턴스가 오며, object에는 클래스, 인스턴스 또는 데이터타입의 값이 올 수 있다. 트리플이 (클래스/인스턴스, predicate, 클래스/인스턴스) 형태일 경우의 predicate를 '오브젝트 프로퍼티'라고 하고, (클래스/인스턴스, predicate, 데이터값) 형태일 경우의 predicate를 '데이터타입 프로퍼티'라고 한다.

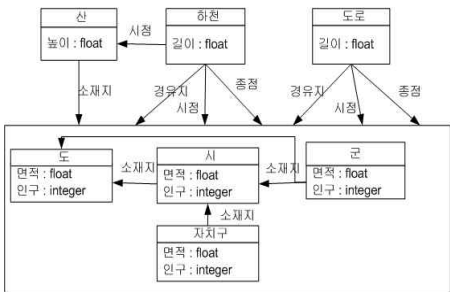

그림 2. 스키마

온톨로지로부터 원하는 지식/정보를 얻기 위해 사용하는 정형 질의 언어인 SPARQL은 기본적으로 RDF triple의 subject,

predicate, object 중 일부를 변수로 나타내고 온톨로지 트리플과 매치하는가의 여부와 SPARQL 언어에서 제공하는 일련의 질의 연산을 통해서 원하는 정보를 검색하게 된다.

```
SELECT ?x1 ?y1 ?z1
{{
     ?x1  rdf:type     :강 .
     ?x1  :기점      ?w1 .
     ?x1  :경유지    ?y1 .
     ?y1 :넓이    ?z1 .
FILTER regex(str(?w1), "태백")
   OPTIONAL
       {  ?x2  rdf:type     :강 .
          ?x2  :경유지    ?y2 .
          ?x2  :기점      ?w2 .
          ?y2 :넓이    ?z2.
          FILTER regex(str(?w2), "태백")
              FILTER (?z2 > ?z1)
            }
            FILTER (!bound (?y2))
  }}
```

그림 3. SPARQL 질의문

그림 3은 SPARQL 질의 언어의 예이다. SPARQL 언어에서는 온톨로지 트리플과 매치하여야 하는 리소스명은 앞에 콜론(:)을 붙여서 표기하고, 변수는 변수명 앞에 물음표(?)를 붙이며 이 밖에 다양한 질의 연산을 수행하기 위한 문법 어휘들을 제공하고 있다. 여기서 질의 연산을 위하여 SPARQL에서 제공하는 여러 문법 어휘들과 변수명은 사용자 질의어와 온톨로지 리소스명 사이의 어휘 상이성을 해소하기 위한 대상이 아니다. 또한 사용자가 문자열을 질의문 안에 직접 표기한 경우는 문자열이 표층적으로 온톨로지 어휘와 일치하기를 의도하는 것으로 해석할 수 있으므로 문자열을 그대로 존치할 필요가 있다. 따라서 변환하여야 할 질의어는 SPARQL 질의 언어에서 사용자가 온톨로지 리소스와 매핑 시키기 위하여 사용된 어휘들 즉 클래스, 인스턴스, 프로퍼티의 표기에 사용된 어휘들이다.

변환 대상 질의어 추출 단계에서는 SPARQL에서 변환 대상 질의어를 추출한다. 또한 질의 언어에 기술된 트리플에서 리소스의 위치에 따라 subject, predicate, object인지를 알 수 있으며 subject, object 위치에 오는 질의어의 경우 온톨로지의 클래스명 또는 인스턴스명에 대응하고 predicate 위치에 오는 질의어는 오브젝트 프로퍼티명, 데이터타입 프로퍼티명과 대응하므로 질의어 변환을 위한 온톨로지 리소스명의 범위를 제한하는데 이용한다.

그림 3의 SPARQL 질의 언어에서 변환 대상이 되는 질의어를 추출하면 다음과 같다. 이 중에서 '경유지'는 온톨로지 리소스명과 동일하므로 변환 대상에서 제외된다.

· 강 (object)　　· 경유지, 기점 (predicate)
· 넓이 (predicate)

2. 유의어 기반 질의어 변환

어떠한 개체가 갖는 속성을 나타내는 단어들, 즉 길이, 색, 면적, 인구, 고도, 크기, 높이, 깊이, 거리, 무게, 수, 부피, 너비, 년도, 날짜 등과 같은 단어들은 중의성이 거의 없다는 점에 착안하여 이러한 단어들은 별도로 유의어 사전을 구축하여 해당 한국어 질의어에 대해 의미 분석 단계까지 가지 않고 바로 해당 질의어를 변환한다. 이러한 어휘에는 길이-length, 면적-area, 인구-population, 고도-altitude, 크기-size, 높이-height, 깊이-depth, 거리-distance, 무게-weight, 개수-number, 부피-volume, 두께-thickness, 너비-breadth, 년도-year, 월-month, 시간-time, 날짜-date, 일시-date_time 등이 있다. 또 이 단어들은 온톨로지 트리플에서 object 위치에 integer, float, boolean, date time, date, time, string과 같은 데이터타입을 갖는 값을 직접 기술하는, 예를 들어 (한국 인구 48,200,000), (한라산 높이 1,950)과 같은 데이터타입 프로퍼티의 명칭으로 많이 나타난다. 표 1은 유의어 사전의 예이다.

표 1. 유의어 사전의 예

질의어	유의어
이름	성명, 성함, 존함, 명칭, 명의
인구	인구수
높이	고도
넓이	면적, 크기
길이	장단, 기장

그림 3의 질의 예제에서 변환 대상인 질의어 중 '넓이'라고 하는 질의어는 유의어 사전에 등록되어 있으므로 {넓이, 면적, 크기}로 질의어가 확장되고 그 중에서 '면적'이라는 어휘가 온톨로지 리소스명과 일치하므로 '넓이'라고 하는 질의어는 '면적'이라고 하는 어휘로 변환된다.

3. 질의어 의미 트리 생성

일반 사용자의 경우 대상 온톨로지에 대한 지식이 없으므로 온톨로지를 구성하고 있는 클래스, 프로퍼티, 인스턴스들을 나타내기 위하여 사용된 어휘들을 모르기 때문에 질의문을 작성할 때 온톨로지에서 사용되지 않은 어휘를 질의어로 사용할 경우가 많다. 예를 들면 온톨로지에서는 '하천'이라는 클래스명으로 작성되었는데 사용자는 질의어로 '강'이라고 하는 어휘를 사용할 경우 사용자가 질의문에 사용한 질의어와 온톨로지 어휘 사이에 차이가 발생하여 검색 결과를 얻지 못하게 된다. 따라서 질의어의 의미와 가장 의미가 유사한 온톨로지 어휘를 찾아 질의어를 변환하여 줌으로써 검색 성공률을 높여야

한다. 이를 위하여 질의문에 포함되어 있는 질의어 중 온톨로지 리소스명과 동일한 질의어를 사용한 경우, 유의어 기반 질의어 변환 단계에서 변환 처리된 질의어를 제외한 질의어들에 대해 한국어 의미체계 사전[16]을 참조하여 해당 질의어가 가질 수 있는 의미 후보들을 추출하고 그로부터 의미 트리를 생성한다. 이것은 추후 온톨로지 리소스명의 표기에 사용된 어휘들과 의미 유사도 계산에 사용된다. 질의어 중에서 "강(object)"을 한국어 의미체계에서 찾아보면 표 2와 같이 총 18개의 의미가 존재한다. 또 '강'의 18개 의미에 대한 상위어들을 의미체계 사전을 찾아보면 그림 4와 같다. 이러한 18개 의미에 대한 상위어들을 통합하여 의미트리를 구성하면 그림 5와 같이 된다.

Sense 1.
강(江).....넓고 길게 흐르는 큰 물줄기.
↑하천(河川).....강과 시내를 아울러 이르는 말.
↑물.....못, 내, 호수, 강, 바다 따위를 두루 이르는 말.
↑곳.....공간적인 또는 추상적인 일정한 자리나 지역.
↑공간(空間).....물리적으로나 심리적으로 널리 퍼져 있는 범위.
↑UWIN.....모든 노드의 구조근원

Sense 2.
강(?).....가옥(家屋) 난방 장치의 하나.
↑난방장치(暖房裝置).....건물이나 방 안의 온도를 따뜻하게 하는 장치.
↑장치(裝置).....어떤 목적에 따라 기능하도록 기계, 도구 따위를 그 장소에 .....
↑기기(機器/器機).....기구, 기계(機械), 기계(器械) 따위를 통틀어 이르는 말.
↑UWIN.....모든 노드의 구조근원

그림 4. '강'의 의미체계 사전 중 일부

표 2. '강'의 18개 의미 중 일부

구분	의미	뜻
1	강(江)	넓고 길게 흐르는 큰 물줄기
2	강	가옥(家屋) 난방 장치의 하나
3	강(羌)	중국의 이민족인 오호(五胡) 가운데, 중국 서방의 변두리에 흩어져 있던 티베트 계통의 유목 민족
4	강(姜)	우리나라 성(姓)의 하나
6	강(剛)	조선 후기에, 서양 음악의 '올림표'를 이르던 말
7	강(康)	우리나라 성(姓)의 하나
13	강(綱)	생물 분류학상의 한 단위
14	강(疆)	우리나라 성(姓)의 하나
15	강(鋼)	강철
16	강(講)	예전에, 서당이나 글방 같은 데서 배운 글을 선생이나 시관 또는 웃어른 앞에서 외던 일
17	강(講)	강의
18	강(講)	불교에서, 사람들이 모여서 경전 따위를 외우고 논의함

한편 질의문 트리플에서 프로퍼티 위치에 오는 '기점'에 대하여 의미트리를 구성하면 그림 6과 같다.

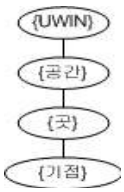

그림 6. '기점'의 의미트리

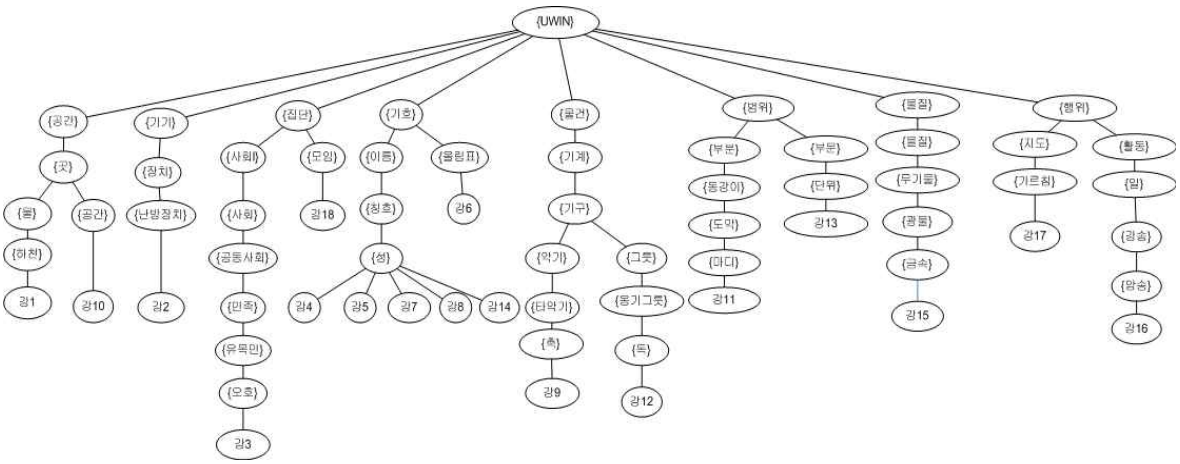

그림 5. 질의어 '강'의 의미트리

4. 클래스, 인스턴스 어휘 의미 결정

온톨로지의 클래스와 그 구성원인 인스턴스들을 나타내기 위하여 사용된 어휘들 하나 하나는 여러 가지 의미를 가질 수 있지만 같은 클래스에 속하고 있다는 사실로부터 이 클래스와 인스턴스들이 어떤 특정한 의미를 공통적으로 갖고 있다는 것을 알 수 있다. 따라서 이 공통적인 의미를 찾게 되면 해당클래스의 중의성이 제거된 의미가 되며, 이것은 다시 각 인스턴스들이 어떤 의미로 온톨로지 클래스의 구성원으로 존재하는지 알 수 있다. 예를 들어 클래스 '하천'의 의미에는 한국어 의미체계를 찾아보면 2개의 의미(河川, 下賤)를 갖고 있다. 그런데 클래스 '하천'과 클래스의 인스턴스들인 '한강', '금강', '낙동강', '영산강'과 공통의미를 찾아보면 '하천'이 '河川'의 의미임을 알 수 있다. '한강', '금강', '낙동강', '영산강'의 의미 체계를 한국어 의미체계를 참조하면 각각에 대하여 그림 7과 같은 결과를 얻을 수 있다. '한강'의 경우 1가지의 의미, 금강의 경우 2가지 의미, 낙동강의 경우 2가지 의미를 갖고 있음을 알 수 있다. '영산강'은 의미체계 사전에 등록되어 있지 않았다. 그림 7의 각각의 의미체계를 통합하면 그림 8과 같다.

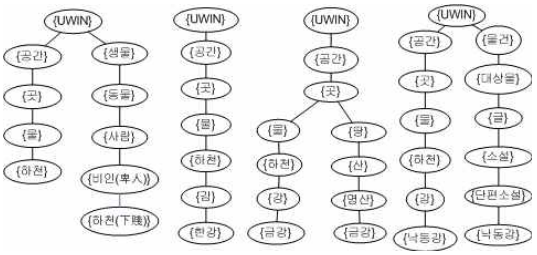

그림 7. 클래스 '하천'과 인스턴스 '한강', '금강', '낙동강', '영산강'의 의미 트리

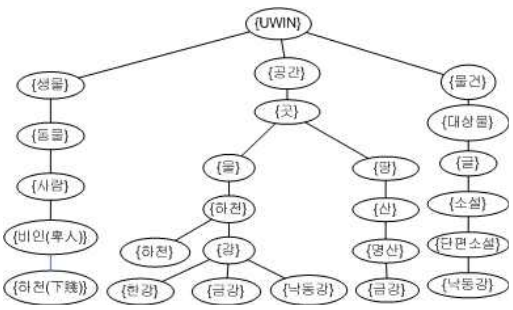

그림 8. 통합 의미 트리

그러나 그림 8에서 보는 바와 같이 단순히 공통적으로 갖는 의미를 찾을 경우 각 어휘의 의미체계에서 최상위 계층에 위치한 {UWIN}이 모든 어휘의 의미 체계에 공통적으로 들어가게 되므로 아무런 변별력이 없는 {UWIN}이 항상 선택될 수밖에 없다는 문제가 발생한다. 따라서 공통적으로 갖는 의미들 중에서 가장 하위 계층에 위치하는, 즉 변별력 있는 의미를 찾는 방법이 필요하다.

공통적이며 변별력 있는 의미를 찾기 위해 언어 의미체계를 사용하여 각 클래스와 각 클래스의 구성원인 인스턴스를 나타내기 위하여 사용된 어휘들에 대하여 그림 8과 같은 통합 의미체계를 만들고 이 통합 의미 체계로부터 수식 (1)을 이용하여 변별력 있는 가장 공통적인 의미를 찾는다.

$$S = \arg max_s \left(\frac{d(s)(e(s)+1)}{D(E+1)} \right) \quad \cdots\cdots (1)$$

여기서 S는 선택된 의미 노드, D는 통합 의미 체계의 각 의미 노드의 depth 중 최대값, 즉 통합 의미 체계의 최대 depth, E는 통합 의미 체계의 각 의미 노드의 진입(아래로부터 위로 향하는) edge들의 차수 중 최대값, s는 임의의 의미 노드, d(s)는 의미 노드 s의 depth, e(s)는 의미 노드의 진입 차수를 나타낸다. 분자와 분모에 +1을 하는 이유는 단말 노드의 경우 진입 edge 차수가 0이므로 S가 항상 0 값으로 고정되어 버리는 현상을 보정하기 위한 것이다. 이 식을 이용하여 변별력을 갖는 공통의미를 찾기 위해 각 의미노드에 대하여 계산한 일부의 예를 보면 아래와 같다.

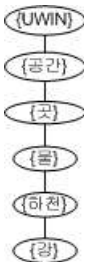

그림 9. 클래스 '하천'의 결정의미 {강}과 의미트리

{하천} : (4(2+1))/(6(3+1)) = 0.5
{강} : (5(3+1))/(6(3+1)) = 0.833
{명산} : (5(1+1))/(6(3+1)) = 0.417
{곳} : (2(2+1))/(6(3+1)) = 0.25

여기서 {강}이 최대값을 갖는 것을 알 수 있다. 따라서 클

래스 '하천'과 그 인스턴스인 '한강', '금강', '낙동강'은 {강}의 의미로 결정되어 의미 중의성이 제거된다. 또 그에 해당하는 의미트리의 구조는 그림 9와 같다. 클래스 '시', '군', '도', '산', '도로'에 대해서 같은 방법으로 각 클래스에 대하여 결정된 의미와 의미트리를 보이면 그림 10과 같다.

하천 : {강}
  강 → 하천 → 물 → 곳 → 공간 → UWIN
시 : {도시(都市)}
  도시 → 지역 → 공간 → UWIN
군 : {군(郡)}
  군 → 행정구역 → 구역 → 지역 → 공간 → UWIN
도 : {도(道)}
  도 → 행정구역 → 구역 → 지역 → 공간 → UWIN
산 : {산(山)}
  산 → 땅 → 곳 → 공간 → UWIN
도로 : {고속도로}
  고속도로 → 도로 → 교통로 → 길 → 공간 → UWIN

그림 10. 클래스 리소스명들의 결정된 의미 및 의미트리

5. 프로퍼티 어휘 의미 트리 생성

트리플에서 object의 위치에 데이터 값이 오는, '길이', '면적', '높이', '인구'와 같은 데이터타입 프로퍼티는 중의성이 거의 없는 경우가 많으므로 유의어 사전에 해당 데이터타입 프로퍼티 어휘들의 유의어들을 등록하여 처리할 수 있다. object의 위치에 클래스를 취하는 오브젝트 프로퍼티 어휘는 클래스와 그 구성원인 인스턴스들로부터 공통 의미를 찾아내는 수식 (1)을 사용하는 방법을 적용할 수 없다. 프로퍼티에는 인스턴스가 존재하지 않아 의미를 결정할 단서가 부족하기 때문이다. 따라서 오브젝트 프로퍼티 어휘는 의미트리를 구성한 후 프로퍼티 질의어 변환 단계에서 처리한다. 온톨로지의 프로퍼티 '시점', '소재지', '종점'에 대하여 의미트리를 구성하면 그림 11과 같다.

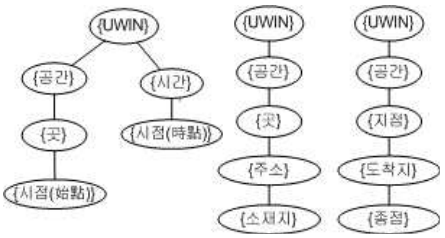

그림 11. 오브젝트 프로퍼티 '시점', '소재지', '종점'의 의미트리

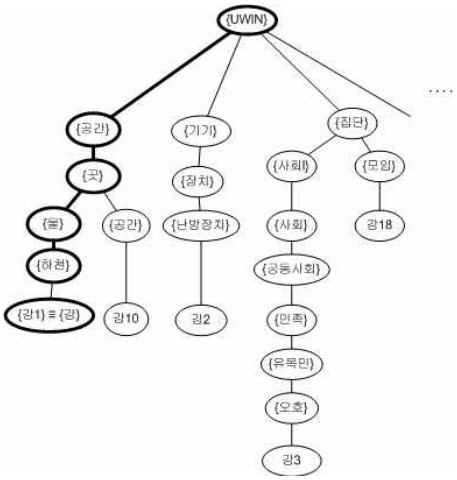

그림 12. 질의어 '강'과 클래스 '하천'의 의미 결정

6. 클래스 질의어 변환

사용자가 SPARQL 질의문에 사용한 질의어 '강'은 온톨로지에서 쓰이지 않은 어휘이나 실제로 '강'의 의미를 가진 '하천' 클래스가 존재하므로 질의어 '강'을 '하천'으로 질의어를 변환하여 주어야 검색 성공률을 높일 수 있다. 온톨로지 클래스에는 '하천', '시', '군', '도', '산', '도로'가 있다. 이 중에서 질의어 '강'과 가장 의미가 가까운 '하천'을 선택하여 변환하여야 한다. 온톨로지 클래스들의 의미를 클래스/인스턴스 중의성 해소 방법을 통하여 구한 결과는 앞의 그림 10과 같다.

질의어 의미후보 트리 생성에서 얻어진 질의어 '강'의 통합 의미트리(그림 5)와 '하천' 클래스 어휘 중의성 제거에서 결정된 의미트리(그림 9)를 통합하면 그림 12와 같은 구조를 구할 수 있다. 이로부터 질의어 {강}과 {강1}, {강}과 {강2}, ...{강}과 {강18} 간의 의미거리를 구하여 보면 {강}-{강1}은 0, {강}-{강2}는 9, {강}-{강3}은 13 등이 된다. 이 중에서 가장 의미 거리가 짧은 {강}-{강1}을 질의어 '강'과 클래스 '하천'의 최종 의미 거리로 한다. 의미거리란 해당 의미에서 다른 의미까지의 최단 경로 길이를 말한다.

다른 클래스 '시', '군', '도', '산'에 대해서도 각각 같은 방법을 적용하여 보면 각 질의어 '강'과 각 클래스 간의 최종 의미거리는 다음과 같다.

'강' - '하천' : 0
'강' - '시' : 5
'강' - '군' : 7
'강' - '도' : 7
'강' - '산' : 4

'강' - '도로' : 5

따라서 최종 의미 거리가 가장 짧은 '하천'으로 질의어 '강'을 변환한다.

7. 프로퍼티 질의어 변환

질의어 '기점'의 의미 트리(그림 6)과 온톨로지의 프로퍼티 어휘 '시점'의 의미 트리(그림 11의 좌측)을 통합하면 그림 13과 같다. 질의어 '기점'과 프로퍼티 '시점' 사이의 의미 거리를 계산하여 보면 {기점}-{시점(始點)}이 2, {기점}-{시점(時點)}이 5가 되므로 질의어 '기점'과 프로퍼티 '시점'간의 최종 의미 거리는 의미 거리가 가장 짧은 {기점}-{시점(始點)}으로 결정된다. 다른 프로퍼티 '소재지', '종점'에 대해서 같은 방법으로 최종 의미거리를 구해보면 {기점}-{소재지}는 3, {기점}-{종점}은 5로 결정된다. 따라서 최종 의미 거리가 가장 짧은 {기점}-{시점}이 선택되어 질의어 '기점'은 '시점'으로 변환된다.

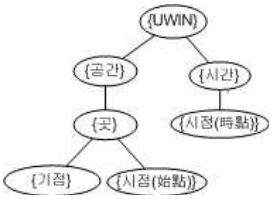

그림 13. '기점'과 '시점'의 통합 의미트리

이상과 같은 처리를 거쳐 SPARQL 질의문에 존재하는 질의어들을 의미가 가장 가까운 온톨로지 어휘로 변환하여 SPARQL 질의문(그림 3)을 재구성하면 그림 14와 같으며 이 질의문을 사용하여 온톨로지 검색을 수행하게 된다.

```
SELECT ?x1 ?y1 ?z1
{{
     ?x1  rdf:type     :하천 .
     ?x1  :시점     ?w1 .
     ?x1  :경유지   ?y1 .
     ?y1 :면적   ?z1 .
FILTER regex(str(?w1), "태백")
    OPTIONAL
        {  ?x2  rdf:type     :하천 .
           ?x2  :경유지   ?y2 .
           ?x2  :시점     ?w2 .
           ?y2 :면적   ?z2.
           FILTER regex(str(?w2), "태백")
                FILTER (?z2 > ?z1)
             }
             FILTER (!bound (?y2))
  }}
```

그림 14. 질의어 변환된 SPARQL 질의문

Ⅳ. 실험 및 고찰

실험에 사용한 SPARQL 질의문은 표 3과 같다. 각 질의문에 사용된 질의어들 중 일부를 의미는 유사하나 온톨로지 리소스명과 다른 어휘를 사용하여 모든 질의문이 검색 실패가 되도록 작성하였다.

표 3. 실험에 사용한 SPARQL 질의문

구분	의미 SPARQL 문
1	태백이 기점인 강의 경유지 중 가장 넓은 곳 SELECT ?x1 ?y1 ?z1 {{ ?x1 rdf:type :강 . ?x1 :기점 ?w1 . ?x1 :경유지 ?y1 . ?y1 :넓이 ?z1 . FILTER regex(str(?w1), "태백") OPTIONAL { ?x2 rdf:type :강 . ?x2 :경유지 ?y2 . ?x2 :기점 ?w2 . ?y2 :넓이 ?z2. FILTER regex(str(?w2), "태백") FILTER (?z2 > ?z1) } FILTER (!bound (?y2)) }}
2	가장 높은 산이 위치한 곳 SELECT ?x1 ?y1 ?z1 { ?x1 rdf:type :산 . ?x1 :고도 ?y1 . ?x1 :위치 ?z1 . OPTIONAL { ?x2 a :산 . ?x2 :고도 ?y2 . FILTER (?y2 > ?y1)} FILTER (!bound (?x2))}
3	대전을 경유하는 고속도로중 가장 긴 고속도로 SELECT ?x1 ?y1 ?z1 {{ ?x1 rdf:type :고속도로 . ?x1 :경유지 ?y1 . ?x1 :길이 ?z1 . FILTER regex(str(?y1), "대전") OPTIONAL {?x2 rdf:type :고속도로 . ?x2 :경유지 ?y2 . ?x2 :길이 ?z2 . FILTER regex(str(?y2), "대전") FILTER (?z2 > ?z1) } FILTER (!bound (?z2)) }}
4	인구가 3백만 이상인 광역시

	SELECT ?x1 ?y1 { ?x1 a :광역시 . ?x1 :인구 ?y1 . FILTER (?y1 >= 3000000) }
5	파주를 경유하는 강의 길이와 시점
	SELECT ?x1 ?y1 ?z1 { ?x1 a :하천 . ?x1 :시점 ?y1 . ?x1 :길이 ?z1 . ?x1 :경로 :파주 . }

실험에 사용한 한국어 의미체계로는 울산대학교의 UOU-Word Intelligent Network(UWIN)[18]을 사용하였으며 유의어 사전은 한컴전자사전으로부터 구축하였고 Jena [19]와 Protege[20]를 사용하였다.

표 4는 각 SPARQL 질의문에서 변환된 질의어들과 변환된 질의어들로 각 SPARQL 질의문을 재구성하여 검색을 수행하였을 때의 검색 결과를 나타낸다. 당초 모든 질의문이 검색 실패였던 것에 비하여 본 논문에서 제안한 방법을 통하여 검색 성공률을 높일 수 있음을 알 수 있다.

표 4. 변환된 SPARQL을 사용한 검색 결과

구분	변환된 질의어	결과		
1	강→하천 기점→시점 넓이→면적	x1	y1	z1
		한강	정선군	1127.6
2	고도→높이 위치→시점	no result		
3	고속도로→도로	x1	y1	z1
		경부고속도로	대전	416.0
4	광역시→시	x1	y1	
		부산	3534822	
		서울	10194327	
5	경로→시점	no result		

실험 결과를 살펴보면 유의어 사전을 이용하여 질의어를 확장하고 그 중에서 온톨로지 어휘와 일치하는 온톨로지 리소스명을 찾아 변환하여 주는, 주로 데이터타입 프로퍼티 어휘의 변환에 사용되는 유의어 기반 변환은 실패가 적고 실패할 경우에도 유의어 사전에 새로운 어휘를 등록함으로써 간단히 해결 될 수 있었다.

클래스 리소스에 대응하는 질의어의 변환은 SPARQL 질의문 1의 '강', 3의 '고속도로', 4의 '광역시'라고 하는 사용자 질의어가 각각 온톨로지 클래스 어휘인 '하천', '도로', '시'로 성공적으로 변환되어 올바른 검색 결과를 출력하고 있다. 이는 한국어 의미체계를 이용하여 클래스와 인스턴스들의 의미를 결정하고 SPARQL 질의문 트리플의 subject 또는 object의 위치에 오는 클래스 리소스에 대응하는 질의어를 온톨로지 클래스 명칭으로 변환하는 클래스 어휘 변환이 효과적임을 보이는 것이다. 단, 임의 클래스의 인스턴스의 수가 충분히 확보되지 않을 경우에는 클래스와 그 인스턴스의 의미가 정확히 결정될 수 없으므로 질의어를 효과적으로 변환할 수 없을 것으로 판단된다.

검색 실패를 보인 질의문 2의 경우 유의어 사전을 사용한 "고도→높이" 변환은 올바르나 "위치→시점"이 올바르지 못하여 검색 실패를 보였다. 이는 '위치'와 '시점'의 의미 거리가 올바른 변환인 '위치'와 '소재지' 사이의 의미거리보다 작아서 생긴 결과이다.

질의문 5의 경우에는 '경로'가 '시점'으로 잘못 변환되었기 때문에 검색 실패를 보였다. 이는 '경유지'에 해당하는 어휘가 한국어 의미체계에 없기 때문에 의미트리를 구성하지 못하였고, 따라서 질의어 '경로'가 다른 오브젝트 프로퍼티 어휘와 의미거리를 계산하여 그 중 가장 짧은 거리를 갖는 '시점'으로 변환되었기 때문이다.

전체적으로 보면 클래스 리소스에 대응하는 질의어의 변환은 클래스와 그 구성원인 인스턴스로부터 공통의미를 결정함으로써 중의성 제거가 비교적 용이한 반면 오브젝트 프로퍼티의 경우에는 보다 정확하게 의미를 결정지을 수 있는 다른 요인이 없기 때문에 질의어 변환의 실패율이 높은 것으로 분석되었으며 나아가서 한국어 의미체계에 존재하지 않는 어휘들의 경우에는 본 논문에서 제안한 방법을 적용할 수 없으므로 실패율을 높이는 것으로 분석되었다.

Ⅴ. 결론

본 논문에서는 사용자가 온톨로지에 사용된 리소스 명을 정확히 알지 못할 경우 온톨로지로부터 원하는 정보를 검색하지 못하는 어휘 상이성을 해소하기 위하여 사용자 질의어를 의미가 가장 가까운 온톨로지 리소스 명으로 변환하여 주는 방법을 제안하였다. 실험 결과 클래스와 그에 속한 복수개의 인스턴스 어휘들로부터 의미를 결정할 수 있는 클래스 어휘 변환과 유의어 사전을 사용하는 데이터타입 프로퍼티 어휘 변환은 성공률이 높았으나 오브젝트 프로퍼티의 경우에는 실패율이 높았다. 따라서 이러한 오브젝트 프로퍼티의 질의어 변환을 위한 연구가 필요하다. 또한 본 논문에서 제안한 사용자 질의어와 온톨로지 어휘 사이의 사용 어휘 상이성에 따른 연구 외에 사용자가 온톨로지의 구조를 정확히 알지 못할 경우에 대한 연구도 수행되어야 할 것이다.

참고문헌

[1] Semantic Web, http://ko.wikipedia.org/wiki/Semantic_Web

[2] Ontology, *http://ko.wikipedia.org/wiki/ontology*

[3] RDQL, *http://www.w3.org/Submission/RDQL/*

[4] SPARQL, *http://www.w3.org/TR/rdf-sparql-query/*

[5] Popescu, A.M., Etzioni, O., Kautz, H.A., "Towards a Theory of Natural Language Interfaces to Databases," Intelligent User Interfaces '03 (IUI'03), pp. 149 - 157, Miami, USA, 2003 January.

[6] Androutsopoulos, I., Ritchie, G., Thanisch, P., " Natural Language Interfaces to Databases - An Introduction," Natural Language Engineering, Vol. 1, No. 1, pp. 29 - 81, 1995.

[7] Copestake, A., Jones, K.S., "Natural Language Interfaces to Databases," Knowledge Engineering Review, Vol. 5, No. 4, pp. 225 - 249, 1990

[8] Bernstein, A., Kaufmann, E., G¨ohring, A., Kiefer, C., "Querying Ontologies: A Controlled English Interface for End-Users," LNCS, Vol. 3279, pp. 112 - 126, 2005.

[9] Bernstein, A., Kaufmann, E., Kaiser, C., "Querying the Semantic Web with Ginseng: A Guided Input Natural Language Search Engine," 15th Workshop on Information Technology and Systems (WITS 2005), pp. 45 - 50, Las Vegas, USA, 2005 December.

[10] Kaufmann, E., Bernstein, A., Zumstein, R., "Querix: A Natural Language Interface to Query Ontologies Based on Clarification Dialogs," 5th International Semantic Web Conference (ISWC 2006), pp. 980 - 981, Athens, Greece, 2006 November.

[11] Spoerri, A, "Infocrystal: A visual tool for information retrieval management," ACM Conf. on Information and Knowledge Management (CIKM 1993), pp. 11-20, Washington D.C., USA, 1993 November.

[12] Kaufmann, E., and Bernstein, A., "How Useful Are Natural Language Interfaces to the Semantic Web for Casual End-Users?," LNCS, Vol. 4825, pp. 281 - 294, 2007.

[13] Qi Zhou, Chong Wang, Miao Xiong, Haofen Wang, and Yong Yu, "SPARK : Adapting Keyword Query to Semantic Search," LNCS, Vol. 4825, pp 694-707, 2007.

[14] Chong Wang, Miao Xiong, Qi Zhou, and Yong Yu, "PANTO : A Portable Natural Language Interface to Ontologies," LNCS, Vol. 4519, pp. 473-487, 2007.

[15] Chandrasekaran, B., Josephson, J.R., Benjamins, V.R., "What Are Ontologies, and Why Do We Need Them?," IEEE Intelligent Systems, Vol. 14, No. 1, pp. 20 - 26, 1999

[15] J. Bhogal, A. Macfarlane, P. Smith, "A review of ontology based query expansion," Information Processing and Management, Vol. 43, pp 866-886, 2007.

[16] Kazuaki Kishida, "Technical issues of cross- language information retrieval: a review," Information Processing and Management, Vol. 41, pp. 433-455, 2005

[17] M. Fernandez, I. Cantador, V. Lopez, D. Vallet, P. Castells, E. Motta, "Semantically enhanced Information Retrieval: an ontology-based approach," Web Semantics: Science, Services and Agents on the WWW(2010), doi:10.1016/j.websem.2010.11.003

[18] Cheolyoung Ok, "U-WIN", KIPONTO/2006, Da-ejeon KISTI, June. 2005.

[19] Jena, *http://jena.sourceforge.net/*

[20] Protege, *http://protegewiki.stanford.edu/wiki/Main_Page*

저 자 소 개

김 태 완

1983년 : 한양대학교 공학사

1985년 : 한양대학교 공학석사

1999년 : 한국과학기술원 공학박사

1985년~2000년 : 한국전자통신연구원 선임연구원

2000년~현재 : 인제대학교 컴퓨터공학부 조교수

관심분야 : 자연언어처리, 정보검색, 시맨틱웹

Email : twkim@inje.ac.kr