온톨로지를 통한 추론형 시멘틱 검색 시스템에 관한 연구

하상범° 박영택
숭실대학교 컴퓨터학과
terrie@ailab.ssu.ac.kr°, park@computing.ssu.ac.kr

Ontology Based Semantic Search System Using Inference

Sangbum Ha° YoungTack Park
Dept. of Computer Science, SoongSil University

요 약

시멘틱웹의 등장으로 온톨로지를 통하여 에이전트가 이해할수 있는 의미(semantic)를 갖는 문서를 생성하는 것이 가능해졌다. 이러한 시멘틱웹의 영역은 비즈니스 업무 효율을 증가시키고 이를 통해 이윤을 극대화 시키는 방법으로 시멘틱 검색을 통한 정보검색시스템으로 확대적용 될 수 있다. 데이터베이스를 활용하여 문서를 저장하고 데이터베이스의 질의문을 사용하거나 일반적인 키워드기반의 정보검색 기법을 사용하여 자료를 검색하는 기존의 시스템은 다양한 분야에서 많이 연구되어 왔다. 본 논문에서는 온톨로지를 기반으로 추론을 적용한 시멘틱 검색시스템에 대하여 문서검색에 초점을 맞추어 연구 결과를 제안한다. 본 논문에서 제안하는 방식은 기존의 데이터베이스 질의문으로 검색이 불가능하거나 정보관리 시스템에서 단순히 키워드 매칭으로 검색되지 않는 문서에 대해서 본 시스템이 온톨로지와 추론을 통하여 문서의 검색이 가능함을 보인다. 이러한 방식은 자연어처리 검색과 유사한 검색영역을 갖는다. 이는 문서의 검색에 있어 단순히 키워드의 유사도에 의존하지 않고 Description Logic을 바탕으로 구성된 온톨로지에 미리 정의 되어있는 의미를 바탕으로 생성된 메타데이타를 가지고 추론을 하기 때문에 가능하다. 또한 기존의 정보관리 시스템에서 채용한 데이터베이스를 통한 질의응답 시스템을 적용하여 온톨로지 표현언어에 대해 질의 응답이 가능한 DQL 인터페이스와 연동을 통하여 본 시스템의 속도와 효율성을 극대화 시킨다.

1. 서 론

인터넷과 검색시스템의 발달로 사용자는 원하는 정보를 많은 부분을 검색시스템에 의존하여 해결한다. 기존의 데이터베이스 질의시스템과 키워드기반의 정보검색시스템에서는 복잡한 질의문의 조합과 여러번의 재검색을 통하여 사용자의 요구에 맞는 결과를 찾아주게 된다. 하지만 예를들어 사용자가 영화에 관한 정보를 검색을 요구할 때 "1996년도 아카데미 작품상을 수상한 영화의 주연남자배우의 골든글러브 수상경력은 무엇인가?" 라는 질의문은 자연어 처리없이는 한번에 검색하기 불가능한 질의문이다. 본 논문에서는 자연어처리 없이 이러한 질의문까지 검색이 가능하는 시멘틱검색 시스템을 제안한다.

2. 관련 연구

2.1 Embedded Grammar Tags(EGTs)

미국의 Maryland대학에서 연구된 EGTs는 기존의 웹 검색이 갖는 한계점, 즉 인간이 이해하는 문서를 기계는 이해할 수 없다는 점을 문서의 의미를 부여하여 기계로 하여금 문서를 이해하게 되는 시멘틱 검색을 가능하게 한다는 것이다. EGTs는 이러한 시멘틱 검색의 공통적인 이슈를 해결하는 방법으로 자연어로 구성된 질의문을 EGT를 사용하여 BNF(Backus-Naur form) 형태로 재표현하고 존재하는 웹 페이지들도 EGT를 사용하여 annotation 함으로써 시멘틱 검색을 하게된다.

예를 들어, 코스닥지수는 얼마인가?(What is the value of Kosdaq?) 라는 질의문을 입력했을때, 일반적인 정보검색으로는 제대로된 답을 얻을수는 없을 것이다. EGTs에서는 다음과 같은 방법으로 이러한 모호한 질의문을 해결한다.
질의문을 EGT를 사용하여 <ROBOTGRAM-IN> * [is][Kosdaq*][the](value|quote|price)[of Kosdaq] * </ROBOTGRAM-IN> 로 변환하고 EGT를 사용하여 annotation 된 웹 정보중에서 중요한 키워드인value와 같은 EGT 매칭이 일어나는 웹페이지의 사용자에게 반환하는 형식이다. 하지만 EGTs는 여러 가지 한계점을 갖는다. 첫 번째로 자연어로 입력된 질의문을 정확히 변환하기 힘들다는 단점이 있고, 두 번째로 annotation된 웹 페이지만 검색이 가능하다는 한계점과 오늘의 코스닥지수는 얼마인가? 와 같은 좀더 모호한 질의문의 결과를 반환하기는 힘들다는 확장성에 단점이 있다.

2.2 XML기반의 시멘틱 검색 XSEarch

독일의 Sara Cohen의 XSEarch는 기존의 시멘틱 검색과는 다른관점에서 XML을 기반으로 문서들을 검색한다. 즉, XML로 작성되어있는 문서와 문서의 계층적인 태그와 키워드를 색인하여 만든 목차 데이터베이스를 사용하여 질의문에 해당하는 문서의 유사도를 계산하여 순위가 높은 문서를 반환한다. 여기서 XSEarch의 한가지 특징은 질의문에 태그와 키워드를 동시에 허용한다는 점이다. 예를 들어 author(태그):victor(키워드) 의 title:database 라는 질의문을 입력하면 모든 문서의 태그의 계층적 색인이 저장되어있는 데이터베이스에서 하나의 상위 태그 안에 하위 태그로 author 태그와 title 태그가 존재하고 해당 키워드가 victor와 database라면 올바른 검색을 하게 되는 것이다. 만약에 victor와 database가 존재하지만 각각 다른 상위 태그아래에 위치하면 연관성이 없는 문서로 판독하고 검색결과

에 반환하지 않게 된다. Cohen의 XSEarch는 기존의 정보검색과 태그를 색인한 데이터베이스 기술을 접목시켜 태그들의 연관성을 파악하여 검색에 효율성을 높힌다. 하지만 이러한 태그들의 계층적 구조만을 사용한 방법은 문서의 의미를 이해하는 완전한 시멘틱 검색이라고 말하기 힘들며, 태그의 계층적 구조가 복잡해지거나 방대해지면 검색의 효율성이 오히려 떨어질 수가 있다.

3. 온톨로지를 통한 시멘틱 검색시스템

본 논문은 효과적인 검색 시스템의 구축을 위해 시멘틱 검색을 사용하는 방법을 제안한다. 본 논문의 시멘틱 검색은 Description Logic 바탕의 온톨로지를 사용하여 메타데이타를 생성하고 추론엔진을 접목시켜 온톨로지의 계층구조와 규칙과 공리를 사용하여 추론을 함으로써, 모호한 질의에 대하여 해당하는 문서의 검색이 가능하다. 다음 그림1 은 시멘틱 검색시스템의 전체적인 구조이다. 그림에서 보듯이 시스템은 크게 세부분으로 나뉘어져 이있다. 시멘틱 검색시스템 내에서 사용할 정보들을 정의하고 있는 온톨로지와 온톨로지를 통해 annotation 한 메타데이타 지식베이스부분, 메타데이터를 트리플로 변환하고 추론엔진을 통해 추론을 적용시키는 부분, 사용자로부터 질의문을 입력받고 결과를 출력하는 인터페이스부분으로 나눌 수 있다.

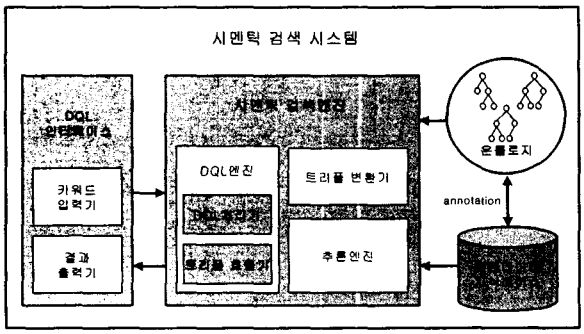

그림 1. 전체적인 시스템 구조

본 논문은 시멘틱 검색시스템의 구축에 있어서 온톨로지를 활용하여 문서들을 annotation 하여 시스템의 메타데이터로 사용함으로써 시멘틱 검색을 유도하여 검색의 효율성을 높이는 방법을 집중 연구하였다. 이러한 시멘틱 검색의 방법론으로 추론엔진을 접목시켜 추론을 수행하고 DQL을 이용하여 질의응답을 통하여 시스템을 운영하게 된다. 이러한 시멘틱 검색시스템의 효율성은 잘 정의된 온톨로지와 빠른속도의 추론엔진에 좌우되게 된다. 먼저 본 논문에서 제안하는 시멘틱 검색시스템의 전체적인 구조를 살펴보고 앞서 언급한 세부분에 대해서 세부적으로 설명하도록 한다.

3.1 온톨로지 구축 및 메타데이터 생성

본 논문에서의 실제적인 검색시스템 구현을 위하여 도메인은 문화컨텐츠로 한정하여 구축한다. 본 논문에서는 OntoEdit를 이용하여 온톨로지를 DAML+OIL의 형태로 구현하였다. DAML+OIL은 현재까지는 가장 표준화에 가까운 표현방식이고 OWL이 출현되고 있으나, DAML+OIL로부터 OWL로 변환이 용이하고 시멘틱 추론을 위한 추론 방식이 DAML+OIL에서 용이하기 때문에 본 논문에서는 DAML+OIL로 온톨로지를 구축한다. 다음 그림2는 시멘틱 검색시스템 내에서 사용될 정보들을 표현하기 위해서 구축할 온톨로지 중에서 영화에 대한 일부 온톨로지를 보여주고 있다. 온톨로지는 class, subclass, property, domain, range등의 vocabulary를 가지고 정보의 흐름을 표현한다.

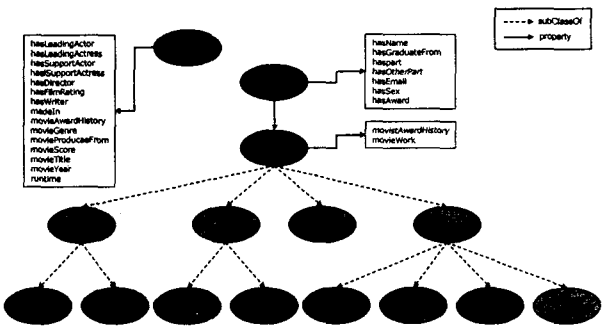

그림 2. 계층적인 온톨로지 구조

본 시스템에서 자연어처리에 가까운 질의문을 수용하기 위해서는 시멘틱 검색의 키워드가 되는 class의 다양한 property를 가지고 검색을 수행하게 된다. 앞서 언급한대로 예를들어 "1996년도 아카데미 작품상을 수상한 영화의 주연남자배우의 골든글러브 수상경력은 무엇인가?" 라는 질의문을 처리한다고 하면 검색키워드가 되는 class에는 영화제class와 주연남자배우class가 된다. 각각의 class의 property를 수상년도는 1996년과 영화제명은 아카데미, 그리고 주연남자배우의 골든글러브 수상경력으로 설정하고 검색을 수행하게 된다. 여기서 온톨로지를 모델링할 때 있어 남자배우의 수상경력처럼 특정 class의 property가 방대해질 수가 있다. 하지만 온톨로지를 모델링할 때 있어서 필요한 부분만 선택하고 나머지 부분은 qualification problem으로 생각할 수 있다. 또한 온톨로지의 class와 property를 설정할 때 있어서는 시멘틱 검색시스템이 특정 도메인에 대하여 구축되기 때문에 해당 도메인 전문가의 의해 주요검색 키워드가 되는 속성을 특정 도메인 사용자에 대한 프로파일 학습기법이나 키워드의 빈도를 가중치로 계산하여 중요 속성을 선택하는 것과 같은 전처리과정을 수행하여 class와 property를 설정하게 된다. 본 논문에서는 이러한 전처리 과정은 온톨로지 모델링에 관한 부분이므로 다루지 않도록 한다. 이렇게 구축된 온톨로지를 바탕으로 기존의 문서와 생성될 문서들에 대해서 annotation을 수행한다.

3.2 추론을 이용한 시멘틱 검색

본 논문에서 제안하는 시스템에서의 시멘틱검색은 검색방법에 추론을 적용하여 사용자의 질의문에 응답을 한다. 이러한 시멘틱검색은 사용자의 의도를 정확히 파악하여 일반적인 키워드 매칭에 의해 나오지 않는 결과를 이끌어낼수 있게 된다. 이러한 방법은 기존의 생성된 메타데이터에 DAML+OIL에서 정의된 axiom들을 적용하여 좀더 풍부한 지식베이스를 생성하거나 방대한 지식베이스를 질의문에 맞게 축약하여 검색의 속도향상에 유리하게 적용될수 있다. 앞에서 언급한대로 시멘틱검색은 Description Logic 기반의 온톨로지를 바탕으로 생성된 메타데이터와 추론엔진을 통한 전방향 추론기법을 사용하여 검색하기 때문에 적용이 가능하다. 본 논문에서 구축된 시스템의 추론엔

진 부분은 미국 스탠포드 대학의 JTP(Java Theorem Prover)를 활용한다. JTP는 온톨로지 언어인 DAML+OIL과 OWL 로 생성된 메타데이터를 KIF(Knowledge Interchange Format) 형태로 읽어 들여 전방향을 추론을 수행하도록 개발되었다. 본 논문에서는 DAML+OIL 형태의 온톨로지를 KIF형태 즉, 트리플 형태로 변환하는 컨버터를 통해서 추론엔진에 적용하게 된다. 다음 그림3은 본 시스템에서 추론엔진을 사용한 검색방법에 대한 개념도이다.

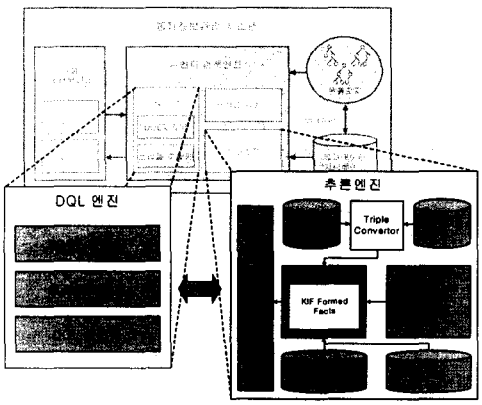

그림 3 시멘틱 검색 시스템의 추론방식 개념도

기존의 구축된 메타데이터를 사용하여 시멘틱 검색을 수행할 때 추론엔진은 DAML+OIL 정의하고 있는 Axiom을 사용하여 기존의 메타데이터를 풍부하게 또는 축약할 수가 있다. 예를들어 남자주연배우에 관한 검색을 수행할 때 남자배우인 Actor class는 남주주연배우인 LeadingActor class를 subClassOf의 관계로 정의하게 되고 추론엔진을 통해서 LeadingActor의 type도 Actor라는 사실이 지식베이스에 추가되게 된다. 또는 메타데이터의 지식베이스에 규칙을 적용시켜 방대한 지식베이스에서 검색을 하는 것 보다 수행시간을 감소시킬 수 있다.

```
(<=> (PropertyValue subClassOf ?csub ?csuper)
     (and (Type ?csub Class)
          (Type ?csuper Class)
          (forall (?x) (=> (Type ?x ?csub)
                           (Type ?x ?csuper)))))
```

- DAML+OIL 에서 정의된 subClassOf Axiom

```
Rule = (and(awardYear ?K 1996)
           (type ?K AcademyFilm)
           (hasPosition ?K 작품상))
```

- 메타데이터 정제를 위한 규칙

3.3 DQL 기반 질의문 생성

본 논문의 시멘틱 검색시스템에서는 DQL(DAML Query Language)을 사용하여 질의문을 생성한다. DQL은 시멘틱웹을 위한 질의 언어로써 DAML+OIL/OWL 로 표현된 지식베이스를 기반으로 질의를 통해 응답을 얻어낼 수 있다. 본 시멘틱 검색 시스템에서는 자연어처리에 가까운 다양한 요구를 온톨로지를 통해 정의된 class와 property를 사용자가 자유롭게 선택함으로써 자동적으로 질의문을 생성하게된다. 앞서 언급한 "1996년도 아카데미 작품상을 수상한 영화의 주연남자배우의 골든글러브 수상경력은 무엇인가?" 라는 질의문을 사용자가 검색인터페이스를 통하여 선택하여 생성된 질의문은 다음과 같다.

```
- 질의 예문
  영화제(영화제이름 : 아카데미) + 영화제(수상년도 : 1996년) + 영화제(수상부문 : 작품상) + 영화(남우주연배우 : A) + 남우주연배우(A의 수상내역 : 골든글러브)
- DQL 질의문
  SELECT ?p
  WHERE  ( ?n <N#AwardName> <N#Academy>)
         ( ?n <N#hasYear> <N#1996>)
         ( ?n <N#hasPosition> <N#작품상>)
         ( ?m <N#movieAwardHistory> ?n)
         ( ?m <N#hasLeadingActor> ?a)
         ( ?a <N#hasAward> ?w)
         ( <N#골든글러브> <N#type> ?w)
         ( ?w <N#hasPosition> ?p)
```

4. 결론

본 논문에서 제안하는 시멘틱 검색시스템으로 자연어처리와 유사한 "1996년도 아카데미 작품상을 수상한 영화의 주연남자배우의 골든글러브 수상경력은 무엇인가?" 라는 질의문을 수행하기 위하여 온톨로지를 구축하고 온톨로지바탕의 class와 다양한 property를 사용하여 DQL 질의문으로 자동생성하고 트리플 컨버터를 구축하여 추론엔진으로 하여금 DAML+OIL에서 정의된 axiom으로 온톨로지와 메타데이터 지식베이스의 추론을 통해 자연어처리 없이 복잡한 질의문의 검색을 수행하였다.

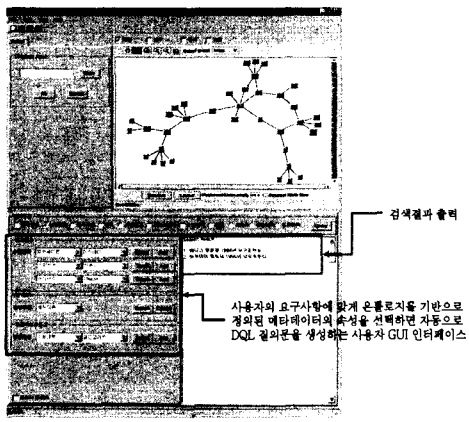

그림 4 시멘틱 검색 사용자 GUI 인터페이스

본 논문에서는 온톨로지를 구축하고 메타데이터를 생성하여 추론엔진에 적용시켜 사용자의 다양한 요구를 자연어처리 없이 시멘틱 검색으로 해결하는 방안을 제안하였다. 향후 온톨로지 모델링시에 class와 property를 적절하게 설정하는 전처리 부분에 대한 연구가 필요하겠으며, 구축된 온톨로지로 하여금 동적으로 메타데이터를 생성하는 부분에 대한 연구가 필요하겠다.

6. 참고문헌

[1] Deborah L. McGuinness and Richard Fikes, James Hendler, Lynn Andrea Stein, "DAML+OIL:An Ontology Language for the Semantic Web", IEEE 2002

[2] Fikes, Richard, Jessica Jenkins, and Gleb Frank. "JTP: A System Architecture and Component Library for Hybrid Reasoning." Proceedings of the Seventh World Multiconference on Systemics, Cybernetics, and Informatics. Orlando, Florida, USA. July 27 - 30, 2003.