메타데이터 기반 시맨틱 검색

최정화° 박영택
숭실대학교 컴퓨터학과
cjh79@ailab.ssu.ac.kr°, park@comp.ssu.ac.kr

Semantic Search based on Metadata

JungHwa Choi° YoungTack Park
Dept. of Computer Science, Soongsil University

요 약

본 논문은 '시맨틱 검색'을 위해서 시맨틱 웹 기술을 사용하여 사용자가 원하는 콘텐츠 제공을 위한 시맨틱 검색 방법을 제안한다. 본 연구는 현재 웹의 단점인 사람 위주의 웹 구성, 단순 텍스트 매칭 기반의 검색, 사람의 필터링이 필요한 대량의 결과, 특정 지식 검색이 불가능한 구조의 웹을 시맨틱 검색이 가능하도록 하기 위해서 다음과 같은 단계로 연구한다. 첫째, 도메인에 따른 정확한 정보의 제공을 위해서 OWL 온톨로지를 이용하여 컨텍스트 모델링한다. 둘째, 도메인 관련 웹 문서를 수집하고 도메인 온톨로지를 기반으로 키워드의 의미를 분석하고 주석 처리(annotation)한다. 셋째, 사용자의 자연어 질의에 의미 있는 컨텍스트를 추가하여 질의를 확장한다. 넷째, 확장된 질의를 규칙기반 추론엔진을 이용하여 결과를 추론한다. 마지막으로, 사용자 프로파일 분석을 이용하여 선호하는 문서를 우선으로 추천하는 방법을 연구한다. 따라서 본 연구는 질의어에 해당하는 결과문서가 존재하지 않더라도 사용자가 선호하는 문서의 추론이 가능하고, 특정 도메인의 전문가 지식을 추가한 메타 데이터 추론을 통해서 검색 패러다임을 변화시킨다.

1. 서 론

시맨틱 검색(Semantic Search)은 웹 검색 서비스의 자동화를 위해서 웹 문서에 세부 정보를 첨가 시켜 사용자가 원하는 서비스 또는 콘텐츠를 비교적 정확하게 검색하기 위한 패러다임이다 [1]. 본 논문은 도메인에 따른 전문가의 지식이 가미된 정보를 제공하기 위하여 시맨틱 웹을 기반으로 시맨틱 검색 기법에 적합한 모델링(modeling)을 제안한다. 현재까지 제공되는 검색 기술은 텍스트 기반으로 사람의 해석과 판단을 요구하는 방법으로서, 검색엔진은 키워드가 포함된 수없이 많은 웹 문서를 결과로 보여주고, 사용자는 문서를 하나하나 열어 확인하며 원하는 문서를 찾아야 한다. 최근 이러한 서비스 질을 향상시키기 위한 대안으로 시맨틱 웹을 이용한 웹 문서의 내용검색을 위한 모델에 대한 연구가 시도되고 있다.

시맨틱 웹(Semantic Web)은 텍스트 위주에서 벗어나 단어의 유사성과 상관관계 등을 파악해서 결과물을 보여 줄 수 있다. 또한 자원(resource)을 효과적으로 검색하기 위해서 메타 데이터의 개념을 통하여 웹 문서에 시맨틱 정보를 덧붙이고 이를 이용하여 소프트웨어 에이전트가 의미 정보를 자동으로 추출할 수 있도록 한다 [2]. 이런 관점에서 본 연구에서는 OWL 온톨로지 언어를 사용하여 웹 문서를 기계가 처리할 수 있는 형태로 변환한다. 온톨로지는 특정 개념에 대한 의미를 표현하기 위해서 개념과 개념들 간의 관계를 이용한 지식 표현 방식이며, OWL은 온톨로지 언어 중 가장 많은 의미 관계를 표현할 수 있어서 웹 문서에 포함되지 않은 새로운 컨텍스트(context)의 생성을 가능하게 한다 [3].

본 논문은 이러한 기술을 기반으로 정보자원을 비교적 정확하게 검색하기 위한 방법으로 웹 문서에 주석을 추가하여 더 정확한 웹 문서의 내용 검색을 위한 모델을 설계한다. 이러한 검색 모델을 통해서 사용자는 원하는 정보 자원에 보다 쉽고 정확하게 접근할 수 있다. 다음은 본 논문에서 제안하는 사용자 질의에 대한 시맨틱 검색 모델링의 단계에 따른 특징이다. 첫째, 도메인에 따른 정확한 정보 제공을 위해서 OWL 온톨로지를 이용하여 컨텍스트 모델링한다. 도메인별 전문가의 지식 검색을 고려한 모델링은 텍스트 기반 검색에서 추출할 수 없는 결과의 추천을 가능하게 한다. 둘째, 도메인 온톨로지를 기반으로 키워드의 의미를 분석하고 주석 처리하여 메타 데이터를 생성한다. 특정 도메인에 포함되는 웹 문서의 정확한 분류와 키워드는 특정 도메인에서 쓰이는 정확한 의미 추출이 가능하다. 셋째, 사용자의 자연어 질의에 의미 있는 컨텍스트를 추가한다. 따라서 정확한 검색을 위한 의미 있는 질의 생성방법을 제안한다. 넷째, 확장된 질의를 규칙 기반 추론 엔진을 이용하여 사용자가 원하는 결과의 정보를 추론한다. 질의 메타 데이터에 부합되는 결과가 없더라도 전문가의 지식이 포함된 메타 데이터의 계층구조 및 관계를 통한 검색 방법을 연구한다. 마지막으로 사용자 프로파일 분석을 통한 랭크 방법을 제안한다. 따라서 본 연구는 사용자 질의의 의도를 정확히 파악하여 사용자가 선호하는 문서를 찾는 시간을 단축시킬 수 있다.

2. 관련 연구

스탠포드 대학의 TAP[4] 시스템은 시맨틱 웹 기술을 사용하여 문서의 검색 영역을 확장한다. 이 연구의 특징은 시맨틱 기술을 통해서 정보들을 온톨로지처럼 계층 구조와 링크된 정보들로 표현한다. 그리고 난 후 사용자가 입력한 질의와 관련된 정보들을 답변 항목들로 구성한다. 예를 들어 사용자가 Yo-Yo-Ma에 대한 정보를 원할 때, 시스템은 Yo-Yo-Ma에 대한 온톨로지 링크 정보들을 통합하여 결과를 생성하게 된다.

TAP 시스템은 시맨틱 웹 기반에서 정보의 검색을 수행하지

만 온톨로지에 표현된 속성을 위주로 결과를 확장하는 방식이다. 이 방식은 온톨로지 속성만 이용한다는 점에서 본 연구와 차이점을 갖는다. 본 연구는 온톨로지를 이용하여 키워드에 의미를 부여하고 키워드간의 계층적인 구조를 갖도록 카테고리(category)로 분류한다. 이 방법은 의미 있는 데이터의 추론과 웹 문서에서 발견되지 않은 키워드에 대한 검색을 가능하게 한다.

미국의 Maryland 대학에서 연구된 EGTs[5]는 기존의 웹 검색이 갖는 인간이 이해하는 문서를 기계는 이해할 수 없다는 문제점에 대한 해결방안을 제시하였다. EGTs는 자연어로 구성된 질의문을 EGT를 사용하여 BNF(Backus-Naur form) 형태로 재 표현하고 존재하는 웹 페이지들도 EGT를 사용하여 annotation 함으로서 시맨틱 검색을 가능하게 한다. 예를 들어, 코스닥 지수는 얼마인가? (What is the value of Kosdaq?) 라는 질의를 입력했을 때, 일반적인 정보검색으로는 제대로 된 답을 얻을 수 없을 것이다. 하지만 이 연구는 질의를 EGT를 사용해서 <ROBOTGRAM-IN> * [is][Kosdaq*][the](value|quote|price) [of Kosdaq] * </ROBOTGRAM-IN>로 변환하고 annotation된 웹 정보 중에서 중요한 키워드인 value와 같은 EGT 매칭이 일어나는 웹 페이지를 사용자에게 반환하는 형식이다. 하지만 EGTs는 여러 가지 한계점을 갖는다. 첫째, 자연어로 입력된 질의를 정확하게 변환하기 힘들다. 둘째, 오늘의 코스닥 지수는 얼마인가? 와 같은 좀 더 모호한 질의문의 결과를 반환하기 힘들다. 본 연구는 이러한 단점을 해결하기 위해서 도메인에 대한 사전을 이용하여 자연어로 입력된 질의의 의미를 해석하고, 사용자가 입력할 수 있는 키워드를 도메인별 카테고리로 구분하여 메타데이터를 분류함으로써 모호한 질의를 의미를 더 정확히 해석할 수 있다.

3. 메타데이터 기반 시맨틱 검색 시스템 모델링

본 논문에서는 웹 문서의 내용에 주석처리를 하여 시맨틱 검색을 할 수 있는 방법을 제안한다. 제안한 방법은 웹 문서를 주석처리 하여 메타데이터를 생성 하는 단계와 시맨틱 검색하는 단계로 구분 된다. 그림 1은 본 연구에서 제안하는 시맨틱 검색 방법의 전체적인 구조도 이다.

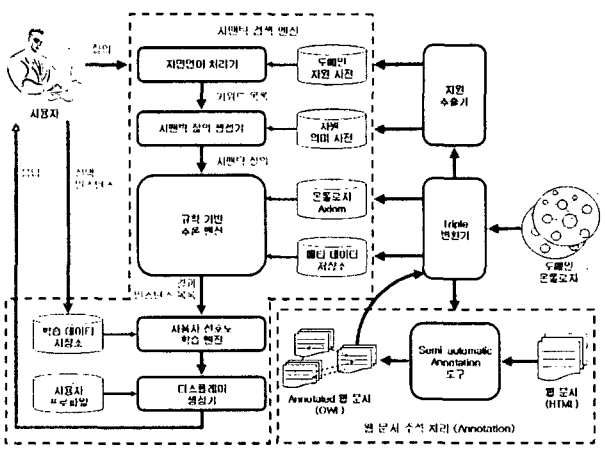

그림 1. 웹 문서 Annotation을 이용한 시맨틱 검색 시스템

3.1 웹 문서의 주석 처리

웹 문서에 주석을 처리하는 단계는 도메인에 따라 모델링한 컨텍스트를 기반으로 웹 문서의 내용에 의미를 부여하여 메타데이터를 생성한다. 본 연구는 컨텍스트 모델링을 위해서 공유와 재사용이 가능한 OWL 온톨로지를 사용한다.

3.1.1 시맨틱 검색을 위한 온톨로지 모델링

본 논문의 온톨로지는 도메인 내에서 사용되는 검색 키워드의 의미를 정의함으로써 시맨틱 검색의 정확도를 향상한다. 시맨틱 검색에서 더 정확한 검색과 효율성의 문제는 도메인 온톨로지를 얼마만큼 정확하고 추론이 가능하게 모델링 할 수 있는가에 있다. 본 연구에서는 사용자의 자연어 질의의 의미를 정확하게 처리하고 더 정확한 검색을 하기 위해서 OWL 온톨로지 언어를 사용하여 온톨로지를 정의한다. 도메인의 범위에 포함되는 질의의 키워드를 클래스(class)와 프로퍼티(property)로 정의하여 의미를 추론할 수 있게 하고, 도메인에 관련된 전문가의 지식이 추론될 수 있도록 온톨로지를 모델링한다.

본 연구의 온톨로지를 이용하면 예를 들어 "사랑을 다룬 여름에 볼만한 협주곡은?" 과 같은 전문가의 지식이 포함된 질의를 처리할 수 있다. 검색 키워드가 되는 클래스는 사랑과 여름이라는 카테고리로 분류되는 감정(emotion)과 계절(seasonal) 클래스, 그리고 협주곡이라는 음악의 장르를 분류하는 장르(genre) 클래스가 된다. 프로퍼티로는 협주곡 중에 감정과 계절의 특성이 사랑과 여름인 것을 선택한다. 이와 같이 온톨로지 상에 정의가 된 키워드에 대해서는 매칭이 가능하지만 온톨로지에 클래스나 프로퍼티로 존재하지 않는 경우에는 키워드로 인식되지 않는다. 본 연구는 이러한 문제를 시맨틱 검색 부분의 질의 확장을 통해서 해결하고, 결과 데이터가 없는 경우를 고려하여 규칙기반 추론엔진을 이용하여 사용자의 요구에 근접한 결과를 제공한다.

3.1.2 메타데이터

메타 데이터는 대량의 정보 중 찾고 있는 정보를 효율적으로 찾아내서 이용하기 위해서 일정한 규칙에 따라 콘텐츠에 대하여 부여되는 데이터이다. 앞에서 언급한 방법으로 도메인 온톨로지를 모델링을 하고 수집한 도메인에 관련된 문서를 semi-automatic semantic annotation 도구를 이용하여 메타데이터로 생성한다. Semantic annotation 도구는 자동과 수동 메타데이터 생성부분으로 나뉜다. 자동으로 메타데이터를 생성하는 부분은 사람의 개입 없이 annotation 시스템을 통해서 메타데이터를 자동으로 생성한다. 예를 들면 문서의 작성일자와 작성자, 문서의 분류, url 등이다. 수동으로 메타데이터를 생성하는 부분은 도메인에 대한 지식을 보유한자가 수집된 도메인에 관련된 문서를 기반으로 하나의 인스턴스에 대해서 메타데이터를 생성한다. 예로는 도메인에 관련된 세부 정보가 포함된다. OWL로 작성된 메타데이터는 추론엔진을 통해서 일차논리형식의 추론을 하기 위해서 <property> <subject> <object> 형태의 트리플 형식으로 변환하여 사용한다.

3.2 메타 데이터 기반 추론을 이용한 시맨틱 검색

시맨틱 검색 단계는 웹 문서가 주석처리 단계를 거쳐 annotation되면 사용자 질의의 의미를 분석하여 사용자가 원하는 웹 문서를 검색한다. 본 연구는 웹 문서의 정확성과 선호도가 높은 문서를 검색하기 위해서 사용자별 프로파일과 학습 데이터를 이용한 온톨로지 추론을 제안한다.

시맨틱 검색 엔진은 사용자의 질의의 의미를 분석하고 콘텐츠가 적합한 웹 문서를 찾아서 이전에 질의 응답한 사용자 히스토리를 분석하여 우선순위를 부여한 결과를 제공한다. 사용자 질의가 입력되면 온톨로지를 기반으로 부여한 자원 사전을 통해서 키워드를 추출한다. 3.1.1 예의 질의와 같이 "사랑을 다룬 여름에 볼만한 협주곡은?"이라는 질의가 입력되면, 도메인 자원 사전을 통해서 온톨로지에 정의된 키워드를 추출한다.

자연어 질의 처리기를 통해서 "사랑", "여름", "협주곡" 자원이 추출되고, 시맨틱 질의 생성기를 통해서 자원 의미 사전에 정의된 자원의 계층정보가 추가된다. 자원에 도메인(domain)과 레인지(range)가 추가되면 시맨틱 추론엔진에서 처리될 수 있는 다음과 같은 predicate 형태의 시맨틱 질의가 생성된다. (mquery (genre concerto) (key1 love) (key2 summer))

본 논문에서의 시맨틱 추론 엔진은 OWL로 생성된 메타 데이터를 KIF(Knowledge Interchange Format) 형태로 읽어 들여서 전방향 추론을 수행한다. OWL에서 정의된 공리는 일차논리(First Order Logic) 표현 방식을 따르고 있다. 따라서 다음으로 KIF형태를 술어(predicate) 형태로 변환하는 작업이 필요하다. 규칙기반 추론엔진은 JESS 언어를 기반으로 하고 Jena의 RDF 파서를 사용하여 술어형태로 변환 한다 [6, 7, 8, 9].

추론엔진으로부터 질의에 맞는 결과가 생성되면, 사용자 선호도 학습 엔진을 통해서 사용자가 원하는 결과에 우선순위를 두어 출력한다. 이 페이지 랭크기법은 사용자 행동을 모델링한 것으로써 일반적인 사용자가 생각하는 특정 페이지의 중요성과 잘 상응하는 객관적인 측정치이다. 본 연구의 페이지 랭크는 단순히 모든 링크를 세는 것에서 더욱 확장해서 사용자의 선호도를 고려한다. 사용자 선호도 학습 엔진은 문서가 갖는 카테고리와 콘텐츠에 대한 정보를 사용자의 정보검색 행위를 모니터링 하여 질의별로 사용자의 특성을 학습한다. 마지막으로, 디스플레이 생성기는 학습된 데이터를 기반으로 사용자의 개인정보를 고려하여 결과를 제공한다.

4. 실 험

본 연구는 실제적인 시맨틱 검색 시스템의 구현을 위해서 도메인을 문화 콘텐츠로 한정하였다. 다음은 클래식 공연에 대한 시맨틱 검색 결과이다.

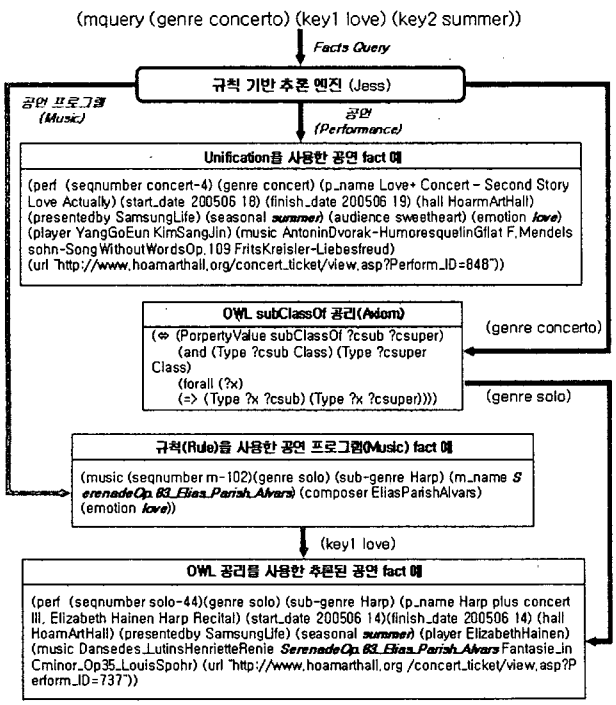

그림 2. 메타데이터 기반의 규칙 기반 추론엔진을 이용한 검색

그림 2는 3.2절에서 생성한 시맨틱 질의를 규칙 기반 추론엔진을 이용해서 결과를 생성하는 과정이다. 이 방법은 사용자의 질의에 웹 문서에 포함되는 텍스트가 없더라도 원하는 결과를 추론할 수 있다. 그림 2는 시맨틱 질의의 fact를 포함하는 공연의 문서와 그에 관련된 문서에 대한 추론, 그리고 지식 추론에 의한 추론 결과를 보여준다. 공연검색에 의한 결과는 웹 문서 annotation시 추가된 공연 메타 데이터에 대해서 일치된 결과이고, 관련 있는 문서의 추론은 OWL axiom을 이용한 concert의 하위 클래스에 대한 추론 결과이다. 마지막으로 지식추론에 의한 결과는 음악에 대해 추가된 카테고리 메타데이터를 통해서 추론된 문서이다. 카테고리는 연주곡, 계절, 감정, 나이, 기간, 그룹, 이벤트, 장소, 악단에 대한 지식 검색이 가능하도록 공연정보에 추가하였다. 이 결과는 사랑과 계절에 관련되는 문서로써 사용자 선호도 학습 엔진을 거치면서 사용자의 선호도를 고려하여 우선순위가 부여 되어 출력된다.

5. 결론 및 향후 연구

본 논문에서는 정보 검색 시스템의 구축에 있어서 온톨로지를 사용하여 문서들을 annotation하고 메타데이터를 생성함으로써 시맨틱 검색을 유도하고, 사용자 요구에 해당하는 시맨틱 질의를 자동 생성하여 검색의 정확성과 효율성을 높이는 방법을 제안한다. 본 논문의 시맨틱 검색을 위한 웹 문서 annotation 방법은 문서의 내용을 도메인 온톨로지에 맞게 해석하여 메타 데이터로 생성하고, 사용자 질의 의도에 맞는 최적의 결과 문서를 찾는데 효과적이다. 그러나 주석처리를 누가 하느냐에 따라 웹 문서의 특성이 다르게 부여될 수 있다. 향후 연구에서는 구체적인 온톨로지를 기반으로 동적으로 메타 데이터를 생성하는 부분과 검색결과의 랭크 방법을 확장하여 질의에 따른 결과의 추천이 연구 되었을 때 현재 웹 검색 방법에서 시맨틱 웹 검색으로의 변환을 고려해 볼 수 있을 것이다.

6. 참고 문헌

[1] Guha, R. and McCool, Rob and Miller, Eric, "Semantic Search," International WWW Conference 2003, pp. 20-24, 2003.

[2] Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web," Scientific American, May 2001.

[3] F. Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider, and L. Stein. Web Ontology Language (OWL) Reference Version 1.0. 2004.

[4] R.Guha and R. McCool. Tap: Towards a web of data. http://tap.stanford.edu/

[5] Gautham K.Dorai, Yaser Yacoob, "Facilitating Semantic Web Search with Embedded Grammar Tags,", University of Maryland-College Park, 2003.

[6] "Knowledge Interchange Format", draft proposed American National Standard (dpANS), http://logic.stanford.edu/kif/dpans.html

[7] Friedman-Hill, E. J., *Jess In Action*, Manning Press, 2003.

[8] Andy Seaborne, "Jena Tutorial: A Programmer's Introduction to RDQL," April 2002.

[9] Frank Manola, Eric Miller, "RDF Primer," W3C Working Draft 23 January 2003.