Journal of Digital Contents Society Vol. 15 No. 5 Oct. 2014(pp. 603-612)
http://dx.doi.org/10.9728/dcs.2014.15.5.603

온톨로지를 이용한 웹문서의 시맨틱 검색

오성균*, 김병곤**

요 약

사용자들에게 좀 더 정확하고 편리한 검색결과를 제공하기 위하여 정보의 구조적인 특징 등을 사용하는 시맨틱 검색의 개념이 널리 연구되고 있다. 이를 위하여, 최근의 정보검색분야와 데이터구축 분야의 연구에서는 데이터의 구조적인 표현과 검색 메카니즘을 구현하기 위하여 온톨로지를 강조하고 있다. 본 연구에서는 웹 환경에서의 검색 정확도와 만족도를 향상시키기 위하여 온톨로지를 이용한 시맨틱 검색 방법을 제안한다. 온톨로지와 KB(KnowledgeBase)를 이용하여 검색 대상을 키워드간의 관계를 유추한 사실(fact)과 관계키워드들을 지니는 웹문서들로 크게 나누고 이들을 서로 유기적으로 검색을 진행하는 시맨틱 검색 질의 처리기법을 제안하였다. 또한 결과에 대한 사용자의 검색 만족도를 높이기 위하여 결과 문서와 사실에 대한 랭킹 방법을 제안하였다. 실험을 통하여 주어진 식의 값을 달리하여 랭킹을 올바로 구현하는 요소로 키워드의 빈도와 온톨로지상의 클래스 레벨이 영향을 미치는 것을 확인 할 수 있었고, 이를 통하여 적합한 형태의 계수 값을 제시하였다.

키워드 : 시맨틱웹, 온톨로지, KB, RDF, 랭킹

Semantic search of web documents using ontology

Sung-Kyun Oh*, Byung-gon Kim**

Abstract

To provide efficient and correct search results, ontology which use the structure of information, is considered as a main mechanism in the semantic web. Therefore, recent research in information retrieval and data construction have emphasized the use of ontologies as a data representation and search mechanism. In this paper, we propose a semantic search method using ontology to improve search ability in web environment. Ontology and knowledge base is used to represent semantic meaning of the data and provide related web documents and facts as results. Also, search result ranking mechanism is proposed. The mechanism use cardinality of the keyword in the contents and structural information of ontology. Experimental results with several query processing indicate that different coefficient value in the expression gives different results in sample ontology system and we propose appropriate values of the coefficient.

Keywords : Semantic Web, Ontology, KB, RDF, Ranking

1. 서론

인터넷을 이용한 정보 검색이 일반화되면서, 이를 이용한 수많은 응용 프로그램과 웹서비스가 더욱 활발히 제공되고 있다. 현재의 웹정보시스템은 대부분 HTML문서 위주의 데이터베이스 검색 메카니즘을 사용하고 있다. 기본적으로 키워드의 조합을 가지고 검색하고, 검색 결과를 중요도에 의해서 순위를 정한다. HTML은 사용 및 관리가 용이하여 웹서비스가 짧은 시간에 일반화 보편화 되는데 크게 기여하였다. 그러나,

※ 교신저자(Corresponding Author): Byung-gon Kim
접수일:2014년 08월 26일, 수정일:2014년 10월 10일
완료일:2014년 10월 24일
* 서일대학교 컴퓨터소프트웨어과
Tel: +82-2-490-7398 , Fax: +82-2-490-7396
email: skoh@seoil.ac.kr
** 부천대학교 e-비즈니스과
Tel: +82-32-610-3460 , Fax: +82-32-610-3207
email: bgkim@bc.ac.kr
▣ 본 논문은 2013년도 서일대학 학술연구비에 의해 수행되었음.

반면에 단순한 구조와 데이터의 의미를 표현하지 못하는 한계성 때문에 더 이상의 기능적 성장을 기대하기는 어려운 상황이다. 그러나, 인터넷상의 데이터의 양이 기하급수적으로 팽창하면서, 너무 많은 데이터를 사용자가 직접 판단하고 분석하기 어려운 상황이 되어 가고 있다. 이러한 한계성을 극복할 수 있는 해결 방안의 하나로 제시되고 있는 것이 시맨틱 검색이며, 시맨틱 검색을 구현하는 웹환경이 시맨틱웹이다. 또한 시맨틱웹의 가장 핵심적인 구성요소가 온톨로지이다.

시맨틱 검색에 대한 정확한 정의는 존재하지 않지만 웹에서의 보다 진보된 형태의 검색의 개념을 의미한다. 사용자의 웹 검색 질의와 다른 형태의 웹데이터로 부터 의미와 구조를 추출하고 이를 이용하여 웹을 검색하는 방식이다. 즉, 검색자의 의도를 파악하고, 검색어의 문맥상의 의미를 이해하여 검색 정확도를 향상 시키는 검색을 의미한다.

이를 위하여 필요한 환경이 시맨틱웹 환경이다. 시맨틱웹은 현재의 인터넷 시스템처럼 검색된 결과를 사람이 눈으로 보고 확인하여 처리하는 시스템이 아닌, 컴퓨터가 스스로 읽고(machine readable) 스스로 이해하여(machine understandable) 처리하는 인터넷 환경을 의미한다.[1] 컴퓨터가 정보를 검색하고 해석하여 사용자에게 필요한 정보만을 추출한 후 이를 가공하여 사용자의 의도에 맞는 예약이나 구매와 같은 정보를 제공하여 인터넷 사용자의 많은 노력을 줄여주는 형태의 인터넷 사용 환경을 제공하게 된다. 따라서, 컴퓨터가 데이터를 이해하고 처리하기 위한 웹 콘텐츠에 대한 표현 구조와 처리 방법을 제시하는 것이 필요하다. 즉, 시맨틱웹의 문서는 자연어 위주의 기존의 웹 문서가 아닌 자동화된 에이전트나 검색엔진이 자동으로 해석 가능하도록 구성되어야 한다. 이를 위하여 필요한 개념이 온톨로지이다.

온톨로지를 통하여 웹검색의 정확도를 향상시킬 수 있으며, 웹페이지의 정보와 연관된 다른 정보와의 추론을 할 수 있다. 다양한 소스로 부터 웹콘텐츠를 수집하는 에이전트 프로그램이 다른 프로그램으로부터의 결과를 서로 교환하여 처리가 가능해지면 상당히 향상된 형태의 에이전트가 가능해진다. 이처럼 온톨로지를 이용한 소프트웨어 에이전트를 통하여 기계가 직접 웹 콘텐츠를 읽고 자동으로 서비스가 가능해지면 인터넷 정보처리 능력이 기하급수적으로 향상될 수 있다.

이를 위하여 정보 자원의 의미와 다른 정보 자원과의 의미적 관계를 이용하여 시맨틱 검색을 처리하는 연구가 필요하다. 또한, 검색된 정보자원들에 대하여 유사도와 정확도의 개념을 도입하여 랭킹을 제공하는 기법이 필요하다. 본 연구는 온톨로지의 시맨틱 검색을 위하여 사람이 정보를 검색하는 것과 같이 의미에 기반을 다단계의 추적 방식을 적용하여 검색자의 만족도를 높이고, 검색된 자원의 랭킹을 제공하는 방식의 시맨틱 검색을 제안하였다.

본 논문은 다음과 같이 구성된다. 2장과 3장에서는 시맨틱 검색을 위해 필수적인 개념인 시맨틱웹 언어와 온톨로지에 대하여 설명하였다. 시맨틱웹 언어를 이용하여 온톨로지가 구축되는 개념을 설명하였다. 4장에서는 기존의 시맨틱 검색 시스템에 대하여 설명하고, 이와 차별하여 본 연구에서 제시한 키워드간의 관계를 유추한 사실(fact)과 관계키워드들을 지니는 웹문서들로 크게 나누고 서로 유기적으로 검색을 진행하는 시맨틱 검색의 개념을 이용한 질의 처리기법을 설명하였다. 5장에서는 이를 이용하여 사용자의 검색 만족도를 높이기 위하여 결과 문서와 사실에 대한 랭킹 방법을 제안하였다. 6장에서는 실험평가 결과를 제시하고 7장에서 결론에 대하여 언급하였다.

2. 시맨틱웹과 온톨로지

시맨틱 웹상에서 중요한 역할을 담당하는 메타데이터와 온톨로지의 데이터에 의미를 부여하기 위해서는 데이터 자체를 표현하고, 나아가 데이터간의 관계를 가지고 유추하여 추론할 수 있는 표현 방법이 필요하다. 이를 위하여 사용되는 언어는 XML, RDF, OWL 등이 있으며, W3C에서 주도적으로 연구가 진행되고 있다.

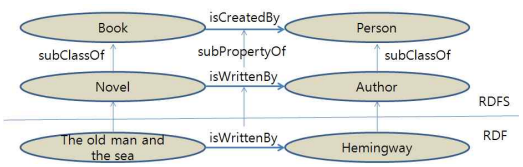
XML은 HTML과 같이 태그를 사용하는 마크업 언어이지만 콘텐츠에 대한 주석을 태그 형태로 생성할 수 있다. 따라서, 문서의 내용에 대한 정보를 구조적으로 구축할 수 있으며, 스크립트

나 프로그램은 이러한 태그를 이용할 수 있다. 그러나, XML의 태그와 정보가 그 정보의 실제적인 의미를 설명해주지는 않는다.

XML의 표현 방식에서 더 나아가 정보의 실제적인 의미를 표현하기 위하여 사용하는 언어가 RDF이다. RDF는 문장을 주어(Subject), 서술어(Verb), 목적어(Object)와 같은 트리플 구조로 표현한다. 주어, 서술어, 목적어는 리소스(Resource), 프로퍼티(Property), 값(Value)으로도 표현 한다.[2]

이러한 RDF 문서구조는 리소스, 프로퍼티, 값이라는 기본 구조를 이용해 메타데이터를 표현한다. 그러나 표현하고자 하는 리소스의 의미를 표현하는데 있어 모호성이 존재하기 때문에 이러한 문제를 해결할 필요가 있다. 이러한 사실들을 표현하고 이용하기 위하여 필요한 것이 스키마의 개념이고 이를 위하여 만들어진 것이 RDFS(RDF Schema)이다.[3] RDFS는 RDF를 기술하기 위해 필요한 키워드들의 의미와 키워드들 간의 의미적 관계를 기술하기 위해 W3C에서 초기에 제안한 언어이며, RDF와 같은 문법적 구조를 가지는 것이 장점이다.

(그림 1) RDF와 RDFS 표현 그래프



(Figure 1) Representation graph of RDF and RDFS

(그림 1)은 문장 “The old man and the sea is written by Hemingway″을 RDF로 표현한 그래프와 이와 연관된 RDFS 그래프의 예이다. RDFS부분에서는 실세계의 개체인 “Book”, “Person”, “Novel”, “Author” “Book”과 이들 간의 관계를 정의하고 있고, RDF는 실제 인스턴스 문장을 표현하고 있다. RDFS에서는 “Book”과 같은 개체를 클래스라 하고 “isCreatedBy”와 같은 관계를 프로퍼티라고 부른다. “Novel” 클래스는 “Book” 클래스에 속하는 서브 클래스의 개념이라는 것을 정의하고 있으며, ″Person″과 ″Author″의 관계도 서브 클래스의 관계에 있음을 정의하고 있다. 이러한 관계는 “isCreatedBy″와 ”isWrittenBy″와 같은 프로퍼티 간에도 성립한다.

이처럼 RDFS를 이용해 정의된 개념과 개념들 간의 관계를 이용하여 RDF로 기술된 메타데이터의 의미를 좀 더 명확하게 표현할 수 있다. 즉, ”The old man and the sea″가 “책”의 분류 중에서 “소설”에 해당한다든지, “Hemingway”가 “작가”이면서 동시에 “사람”이라는 사실을 표현할 수 있다.

이외에 스키마 표현 언어로는 DAML+OIL과 OWL 등이 있다. DAML+OIL은 RDF와 RDFS 표준에 기반을 두고 확장한 프레임 기반의 온톨로지 표현 언어이다. 서술논리(Description Logic)의 추론 능력과 표현력을 가지며, 인공 지능적 측면에서의 관련 연구에 활용이 가능하다. OWL은 DAML+OIL 언어를 기반으로 확장한 언어 이다. DAML+OIL의 네임스페이스와 속성 클래스 이름 등을 변경하여 사용하며, RDF 및 RDFS의 변화를 수용하여 현재 온톨로지 개발에 많이 사용되고 있는 언어이다.

온톨로지는 앞에서 언급한 대로 인터넷상의 문서에 포함된 개체들에 대한 개념과 개체들 간의 관계를 정의한 것이다. 온톨로지의 정의를 살펴보면 “공유된 개념에 대한 형식적이고 명시적인 설명 혹은 서술(A formal explicit specification of a shared body of concepts)”이다.[4] 이에 대한 각각의 의미는 다음과 같다.

- 공유된 개념(shared body of concepts) : 실세계에 존재하는 개체나 현상 등에 대하여 혼자만들고 혼자 사용하는 개념이 아닌 특정 그룹내의 구성원들이 동의하여 함께 사용되는 개념이다.
- 형식적(formal) : 개념에 대한 정의를 기계가 스스로 읽고 해석할 수 있는 형태이여야 한다.
- 명시적(explicit) : 개념과 더불어 개념에 대한 제약 조건 등이 정의되어 있어야 한다.

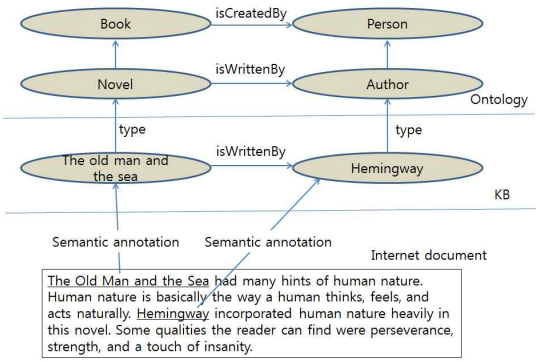
온톨로지를 구성하는 구성요소로는 개념(concepts), 관계(relations), 인스턴스(instances), 공리(axioms)가 있다. 개념은 도메인에 존재하는 엔티티의 집합을 의미하며, 관계는 개념들 혹은

개념의 성질(Property)간의 상호 관계를 의미한다. 인스턴스는 도메인의 개념을 반영하는 실세계의 실제 예이며, 공리는 개념의 사용을 제약하는 명시적인 규칙을 의미한다. 예를 들어, 와인이라는 개념을 온톨로지로 표현하고자 한다면, 먼저 와인, 와이너리, 보르도, 보졸레와 같은 개념을 정의한 후, “와이너리는 와인을 생산한다.” 와 같은 관계들을 정립한다. 다음으로 “와이너리는 반드시 하나 이상의 와인을 생산해야 한다.” 등의 공리규칙을 명시할 수 있다.

온톨로지는 실제 웹상에서 사람이나 소프트웨어 에이전트로 하여금 개념에 대한 구조를 제공하여 공유할 수 있으며, 이는 온톨로지를 통하여 구축된 지식을 다른 사람이나 그룹에서 재사용이 가능하도록 할 수 있다. 온톨로지가 완성되면 에이전트는 이를 이용하여 사용자가 원하는 검색결과나 처리를 제공할 수 있다. 예를 들면, 와인 온톨로지를 이용하여 식당 고객이 가장 원하는 와인을 정확하게 추천할 수 있다. 이와 같이 사람과 사람간의 의사소통을 위해서 온톨로지를 이용할 수 있지만, 더 넓은 범위에서는 시스템과 시스템간의 상호작용과 시스템 공학 측면에서의 이용이 가능하다. 즉, 시스템이 요구하는 도메인의 요구사항 명시, 일관성체크, 정보획득 등에 사용될 수 있다.

온톨로지는 주석의 개념으로 웹상의 문서와 관계를 맺는다. 주석은 데이터의 내용과 구조를 모두 표현하는 자기 기술적 특징을 가지고 있고 문서 내용에 대한 의미 정보를 나타낼 수 있으므로 논리적 구조 및 내용 정보를 이용하여 보다 정확한 검색이 가능하다. 웹문서에 대한 온톨로지 기반의 검색은 규칙 추론과 같이 정의된 규칙에 의해 기존에 존재하지 않는 새로운 정보를 생성하는 기능과 관계 추론과 같이 온톨로지 관계(상하관계, 유사개념 관계 등)를 이용하여 검색의 범위를 넓히거나 좁히는 기능, 그리고 온톨로지의 내용을 검색결과로 출력하고 온톨로지의 관계를 계속 브라우징 할 수 있는 기능 등을 이용하여 검색이 수행된다.

(그림 2) 온톨로지를 사용한 검색환경



(Figure 2) Search environment using ontology

본 연구에서는 (그림 2)에 나타난 바와 같이 온톨로지를 이용한 검색을 수행하기 위한 환경을 다음과 같이 가정한다. 검색의 대상은 웹문서이고, 시맨틱주석(Semantic annotation)은 인터넷 문서에 나타난 개체 즉, 단어에 대하여 온톨로지와 KB(Knowlwdge Base)의 정보와 연결하는 단계를 의미한다. 그림은 인터넷 문서의 일부분에 나타난 단어에 대하여 시맨틱주석을 연결한 것을 보여준다. “The old man and the sea″와 ”Hemingway″라는 단어는 문서상에서는 별다른 연관 관계를 추론할 수 없지만, 시맨틱 주석에 따라 KB와 온톨로지를 따라가면 “isWrittenBy″의 관계를 유추할 수 있다. 추가로 ”Hemingway″의 다른 작품들의 정보를 수합하여 다른 웹문서의 정보를 질의의 결과로 제공할 수 있게 된다. 또한, KB에 존재하는 “The old man and the sea″와 관련된 다른 사실들 예를 들면, 소설을 소재로 한 영화나 소설에 대한 서평 등의 정보들을 검색의 결과로 제공할 수 있게 된다. 이처럼 온톨로지와 KB를 바탕으로 구현된 검색 방법은 기존의 키워드 인덱스 방식의 검색과 상당한 차이를 지닌다. 가장 큰 차이점은 정보의 나열이 아닌, 개체간의 관계와 유추를 통해 검색의 범위를 넓힐 수 있으며, 새로운 사실들을 검색의 결과로 제공하게 된다는 것이다.

본 연구에서 제시하는 시맨틱 검색의 개념은 지금 설명한 것처럼 온톨로지를 이용한 검색 메카니즘과 검색 결과에 대한 랭킹 기법이다. 다음 장에서는 기존의 시맨틱 검색 시스템을 살펴보고, 새로운 개념의 시맨틱 검색과 랭킹 기법을 제시한다. 온톨로지를 이용한 응용 연구는 추론

기능을 이용한 검색에 대한 분야 등에서 진행되고 있다.[5]

3. 시맨틱 검색 시스템

앞서 언급한대로 시맨틱 검색에 대한 정의가 정확하게 존재하지는 않지만 지금까지 각 시스템별로 각자의 개념을 가지고 연구를 이어 왔다.

SHOE시스템[6]은 웹에 시맨틱한 질의를 처리하는 초기의 시도이다. 웹페이지에 주석을 다는 툴을 제공하고, 사용자로 하여금 온톨로지와 클래스와 프로퍼티를 선택하여 마크업을 추가하도록 하며, 웹크롤러는 이를 바탕으로 검색을 수행하는 방식을 사용하였다. 추론 엔진(inference engine)은 추론 규칙을 통하여 새로운 마크업을 제공하는 역할을 한다. 여러 질의 툴은 사용자가 온톨로지에 대하여 구조적인 질의가 가능하도록 설계하였다. 검색 툴은 온톨로지로부터 클래스와 프로퍼티를 선택하여 KB에 질의를 던지고 결과를 테이블 형태로 받도록 하였다.

Swoogle시스템[7]은 RDF를 기본으로 하는 접근하는 방식이며, RDF 문서를 대상으로 인덱스하고 질의하고 검색하는 웹크롤러방식의 검색시스템이다. 기본적으로 시맨틱웹 개념의 문서에 대하여 검색을 수행하며, 문서단계에서의 메타데이터의 조건들을 지니는 질의들을 명시하도록 한다. 검색된 문서들은 시맨틱웹의 관점에서 중요도에 따라 랭킹이 되도록 설계하였다.

Corese시스템[8]도 시맨틱웹에서 온톨로지 개념을 이용하는 검색엔진이다. RDF 형태로 주석되어 있는 웹자원을 검색하는데 RDF 형태의 질의어를 사용한다. 검색 수행시 추론규칙을 사용하는데, 온톨로지에서의 클래스와 프로퍼티 간의 시맨틱한 거리를 계산하여 적용한다.

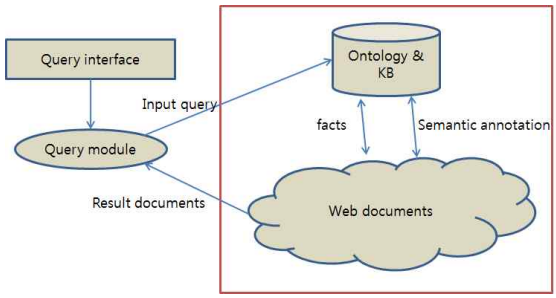
ONTOSEARCH2[9]는 시맨틱웹의 온톨로지에 대한 검색과 질의 엔진이다. SPARQL[10]을 이용하여 질의하고 온톨로지의 인스턴스와 구조에 대한 질의를 모두 제공한다.

NAGA 시맨틱 검색 엔진[11]은 그래프 형태의 질의 언어를 사용한다. KB를 그래프 형태로 나타내며, KB는 웹자원으로 부터 지식을 추출하는 툴에 의하여 자동으로 형성되도록 설계하였다. 그래프의 노드와 에지는 엔티티와 엔티티간의 관계를 나타내고, 질의어는 SPARQL을 확장하여 복잡한 그래프 질의를 사용한다. 질의의 답은 지식 그래프의 부분 그래프로 나타나며 가중치를 적용한 그래프를 이용하여 순위를 매기는 방식을 사용하였다.

4. 시맨틱 검색 제안

본 연구에서 제시하는 시맨틱 검색의 개념은 온톨로지와 KB를 바탕으로 시맨틱주석으로 연결된 정보들을 통하여, 사용자가 원하는 정보들을 추출하여 제공하는 것을 목표로 한다. 위에서 언급된 연구들과 달리 제공되는 정보들을 키워드간의 관계를 유추한 사실(fact)과 관계키워드들을 지니는 웹문서들로 크게 나누고 서로 유기적으로 검색을 진행하도록 설계하였다.

(그림 3) 시맨틱 검색 개념



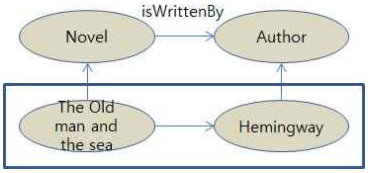
(Figure 3) Semantic search concept

(그림 3)은 사용자의 질의인터페이스로부터 생성된 질의가 질의모듈을 거쳐 온톨로지와 KB에 대하여 사실(fact)이 검색되고 검색된 사실을 가지고 웹을 검색하는 개념의 다이어그램이다. 예를 들어, 미국의 작가 “헤밍웨이”와 그의 작품 “노인과 바다”를 주제로 검색을 수행한다고 가정하면, 시맨틱 검색 절차는 다음과 같이 수행된다.

1단계 : 입력질의는 사용자의 질문으로부터 생성된다. 생성된 질의는 먼저 온톨로지와 KB를 대상으로 실행된다. 질의중의 키워드에 대하여 온톨로지와 KB로부터 Author와 Novel 개체클래스를 검색하고, 이와 연결된 인스턴스를 가지고 인스턴스들 간의 사실들을 찾아낸다. 즉, “The

old man and the sea"라는 소설은 "Hemingway"라는 작가가 쓴 작품이라는 사실을 알 수 있다.

(그림 4) 검색된 사실

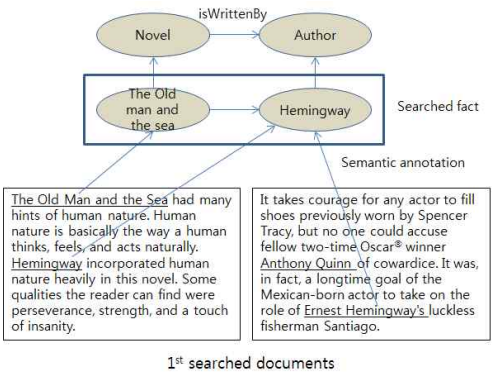


(Figure 4) Searched fact

2단계 : 1단계의 KB에서 검색된 사실들과 관계를 맺고 있는 사실들을 검색한다. 검색 환경에 나타난바와 같이 특정 사실과 직접적인 연관 관계를 맺고 있는 사실은 질의 결과에 필요한 정보를 지니고 있을 가능성이 높다.

3단계 : 1, 2단계의 결과로 질의의 조건에 만족하는 KB의 사실(fact)들이 반환되고, 반환된 사실들을 시맨틱 주석으로 지니는 웹문서들을 검색한다. (그림 5)는 검색된 사실을 시맨틱 주석으로 지니는 문서들을 보여준다.

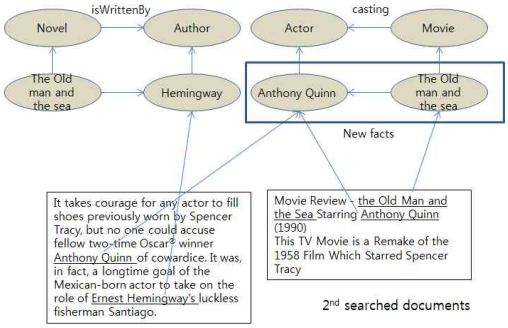
(그림 5) 검색된 사실을 시맨틱 주석으로 지니는 문서들



(Figere 5) Documents containing searched fact

4단계 : 1차 검색 문서를 대상으로 2차 검색을 수행한다. 2차 검색의 의미는 더 많은 연관 문서를 검색하여, 새로이 검색된 연관 문서를 통하여 새로운 사실을 추출하는 것이다. (그림 6)에서는 "Anthony Quinn"을 주석으로 지니는 사실을 추출하는 과정을 보여준다. 새로운 사실은 기존의 사실과 결합하여 새로운 사실을 추가하여, 제공하는데 사용된다.

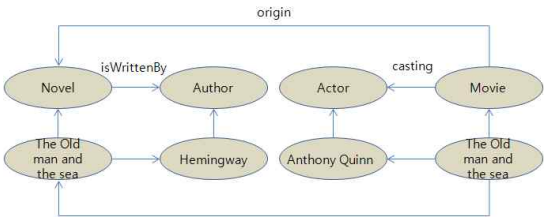
(그림 6) 1차 검색 문서를 통한 2차 문서 검색



(Figure 6) 2nd search through 1st search

5단계 : 4단계에서 검색된 사실들을 가지고, 온톨로지의 관계를 바탕으로 새로운 사실을 추가하여 결과에 반영한다. (그림 7)은 검색된 사실들을 통하여 "Hemingway"의 소설 "The old man and the sea"를 원작으로 하는 영화 "The old man and the sea"의 주연 배우가 "안소니 퀸"이라는 사실을 보여준다.

(그림 7) 검색된 사실들을 통한 결과 도출



(Figure 7) Results from searched facts

검색의 결과물은 사실들을 통하여 확인된 사실들과 인터넷문서들이다. 이들 결과물은 사용자의 의도에 따라 여러 가지 형태로 제공하도록 할 수 있다. 이때, 검색된 결과물들 즉 사실과 인터넷 문서들의 랭킹을 통하여 검색의 정확도와 제공 순위를 결정 할 수 있다. 다음 장에서는 이를 위한 랭킹 기법을 제시하도록 한다.

5. 검색자원 랭킹 기법

앞에서 제안한 시맨틱검색 방법으로 질의를 수행하면 주제와 직접 혹은 간접적으로 연관된 사실들과 인터넷 문서들을 검색할 수 있다. 검색 결과에 직접적으로 연관된 자원뿐 아니라 간접적으로 연관된 자원들도 검색되기 때문에 검색되는 정보의 양이 상당히 많을 수 있다. 따라서, 사용자의 검색 만족도를 높이기 위해서는 검색 결과에 대한 랭킹(Ranking)기법이 필요하다. 랭킹의 의미는 질의 주제와 관련성이 높고 사용자가 원하는 정보를 많이 포함하고 있는 문서를 검색 결과의 상위에 위치 시켜서 검색 만족도를 높이고자 하는 것이다.

본 논문에서 제안한 랭킹 기준은 먼저, 검색문서중에 포함된 질의 키워드의 빈도수를 기준으로 제시한다. 질의 키워드를 많이 포함한 문서가 적게 포함된 문서에 비하여 주제와의 연관성이 높다고 보는 것이다. 이를 위하여, 키워드 k에 대한 문서 d의 문서중요도(DocumentFactor)를 식으로 표현하면 다음과 같다.

$$DocumentFactor(d, k) = KeywordFreq(d, k) / maxKeywordFreq(D, k) \quad (1)$$

KeywordFreq(d, k)는 문서 d중에 질의키워드 k가 출현하는 빈도수(frequency)를 의미한다. maxKeywordFreq(D, k)는 도메인 문서의 집합 D의 문서들에서 출현하는 질의 키워드 k의 빈도수중에서 가장 많은 빈도수를 의미한다. KeywordFreq(d, k)를 maxKeywordFreq(D, k)로 나누어 키워드에 대한 중요도를 표현한다. 따라서, 키워드에 대한 중요도는 0부터 1사이의 값을 가지게 된다.

다음으로 제시하는 기준은 도메인중 키워드 중요도이다. 도메인에 존재하는 수많은 문서에는 많은 키워드들이 있다. 많은 키워드들 중에는 중요하고 비중 있는 키워드들도 있고, 중요하지 않고 문서 중에 거의 출현하지 않는 키워드들도 있다. 따라서, 문서 중에 출현한 횟수가 많은 키워드가 횟수가 적은 키워드에 비하여 중요도가 높다고 판단하고자 한다. 키워드중요도에서 한가지 요소를 더 추가한다. 키워드가 연결되어 있는 온톨로지의 클래스의 레벨을 고려한다. 클래스 계층구조에서 하위 클래스가 상위클래스에 비교하여 좀 더 자세하고 명확한 개념을 지니므로, 하위 클래스에 소속한 키워드에 대하여 더 높은 중요도를 부여한다. 이를 위하여, 도메인중 키워드중요도를 다음과 같이 제시한다.

$$KeywordFactor(k) = a * NumberOfDoc(k)/NumberOfDoc(Domain) + (1-a) * classLevel(k)/maxClassLevel(ontology) \quad (2)$$

NumberOfDoc(k)는 키워드 k가 출현한 문서의 수를 의미하며, NumberOfDoc(Domain)는 전체 도메인에서 가장 많이 출현하는 키워드의 문서수를 의미한다. classLevel(k)는 키워드 k의 클래스계층 레벨을 의미하며, maxClassLevel(ontology)는 온톨로지 계층구조에서 가장 높은 레벨을 지니는 클래스 레벨을 의미한다. α는 0부터 1 사이의 값을 배정한다. 따라서, 도메인중 키워드중요도는 0부터 1사이의 값을 가지게 된다.

키워드에 대한 문서중요도와 도메인중 키워드 중요도를 이용하여 키워드집합 K의 원소를 지닌 문서 d의 문서랭킹(document ranking)를 식(1)과 식(2)를 이용하여 표현하면 다음과 같다.

$$documentRanking(d, K) = \sum_{k=1 \text{ to } n} (DocumentFactor(d, k) * KeywordFactor(k)) \quad (3)$$

식(3)에 나타난 바와 같이 문서 d의 중요도는 더 많은 키워드를 지니면서, 동시에 중요한 키워드를 더 많이 지니는 문서일수록 높게 나타나도록 하였다.

검색을 통하여 추출된 여러 사실(fact)의 우선순위는 다음과 같이 정한다. 가장 많은 문서에 링크된 사실이 우선순위가 높고, 낮은 레벨의 인스턴스를 기반으로 된 사실이 더 자세한 정보를 지니는 것으로 간주한다.

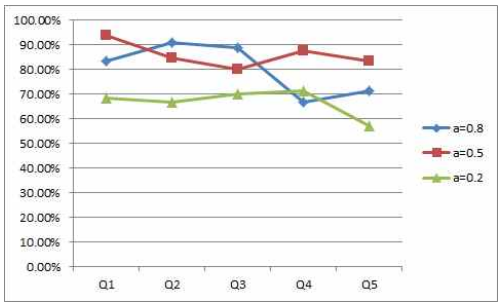
$$FactFactor(f) = \beta * link(f,D)/maxLink(D) + (1-\beta) * classLevel(f)/maxClassLevel(ontology) \quad (4)$$

6. 실험

본 논문에서 제안한 온톨로지를 이용한 시맨틱 검색과 검색 결과에 대한 랭킹 결정 알고리즘을 시뮬레이션 형태로 검증하기 위하여 영화 도메인에 대하여 온톨로지와 메타데이터를 OWL 문서 형태로 구축하였다. 영화온톨로지(Movieontology.org)를 참조하여 영화와 관련된 개념에 대한 76개의 클래스와 42개의 프로퍼티를 정의하고, 클래스간의 의미적인 관계를 고려하여 계층관계를 정립하였다. 정의된 온톨로지와 관련하여 실제 오브젝트들로 구성된 트리플 형태의 사실들을 형성하고 이를 온톨로지와 연결하여 KB를 구성하였다. 또한, 한국 영화와 관련된 국내 웹상의 500개 이상의 문서 페이지들을 수집하여 문서중의 키워드를 KB내의 오브젝트와 연결하였다.

실험은 구성된 데이터 집합에 대하여 영화 제목과 감독, 배우 등의 키워드를 조합한 형태의 Q1~Q5 질의를 생성하여 질의를 수행하고, 반환된 문서들에 대하여 문서랭킹 값을 기준으로 랭킹을 부여한 후에, 이 결과에 대하여 평가를 하였다. 랭킹에 의하여 순서화된 문서들의 정확성을 판단하기 위하여 정확률(Precision)과 재현률(Recall)을 사용하였다. 정확률은 검색된 결과 집합 중에서 얼마나 많은 정답이 포함되어 있는지를 나타내는 지표이며, 전체 검색 결과중에서 정답문서의 수로 계산하였다. 재현률이란 정답 혹은 적합 문서 집합을 현재 결과 집합과 비교하여 얼마나 누락이 없는지를 나타내는 지표이며, 전체 정답 문서중에서 검색된 정답문서의 수로 계산된다. 정확률과 재현률의 측정의 기준이 되는 정답문서의 추출은 수작업으로 수행되었고, 질의 처리 방식은 4장에서 제시한 방법으로 결과 문서들을 추출하고, 다음으로 5장의 식 (3)에 대한 결과를 가지고 문서랭킹 값을 매기고 이를 다시 결과에 반영하여 측정하였다. 문서랭킹이 일정 비율인 50% 이하인 경우에 결과에서 제거하는 방식을 사용하였다. 문서랭킹의 값이 일정 비율 이하인 문서를 제거 할 때 이 값의 변형에 따라 다른 결과를 나타낼 수 있다. 아래 측정된 그래프는 문서랭킹값의 측정 때 문서중요도와 키워드중요도 중에서 키워드중여도의 α값의 변형에 따른 결과의 변화에 대하여 중점적으로 측정하였다.

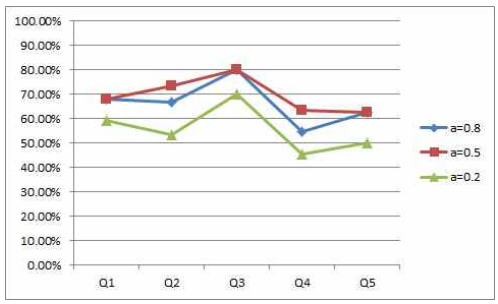
(그림 8) 제안된 처리 기법의 정확률



(Figure 8) Precision of the proposed method

(그림 8)는 제안된 검색 기법으로 질의를 수행 한 후에 관련이 있다고 검색된 문서의 순위를 고려하여 정답문서집합과 비교하여 정확률을 측정한 결과이다. 5장에서 제시된 식 (2)의 α값을 달리 하면서 결과를 측정하였다. 제시된 결과를 보면 α=0.8과 α=0.5 인 경우에 비슷한 결과를 보이고 있으며, α=0.2인 경우에는 다소 낮은 결과를 보이고 있다. 이는 검색결과의 정확도에 미치는 영향이 키워드의 클래스 레벨에 의한 평가보다는 출현 문서수에 의하여 더욱 정확히 판단될 수 있음을 나타낸다. α=0.8과 α=0.5가 경우에 따라 서로 다른 결과를 보이는 것은 더 많은 샘플과 질의를 처리하여야 정확한 이유를 판단 할 수 있을 것으로 보인다.

(그림 9) 제안된 처리 기법의 재현률



(Figure 9) Recall of the proposed method

(그림 9)는 제안된 검색 기법으로 질의를 수행 한 후에 관련이 있다고 검색된 문서의 순위를 고려하여 정답문서집합과 비교하여 재현률을

측정한 결과이다. 정확률의 측정과 마찬가지로 5장에서 제시된 식 (2)의 α값을 달리 하면서 결과를 측정하였다. 제시된 결과를 보면 α=0.8인 경우에 α=0.5 인 경우보다 다소 우수한 결과를 보이고 있다. 하지만 월등히 높은 수치가 아니므로 두가지 경우의 우위를 판단하기는 힘들다. α=0.2인 경우에는 정확률의 경우와 마찬가지로 다소 낮은 결과를 보이고 있다. 이는 키워드의 빈도가 클래스의 레벨에 비하여 더 중요한 요소로 작용하고 있음을 보여주는 결과이다.

두 가지 그래프를 전체적으로 판단해보면 평균적으로 정확률이 77.6%이며, 재현률 63.8%로 나타나 정확률이 더 좋은 결과를 보였다. 이는 검색된 결과에 비하여 선정된 정답문서의 개수가 많았음을 나타낸다. 정답문서의 선정은 다소 주관적인 판단이 개입되므로 정답문서의 선정을 어떻게 하느냐에 따라 결과가 달라질 수 있다. 그러나, 전체적으로 제시된 방법의 질의 처리와 우선순위의 부여가 웹상의 관련 문서를 찾는데 유용한 방법임을 알 수 있다.

7. 결론

갈수록 정보 검색의 중요성이 증대되고 있는 가운데 온톨로지와 같은 새로운 개념의 메카니즘을 이용하여 좀 더 정확한 검색 결과를 얻기 위한 시도들이 있다. 본 연구에서는 온톨로지를 이용하여 구축된 시맨틱 웹 환경에서 시맨틱 검색을 수행하기 위한 처리 과정과 메카니즘을 제안하였다.

RDF와 같은 시맨틱웹 언어에 대하여 살펴보았고, 이를 기반으로 구성한 온톨로지의 개념과 예를 살펴보았다. 온톨로지와 KB를 통하여 구축된 시맨틱웹과 시맨틱 주석을 통하여 연결된 웹문서에 대한 질의 처리 과정을 제시하였다. 온톨로지와 KB는 스키마와 사실 관계로 연결하였으며, KB와 웹문서는 시맨틱 주석의 개념을 이용하여 연결하였다. 처음 질의는 온톨로지와 KB를 대상으로 수행하여 이를 바탕으로 웹상의 문서를 검색하여 새로운 사실들을 유추하여 제공하도록 하였다. 검색된 문서들을 대상으로 랭킹을 구현하기 위한 식을 제시하였으며, 이를 정확률과 재현률 요소를 가지고 실제로 수치적으로 질의 처리된 결과를 제시하였다. 랭킹을 올바로 구현하는 요소로 키워드의 빈도와 온톨로지상의 클래스 레벨이 영향을 미치는 것을 확인 할 수 있었다. 제안된 방법은 추후 시맨틱웹의 검색 연구에 요소기술로 사용가능하며, 추후로 사실에 대한 순위를 고려한 실험을 진행하여 종합적으로 판단하여 좀 더 정확한 결과를 산출하도록 할 것이다.

References

[1] Berners-Lee. Tim, Heldler. J, Lassila. O, "The Semantic Web," Scientific American, pp.28-37, May 2001,

[2] Resource Description Framework(RDF) Model and Syntax Specification, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[3] Resource Description Framework(RDF) Schema Specification 1.0, http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

[4] Gruber, T.R., "A Translation Approach to Portable Ontology Specifications", Knowl. Acquis., Vol. 5, No. 2., pp.199-220, 1993,

[5] Kim, Youn Hee, Lee, AeJung, "OWL Storage Model to Support Efficient Ontology Reasoning Query," Journal of the Korea Society of Digital Industry and Information Management, Vol. 7, No. 3, pp. 25-35, 2011

[6] J. Heflin, J. A. Hendler, and S. Luke. SHOE: A blueprint for the Semantic Web. In D. Fensel, W. Wahlster, and H. Lieberman, editors, Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, .MIT Press, pp. 29 - 63, 2003.

[7] T. W. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng. Swoogle: Searching for knowledge on the Semantic Web. In Proc. AAAI-2005, AAAI Press / MIT Press, pp. 1682 - 1683, 2005.

[8] O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker. Querying the SemanticWeb with Corese search engine. In Proc. ECAI-2004, pp. 705 - 709. 2004.

[9] E. Thomas, J. Z. Pan, and D. H. Sleeman. ONTOSEARCH2: Searching ontologies semantically. In Proc. OWLED-2007, CEUR Workshop Proceedings 258. CEUR-WS.org, 2007.

[10] W3C. SPARQL Query Language for RDF, 2008. W3C Recommendation (15 January 2008). Available at http://www.w3.org/TR/rdf-sparql-query

[11] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. In Proc. ICDE-2008, pp. 953 - 962. 2008.

오 성 균

1981년 : 홍익대학교 이공대학 전자계산학과 이학사
1984년 : 연세대학교 산업대학원 전자계산학과 공학석사
1999년 : 홍익대학교 공과대학 전자계산학과 이학박사
1987년～현재 : 서일대학교 컴퓨터소프트웨어과 교수
관심분야 : 능동데이터베이스, XML모델링, 소프트웨어공학

김 병 곤

1990년 : 홍익대학교 공과대학 전자계산학과 이학사
1992년 : 홍익대학교 공과대학 전자계산학과 이학석사
2001년 : 홍익대학교 공과대학 전자계산학과 이학박사
1992년～1998년 : 국방과학연구소 연구원
2001년～현재 : 부천대학교 e-비즈니스과 부교수
관심분야 : 다차원 인덱싱, 시맨틱 웹, 온톨로지