자동화된 통합 프레임워크를 위한 시맨틱 웹 기반의 정보 검색 시스템

최 옥 경[†] · 한 상 용[††]

요 약

정보 검색 시스템은 사용자가 찾고자 하는 지식 정보를 보다 정확하고 빠르게 전달하는 데 그 목적이 있다. 그러나 현재의 검색 시스템은 단순 구문 분석 방식으로 사용자가 원하는 정확한 정보를 제공하지 못한다. 따라서 본 논문에서는 온톨로지 서버를 이용한 SW-IRS(Semantic Web based Information Retrieval System)를 제안한다. 제안한 시스템은 에이전트 기반의 자동 분류 기술과 시맨틱 웹 기반의 정보 검색 기법들을 이용하여 반구조(semi-structured) 문서뿐만 아니라 비구조(unstructured) 문서의 처리를 극대화 시키고자 한다. 또한 상호 운용성 및 데이터 통합을 위해 RDF(Resource Description Framework) 방식의 문서 저장 서버를 지원하며 웹 페이지들간에 검색 순위를 두어 보다 신속하고 정확한 정보 검색이 가능하도록 하고자 한다. 마지막으로 새로운 순위 측정 알고리즘을 제안하고 이를 이용한 성능 평가를 실시하여 그 효율성과 정확성을 검증해 보이고자 한다.

키워드 : 시맨틱 웹, 온톨로지, 정보검색, RDF(Resource Description Framework), RQL(RDF Query Ranguage)

Semantic Web based Information Retrieval System for the automatic integration framework

Okkyung Choi[†] · Sangyong Han[††]

ABSTRACT

Information Retrieval System aims towards providing fast and accurate information to users. However, current search systems are based on plain syntactic analysis which makes it difficult for the user to find the exact required information. This paper proposes the SW-IRS (Semantic Web-based Information Retrieval System) using an Ontology Server. The proposed system is purposed to maximize efficiency and accuracy of information retrieval of unstructured and semi-structured documents by using an agent-based automatic classification technology and semantic web based information retrieval methods. For interoperability and easy integration, RDF based repository system is supported, and the newly developed ranking algorithm was applied to rank search results and provide more accurate and reliable information. Finally, a new ranking algorithm is suggested to be used to evaluate performance and verify the efficiency and accuracy of the proposed retrieval system.

Key Words : Semantic Web, Ontology, Information Retrieval, RDF(Resource Description Framework), RQL(RDF Query Ranguage)

1. 서 론

정보 검색 시스템은 사용자가 원하는 지식 정보를 얼마나 정확하고 빠르게 검색하여 의미 있는 지식 정보를 제공할 수 있는가에 따라 시스템의 성능과 평가가 좌우된다고 할 수 있다[17]. 그러나 현재의 검색 시스템은 단순한 구문 처리 방식으로 데이터들간의 관련성만을 검사하므로 의미론적인 해석이 전혀 이루어지지 않고 있다. 시맨틱 웹(Semantic Web)은 이와 관련된 문제점들을 해결한 차세대 웹 기술로 자원간의 관계를 연결시키는 온톨로지 기술과 자동화 에이전트 기술을 접목시켜 의미론적 검색, 상호 운용성 보장, 어플리케이션간의 통합 등의 효과를 가져 올 수 있다. 따라서 이러한 에이전트들이 구조화(structured) 또는 반구조화(semistructured) 문서로부터 온톨로지에 기반한 메타 정보(metadata)를 추출하기 위해 다양한 시맨틱 기술과 결합된 규칙을 사용하는 것이 필요하다.

본 연구에서는 시맨틱 웹 기반 기술을 하위 구조로 설계

※ 이 논문은 IITA(Institute of Information Technology Assessment)에서 후원하는 홈 네트워크 연구 센터(HNRC-ITRC (Home Network Research Center)) 산하 중앙대학교 MIC(Ministry of Information and Communication)의 연구비 지원에 의한 것임.
† 준 회 원 : 중앙대학교 대학원 컴퓨터공학과 박사과정
†† 종신회원 : 중앙대학교 컴퓨터공학과 교수
논문접수 : 2005년 9월 7일, 심사완료 : 2005년 12월 5일

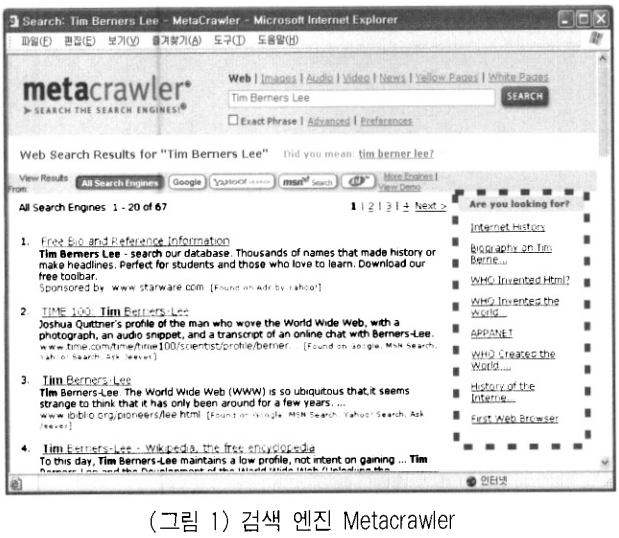
한 후 여기에 에이전트 기반의 자동 분류 기술과 시맨틱 웹 기반의 정보 검색 기법을 접목시킨 SW-IRS(Semantic Web based Information Retrieval System)를 제안하여 현재 검색 시스템이 가지고 있는 정확성과 효율성이 떨어지는 문제점을 해결하고자 한다. 제안 시스템은 3가지 관점에서 기존 연구와의 차별화를 보이고자 한다. 첫째, 기존 검색 시스템의 단순 구문 처리 방식을 보완하기 위해RDF(Resource Description Language)[2]와 DAML+OIL(DARPA Agent Markup Language + OIL)[1] 기반의 온톨로지 서버를 설계하고 이를 바탕으로 의미론적 데이터의 해석 및 분석이 가능하도록 하였다. 둘째, 현재 시맨틱 웹 기반 검색 시스템의 수행 능력을 평가할 만한 성능 측정 알고리즘이 미비한데 본 연구에서는 새로운 순위 측정 알고리즘(3장)을 제시하여 이러한 문제점을 해결하였다. 셋째, 제안한 순위 측정 알고리즘과 자동 분류 기술을 적용하여 성능 평가를 실시함으로써 사용자에게 보다 정확하고 효율적인 정보 전달이 가능하도록 하였다.

본 연구의 구성은 먼저 2장에서 기존 시스템을 분석하고, 3장에서는 순위 측정 알고리즘을 기술하고, 4장에서는 본 연구에서 제안하는 통합 정보 검색 시스템인 SW-IRS의 설계 기법, 구조, 모듈별 기능 및 특징에 대해 논하고, 5장에서는 3장의 성능 측정 알고리즘을 바탕으로 성능 평가를 실시하며, 마지막으로 결론 및 향후 연구 과제를 6장에서 언급하였다.

2. 기존 시스템 분석

본 장에서는 효율적인 정보 검색을 위한 에이전트 기반의 정보 검색 시스템을 비교 및 분석 한다.

2.1 멀티 쓰레드 기반의 검색 엔진 Metacrawler



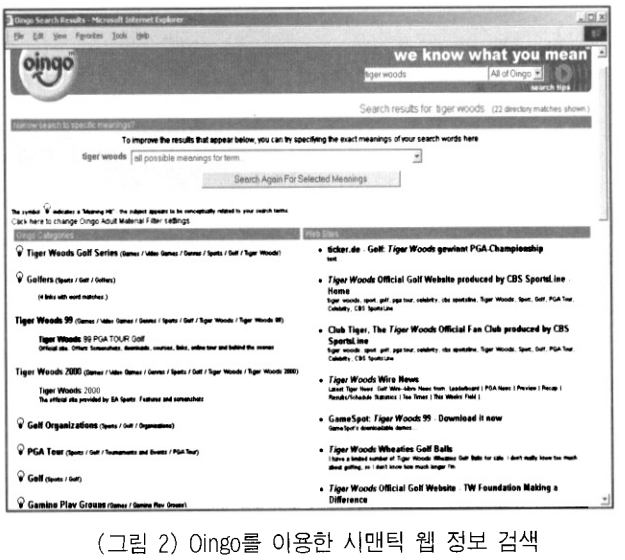
(그림 1) 검색 엔진 Metacrawler

자체적으로 검색 시스템을 운영하지 않고 Google[15], Yahoo[3], MSN[4], LookSmart[5] 등과 같은 검색엔진에 검색을 의뢰한 후 반환 받은 결과물을 웹 상에 보여주는 방식을 채택하고 있는 검색 기법인 Metacrawler[6]는 (그림 1)에서 보는 바와 같이 "Tim Berners Lee"라는 검색어를 입력하였을 때 해당 결과 사이트를 좌측에 보여주고 우측에는 에이전트 기법을 이용하여 사용자가 관심 있어 하는 항목을 찾을 수 있도록 리스트를 보여준다. Metacrawler는 에이전트를 이용하여 관심있어 할 만한 항목들을 리스트로 제시하는 방식으로 의미론적 검색이 가능한 것 같아 보이지만 실제로 제시해 주는 관심사 리스트는 사용자 의도와는 전혀 무관한 정보를 제공해 주는 경우가 있다.

2.2 Google

Google[15]은 1999년 등장한 검색 엔진으로서 여러 개 나라의 언어를 지원한다. 기존 검색엔진들은 키워드 방식의 순위 결정 방법을 수행하지만 Google에선 각 웹 페이지마다 참조한 페이지 수에 의거하여 검색 순위를 부여하기에 다른 검색엔진에 비해 좀 더 정확하고 많은 검색 결과를 가져오는 장점이 있다. 다른 사이트와의 차별화된 검색 순위 결정 방법으로 학내나 연구소에서 가장 많이 이용하는 검색엔진으로 알려져 있지만 존재하지 않는 사이트(broken link)가 검색되거나 전혀 사용자가 의도하는 바와는 무관한 정보가 검색되는 경우도 많다.

2003년 구글은 Oingo사의 온라인 광고 시맨틱 기법을 도입시켜 광고 에이전트가 일반 사용자에게 내용 기반의 광고를 제공시켜 주는 부분적인 의미론적 검색 방식을 추가하였다. 아래 (그림 2)는 Tiger Woods 라는 검색어를 입력하고 이와 관련된 Clustering 된 Category 와 Web Site를 조회하는 화면으로 "Tiger Woods"라는 검색어에 대해서 에이전트가 사용자가 관심 있어할 만한 내용들을 추가하여 제시해 주는 부분이다.



(그림 2) Oingo를 이용한 시맨틱 웹 정보 검색

끝으로 Score (Semantic Content Organization and Retrieval Engine)[8]는 조지아 대학(Univ. of Georgia)의 대

형 분산 시스템 연구소(Large Scale Distributed Information Systems Laboratory)에서 정보 검색 시스템에 시맨틱 웹 기술을 접목시킨 차세대 기술로 소프트웨어 에이전트(agent)들이 유지할 수 있는 온톨로지(Ontology)의 구성 요소를 지정하는 기능을 제공한다.

에이전트들은 구조화(structured) 되거나 반구조화(semi-structured)된 내용들(contents)로부터 온톨로지에 기반한 메타 정보(metadata)를 추출하기 위해 다양한 시맨틱 기술과 결합된 규칙(rule)을 사용한다. 또한 자동 분류와 정보 추출 기술을 이용하여 비구조화된(unstructured) 문서의 처리가 가능하다.

3. 순위 측정 알고리즘(Ranking Algorithm)

정확한 정보의 검색을 위해선 사용자가 원하는 정보의 순위화가 반드시 요구된다. 3장에서는 이러한 순위화를 위한 개선된 순위 측정 알고리즘을 제안한다

3.1 순위 측정 알고리즘(Ranking Algorithm)

기존 논문[19]에서 제시한 연구 방법의 유사도 측정 기법을 이용하여 비교 및 분석을 한 결과, 단순히 단어의 빈도수만을 반영한 기존 검색 방식의 문제점을 해결하여 사용자가 의도하는 결과의 의미론적 검색이 가능하게 된다. 그러나 시맨틱 정보(RDF 문서)의 유무만을 판단하여 최종 유사도 및 순위를 부여하기에 실제적으로 의미 있는 정보가 검색될 가능성이 있는 반구조적 문서나 비구조적 문서들을 제외시킴으로써 순위의 정확성을 떨어뜨리는 역효과를 가져오게 된다. 이에 기존 벡터 모델을 새롭게 보안한 순위 측정 알고리즘을 제안하여 이러한 문제점을 해결하고자 한다.

기존 방법[19]에서는 HTML, XML 과 같은 문서 형식과 RDF 형식이 포함된 문서들을 구분하여 성능 평가를 실시하였다. 그러나 본 논문에서 제안하는 방법은 이러한 문서들의 구분을 두지 않고 개선된 순위 측정 알고리즘을 이용하여 문서에 대한 정확한 순위화가 가능하도록 하였다. 수식 (1)은 의미론적 메타 정보를 사용한 RDF 문서와 XML, HTML과 같은 문서들의 자동 분류 및 순위를 부여하기 위한 순위 측정 알고리즘으로 기존 벡터모델의 코사인 유사도와 3.2절의 정의 1)에서 제안한 가중치 부여 비례 반영치(k_j)로 이루어진다.

$$sim(d_j q) = k_j \times \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} k_j = \frac{R_j}{D_j}. \quad (1)$$

3.2 가중치 부여 비례 반영치(kj)

이번 단원에서는 3.1의 순위 측정 알고리즘(ranking algorithm)에 적용시킬 가중치 부여 비례 반영치를 기술한다. 자동 분류 및 순위화를 위한 가중치 부여 비례 반영치(k_j)는 다음과 같다.

[정의 1] 가중치 부여 비례 반영치(k_j)

$$k_j = \frac{R_j}{D_j}.$$

R_j : 각 용어들(j)간의 동의어 관계를 측정한 term relationship 변수
D_j : 용어간에 관계성(relationship)을 측정한 Semantic Distance 변수

정의 2, 3은 정의 1)의 요소인 각 용어들(j)간의 동의어 관계를 측정한 term relationship 변수, 용어간에 관계성(relationship), 즉 거리에 따른 근접도를 측정한 Semantic Distance 변수를 이용한다.

term relationship 변수는 각 용어가 가지는 similarity level(유사도 범위)를 이용하여 측정하는 데 범위는 1-9 사이의 값을 가지며 유사성이 높을수록 1에 가깝고 유사성이 떨어질수록 9에 가까워 진다. 각 범위(level)의 비교 대상 요소는 검색어, 추출된 문서에 포함된 용어, 온톨로지를 통해 추출된 검색어의 동의어, 온톨로지를 통해 추출된 문서에 포함된 용어의 동의어다.

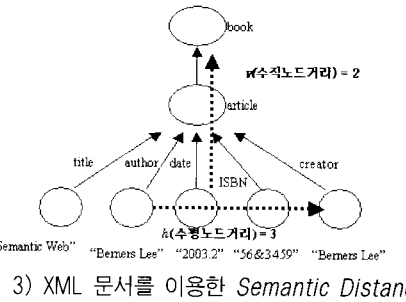
term relationship 의 정의는 다음과 같다.

[정의 2] *term relationship* (용어들간 동의어 관계)

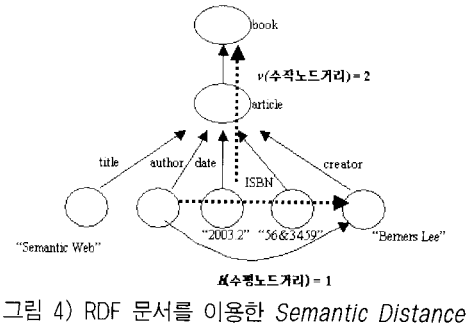
$$R_j = \frac{f_{ij}}{t_r}.$$

f_{ij} : 문서(i)에서 용어(j)의 발생수
t_r : 용어(j)들간의 유사도 측정 변수 = level(tr)

Semantic Distance 변수는 각 문서가 가지는 구조들의 각 수평 노드간의 근접도(H_p)와 각 수직 노드간의 근접도(V_p)를 이용하여 가중치를 결정한다. 다음 (그림 3,4)는 XML 문서와 RDF 문서를 분리해서 ***Semantic Distance*** 값을 측정한 것을 비교 분석 한 예이다.



(그림 3) XML 문서를 이용한 *Semantic Distance*



(그림 4) RDF 문서를 이용한 *Semantic Distance*

여기서 XML 문서와 RDF 문서간의 *Semantic Distance* 변수 값이 다르게 나타나는 이유는 XML 문서는 트리 구조의 계층적 방식이고 RDF 문서는 그래프 구조 방식으로, "Author와 Publisher가 모두 Berners Lee인 book"을 찾는다고 했을 때 RDF 문서에선 author와 creator사이의 수평 노드간 거리가 "1"로 매우 밀접한 관련이 있지만 XML 문서에선 수평 노드간 거리가 "3"으로 관련성이 떨어지게 된다.

*Semantic Distance*의 정의는 다음과 같다.

[정의 3] *Semantic Distance* (용어간에 관계성, 즉 거리에 따른 근접도를 측정)

$$D_j = H_p * V_p$$

$$H_p = \frac{1}{C^h}, h = |k - j|, C = \frac{level(i_j)}{\max V(i)}$$

H_p : 각 노드간의 수평 근접도,
^{h} : 각 용어간의 수평 근접도
C : 문서내의 각 트리의 level 측정 변수
$level(i_j)$: 문서(i)에서 용어(j)가 위치한 곳의 level 값
$\max V(i)$: 문서((i)에서 최대 level 값

$$V_p = \frac{1}{F^v}$$

V_p : 각 노드간의 수직 근접도,
F : 수직 근접도 결정 인자(0<F<1)
$v = level(i_k) - level(i_j)$: 각 용어간의 수직 노드 거리

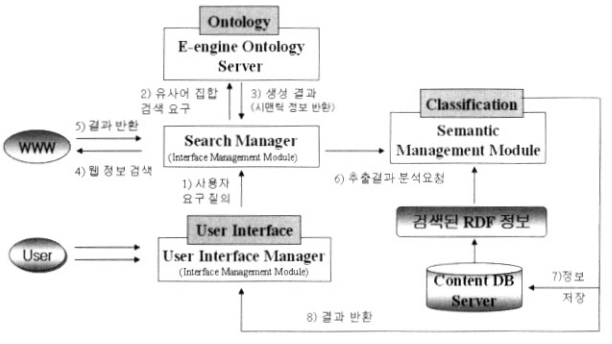
4. SW-IRS(Semantic Web based Information Retrieval System)

4.1 필요성 및 연구배경

현재 웹 기술의 발전동향을 살펴 보면, XML의 등장으로 사용자가 임의로 태그를 생성할 수 있는 기반이 구축되었으나, XML은 동일한 의미의 메타 데이터를 서로 다르게 작성하여 동일 문서가 상이한 문서로 분류되는 문제점을 야기시켰다. 또한, XML은 구조적 메타 데이터에만 치중한 나머지 내용에 의미가 결합되지 않고 의미론적 연결이 전혀 이루어지지 않았다. 이에 의미 있는 웹으로의 전환의 필요성이 점점 대두되기 시작하였다. 과거의 웹에 대한 기술 개발 및 표준 개발과 더불어, 데이터 또는 문서들이 가지고 있는 정보 자원들에 대하여 상호 운용이 가능하고 의미론적 통합이 가능한 형태로 개발된다면, 같은 도메인 내에서 사용하는 메타 데이터가 일정한 규칙을 적용하여 새로운 정보를 도출해 낼 수 있을 것이다. 따라서 이러한 정보는 지금의 검색 엔진에서 찾아내는 무의미하고, 부정확한 검색 결과가 아니라 사실에 기반한 지식이 될 수 있을 것이다. 이에 본 연구에서 제안하는 의미론적 데이터 검색을 위한 온톨로지 기반의 시맨틱 웹 정보 검색 시스템을 제안하는 것이다.

4.2 방법론 및 절차

(그림 5)는 시스템의 전체 흐름도로 사용자가 검색 질의를 하고 최종 검색 결과를 제공해주기까지의 과정을 나타낸 것이다. 8단계의 절차를 통해 시맨틱 웹 검색을 수행하며 각각의 절차를 살펴 보면 다음과 같다.



(그림 5) 시스템 전체 흐름도

1단계 : 사용자 요구 질의: 사용자가 용어들의 집합과 연산자들을 이용해서 질의를 한다.

2단계 : 관련 온톨로지 용어 선택: Interface Management Module은 E-engine Ontology Server를 이용하여 복수개의 온톨로지들을 제공한다. 사용자는 제공된 온톨로지들 중에서 자신이 원하는 정보를 가진 온톨로지를 선택한다. E-engine Ontology Server는 사용자가 올바르게 온톨로지를 선택할 수 있도록 온톨로지 용어와 설명을 함께 제공해 준다.

3단계 : 유사어 생성: 선택한 온톨로지 용어를 바탕으로 사용자의 요구 질의와 관련이 있는 유사어 집합을 E-engine Ontology Server에서 가져온다. 예를 들어 음식, 식료품, 요리등은 식품과 관련이 있고, 주문자 부착 상표 방식, original equipment manufacturer 등은 OEM과 관련이 있다.

4, 5단계 : 외부 검색 엔진을 통한 질의 형성: 외부 검색 엔진을 이용하여 사용자가 요구하는 질의를 주고 검색 결과를 얻어 온다. 여기서 검색이 효율적으로 이루어질 수 있도록 적합한 질의를 형성하여 정확한 검색 결과를 유도한다.

6단계 : 관련 페이지 결정: 외부 검색 엔진을 통해 관련 URI 정보를 얻게 된다. 여기서 URI 정보는 일부 문맥에 질의어가 포함되어 나온 결과물로 사용자가 의도하는 바와는 전혀 상관 없는 URI도 얻어지게 된다. 따라서 각각의 URI가 가지고 있는 부분적으로 인용된 문장들을 임계점(threshold) 방식의 클러스터링 기법을 이용해서 유사도를 측정하고 0.05 이하의 임계점을 가진 URI는 제외시킨다.

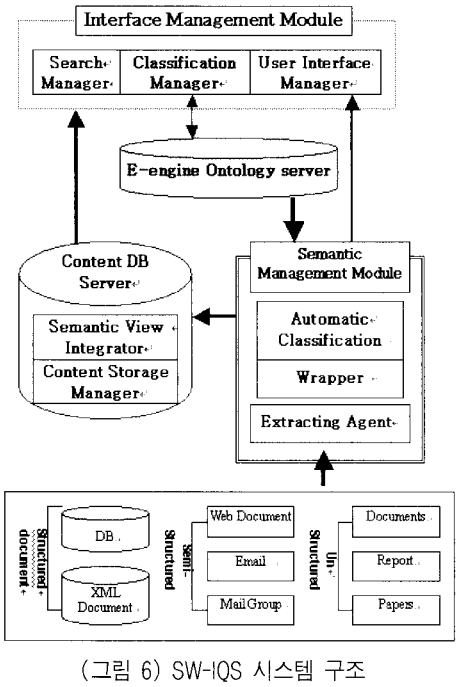
7단계 : 자동 분류 및 저장: 클러스터링를 통해 얻어진 URI 검색 결과를 Semantic Management Module를 통해 의미론적, 구조적 메타 정보를 자동 분류하고 분류된 정보를 Content DB Server에 저장한다. Content DB Server는 이러한 메타 정보를 RDF와 DAML+OIL의 형식으로 저장한다.

8단계 : 페이지 순서화 및 최종 결과 반환: 유사도 측정 모델

을 이용하여 URI 검색 결과와 Content DB Server에서 가져온 RDF 문서의 순위를 부여한다. 여기서 사용자는 10개의 웹 페이지를 검색하기를 원했기 때문에 1위-10위까지의 웹 문서 결과를 제공해 준다.

4.3 시스템 구조 및 모듈별 기능

전체적인 시스템 구조는 (그림 6)에서 보는 바와 같이 Interface Management Module, E-engine Ontology Server, Content DB Server, Semantic Management Module로 구성된다. 본 단원에서는 각 모듈별 상세 기능 및 특징을 살펴보고자 한다.



(그림 6) SW-IQS 시스템 구조

4.3.1 E-engine Ontology Server

E-engine Ontology Server는 World Map이라고도 불리며syntactic layer(XML), semantic layer(RDF)의 요소를 포함하여 웹상의 정보를 단순한 데이터 차원에서 처리하여 사람이 의미를 부여하는 현재의 상태에서, 정보 생성 단계까지 정보가 지식으로서의 가치를 지닌 상태로 향상시킬 수 있는 지식의 체계적 표현 방안이다.

E-engine Ontology Server는 Content Manger, Schema Manager, Thesaurus Manager의 3개의 층(layer)으로 구성된다. Content Manager는 시맨틱 메타 데이터에 대한 정의, 의미론적 데이터 검색을 위한 분류 모델 정의, 상속, 대등 등을 이용하여 메타 데이터 간에 관계를 정의한다. Thesaurus Manger는 전자상거래 국제 표준에 따라 식별, 속성 표준을 정의한 일종의 백과사전으로 스키마 통합이나 유사 용어들에 대한 통일 및 재구성의 역할을 한다. Schema Manager는 Content Manger의 표준 분류 모델과 Thesaurus Manager의 의미론적 통합 모델에 대한 표준 데이터 타입과 형식이 정의되어 있다.

4.3.2 Interface Management Module

Interface Management Module은 검색의 정확성을 높이기 위해 Ontology Server의 도메인 시맨틱 정보를 가져와 사용자에게 재질의를 통한 정확한 검색 결과를 유도하며 또한 Search Manager, Classification Manager, User Interface Manager를 통해 검색의 효율성을 증진시키고자 한다. 본 단원에서는 SW-IQS의 최상위 층인 Interface Management Module의 각 모듈별 기능 및 특징에 대해 서술한다.

(1) Search Manager

Search Manager는 사용자가 입력한 검색어를 기준으로 온톨로지 서버로 부터 관련 도메인 시맨틱 정보를 가져와 사용자에게 재질의를 통한 정확한 검색 결과를 유도한다. 즉 여러 개의 도메인 시맨틱 정보가 검색되었다면 사용자에게 해당 시맨틱 정보에 대한 주제어와 설명을 제시하여 원하는 메타 데이터를 선택할 수 있도록 한다.

(2) Classification Manager

사용자가 원하는 정보를 찾는 검색방법으로는 특정 검색어나 주제어를 입력하여 관련 웹 페이지들을 찾는 검색어 입력 방식과 찾고자 하는 단어를 모르거나, 찾고자 하는 정보 등이 광범위 할 때 이용할 수 있는 주제별 검색 방식이 있다. Classification Manager의 주제별 검색 방식은 계층적 구조 방식의 기존 검색 기법과는 차별화하여 Content DB Server가 보유하고 있는 문서 정보를 바탕으로 유연한 구조의 네트워크 방식을 택한다. XML 문서는 계층적 구조의 분류학적 방식으로 제품 간의 상호 연관관계를 표시해 주기 힘들다. 따라서 RDF 문서를 바탕으로 한 용어간의 관계성을 구분해 주는 유연한 네트워크 구조 방식을 택하여 보다 정확하고 효율적인 문서 검색이 가능하도록 한다.

(3) User Interface Manger

다양한 사용자 검색 입력 화면과 온톨로지 정보 선택 화면을 제공하며 최종 정보 검색 결과 단계에서는 Semantic Management Module로부터 자동 분류 및 순위 화한 결과값을 반환 받은 후 최종 결과를 나타내 주는 화면을 사용자에게 제공한다.

4.3.3 Semantic Management Module

Semantic Management Module은 정보 추출 에이전트를 이용하여 관련 웹 페이지들을 추출하고 Wrapper를 이용하여 자료 중심의 XML 문서로 변환 시킨 후 Automatic Classification Module을 이용하여 페이지를 자동 분류하고 그 결과를 Content DB Server에 저장한다. 여기서 정보를 자동 분류하고 순위를 부여하기 위해선 관련 페이지들의 유사도를 측정하여야 하는데 이러한 유사도 측정을 위해 본 연구에서는 각 용어들(i)간의 동의어 관계와 용어간에 관계성을 이용한 순위 측정 알고리즘(3장)을 이용한다.

4.3.4 Content DB Server

Content DB Server는 2개의 모듈로 구성된다. 시맨틱 검색 엔진과 Storage Manager로 구성된다. 시맨틱 검색 엔진을 통해 나온 결과는 Semantic Management Module로 보내지며 여기서 순위 적용을 한 후 검색 결과가 사용자에게 보내지게 된다.

4.3.4.1 시맨틱 검색 엔진

(1) RQL Converter

RQL(RDF Query Language)를 사용하여 질의문을 생성한다. RQL은 RDF와 RDF 스키마를 위한 질의 언어로 RDF/RDF 스키마로 표현된 지식을 기반으로 에이전트 간에 질의를 던져서 사용자가 원하는 응답을 받아내는 방식이다. Content DB Server가 가지고 있는 RDF 문서를 사용자의 질의 정보를 이용하여 검색하기 위해 RQL 질의로 확장하여 검색한다. RQL을 이용하면 우선 만족하는 리소스들을 검색하고 사용자 질의에 기반한 일반적인 웹 검색을 실시한다.

(2) 엔진 분석기

엔진 분석기는 RQL Converter를 통해 나온 결과를 Content Storage Manager로 보낸 후 각 구성 요소를 분석 한다.

4.3.4.2 Content Storage Manager

E-engine Ontology Server의 온톨로지 정보를 바탕으로 Semantic Management Module 을 통해 분류된 웹 문서들을 RDF(S)와 DAML+OIL 형식으로 변환하여 각 페이지의 URI 정보와 함께 Content Storage Manager에 저장된다. 저장된 정보는 Semantic View Integrator에 있는 상태 View를 통해 저장 상태를 확인할 수 있으며 여기서 해당 정보의 수정이 가능하다.

5. 성능 평가

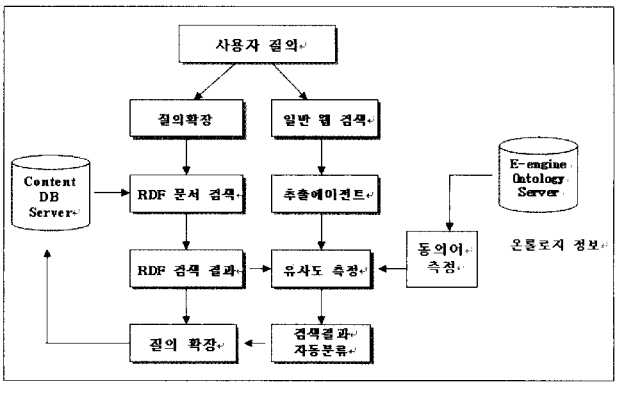
5장에서는 3장에서 제안한 새로운 순위 측정 알고리즘을 바탕으로 성능 평가를 실시한다.

〈표 1〉 테스트 대상 웹 페이지

번호	URL
0	http://ec.cse.cau.ac.kr/okchoi/rdf.xml
1	http://ec.cse.cau.ac.kr/okchoi/person_book.rdf
2	http://static.userland.com/userLandDiscussArchive/msg011396.html
3	http://www.amazon.co.uk/exec/obidos/ASIN/1587990180/202-4832593-9905459
4	http://www.w3.org/People/Berners-Lee/1996/ppf.html
5	http://lists.w3.org/Archives/Public/www-talk/1992MarApr/0030.html
6	http://www.aarp.org/computers_books/Articles/a2002-06-26-weaving
7	http://www.bookbrowse.com/dyn_/title/titleID/125.htm
8	http://www.businessweek.com/bwplus/books/bookauth1.htm
9	http://www.freelists.org/archives/hacknot/07_2002/msg00004.html
10	http://www.addall.com/Browse/Author/2688303-1
11	http://search.barnesandnoble.com/booksearch/

먼저 구글[15]에서 'the book which the author is berners Lee'이라는 검색어를 가지고 검색한 페이지 중 10위 안에 있는 10개의 문서와 2개의 XML, RDF 문서를 가지고 성능 분석을 하였다. 검색한 웹 페이지의 결과 문서들은 <표 1>과 같으며 각 번호는 문서 번호에 해당한다.

(그림 7)은 순위 측정 알고리즘을 이용한 단계별 검색 기법으로 성능 비교는 단계별로 수행되며 진행 순서를 살펴보면 다음과 같다.



(그림 7) 단계별 검색 기법

1단계 : 일반 검색엔진[Google]을 이용하여 상위 10개의 문서를 검색한다. 검색된 문서들을 추출 에이전트를 이용하여 불필요한 웹 페이지를 제거한다. 여기서 3, 9, 11번의 경우 잘못된 링크 정보로 웹 페이지 리스트에서 제거된다.

2단계 : Content DB Server가 가지고 있는 RDF 문서를 사용자의 질의 정보를 이용하여 검색하기 위해 RQL 질의로 확장하여 검색한다. RQL을 이용하면 우선 만족하는 리소스들을 검색하고 사용자 질의에 기반한 일반적인 웹 검색을 실시한다. 검색 결과 person_book.rdf 문서[7]가 검색되었다.

3단계 : 추출 에이전트를 통해 얻어진 9개의 일반 웹 문서와 RDF 문서를 새로운 유사도 측정 모델을 이용하여 실시한 후 그 결과를 가지고 순위를 매긴다.

4단계 : 용어간의 관계성을 측정하기 위해 Html 문서를 XML 문서로 변환한다. 이때 Html 문서를 XML 문서로 변환하기 위해 HtmltoXML Wrapper를 이용하며 여기서 나온 XML 문서들과 Content DB Server에서 가져온 RDF문서를 가지고 순위 측정 알고리즘을 이용한 유사도 측정을 한다. 그 결과 <표 2>와 같이 기존 코사인 유사도를 이용한 순위 결과와 다른 결과가 나온 것을 볼 수 있으며 이를 바탕으로 각각의 문서에 대한 순위를 재조정하고 문서에 대한 자동 분류가 이루어 진다. 현재의 검색 엔진은 문장에 포함된 단어의 가중치 뿐만 아니라 동의어에 대한 가중치 및 관련성을 전혀 고려하지 않고 있다. 또한 벡터 기반의 코사인 유사도를 이용한 경우 <표 2>에서 보는 바와 같이 RDF 문서가 4위로 일반 웹

〈표 2〉 유사도 측정 결과

문서 번호	용어 가중치				코사인유사도	순위		유사도측정 (제안모델)	순위 (제안모델)
	Book	Author	Berners Lee	Book Author Berners Lee					
0	0	0	0	0	0	8	0	0	8
1	0	0.00117450	0.00033550	0.001510110	0.00075505	4	0.002064926	0.001409910	1
3	0.00695923	0.00046600	0.00232970	0.009755110	0.00487755	0	0.001553185	0.003215370	0
5	0.00067344	0.00050500	0.00018930	0.001367778	0.00068388	5	0.000399741	0.000541815	6
6	0.00153377	0.00089855	0.00089850	0.003330889	0.00166544	1	0.000994426	0.001329935	3
7	0.00054144	0.00008000	0.00015220	0.000769778	0.00038488	6	0.001152593	0.001329935	5
8	0.00165611	0.00058988	0.00052777	0.002773778	0.00138688	3	0.000910704	0.000768741	1
9	0.00004300	0.00004000	0.00004000	0.000122000	0.00006100	7	0.000053000	0.000057000	7
11	0.00037555	0.00223922	0.00006890	0.003295778	0.00149486	2	0.001205704	0.001350284	2

K_j - 가중치 부여 비례 반영치

문서에 비해 순위가 낮게 나온 것을 볼 수 있다. 그 이유는 벡터 모델의 코사인 유사도를 이용할 경우 각 용어들(i)간의 동의어 관계를 측정한 ***term relationship*** 변수, 용어간에 관계성(relationship), 즉 거리에 따른 근접도를 측정한 ***Semantic Distance*** 변수가 검색 모델에 전혀 반영되지 않았기 때문이다. 이에 기존 벡터 모델을 개선한 새로운 순위 측정 알고리즘을 이용하여 유사도 측정을 한 결과 기존에 4순위 였던 RDF 문서가 1순위로 올라가고 기존 1순위 였던 6번 문서가 3순위를 기록한 결과를 볼 수 있어 보다 정확하고 신뢰성 있는 정보 검색이 가능하게 된다.

먼저 기존 벡터 모델의 유사도 측정 공식을 이용한 순위 측정을 위해 위에서 명시한 7개의 웹 문서와 2개의 XML ,RDF 문서를 바탕으로 각각의 벡터 기반 코사인 유사도 값을 계산하였다.

6. 결론 및 향후 연구

본 연구에서는 검색의 효율성과 정확성을 증진시키기 위해 SW-IRS(Semantic Web based Information Query System)을 제안하였다. 기존의 검색 모델이 가지고 있는 문제점을 해결하기 위한 방안으로, 차세대 웹으로 대두되고 있는 시맨틱 웹 요소들을 이용한 통합 정보 검색 시스템을 제시 하여 정보 추출 기법과 자동 분류 기법을 이용한 검색의 효율성과 정확성을 증진시키고 반구조(semistructured) 문서 뿐만 아니라 비구조(unstructured) 문서의 처리를 극대화 시키는 효과를 가져 오고자 한다. 제안한 시스템은 온톨로지의 확립, 데이터 표준화, 데이터 통합화, 시맨틱 연결 방법을 통해 의미론적 데이터 검색 및 통합이 가능하다.

제안한 통합 검색 시스템의 성능 측정을 하기 위한 방안으로 기존에 제안한 방법[19]을 보안한 새로운 성능 측정 알고리즘을 제시 하여 그 효율성과 정확성을 검증해 보였다. 기존 방법에서 제안한 RDF의 의미론적 메타 정보를 이용한 검색 기법을 개선하기 위해, RQL을 이용한 이진적 가중치를 부여하는 불리언 모델의 방식을 보완하여 비이진 가중치의 유사도 측정이 가능한 새로운 의미론적 벡터 모델을 제시하였다. 제안한 순위 측정 알고리즘을 이용하여 본 시스템을 분석한 결과 웹상에서 추출된 문서와 Content DB Server가 보유하고 있는 문서들의 순위 측정 결과가 기존 방법에 비해서 향상된 점을 볼 수 있다. 또한 HTML, XML, RDF 와 같은 문서 유형의 구분을 두지 않고 의미론적 통합 검색이 가능하게 된다.

또한 추출된 웹 페이지들의 관련성을 증진시키기 위해 문서의 구조, 동의어, 문맥어의 형태를 이용한 Semantic Management Module의 자동 분류 기법을 이용하였으며, E-engine Ontology Server는 검색의 정확률과 재현율을 높이기 위해 기존 계층적 구조 방식과는 달리 그래픽 방식의 유연한 구조 방식을 채택하여 유연성, 확장성, 상호 운용성을 증진시켰다.

향후 본 시스템의 Content DB Server를 보완하여 e-business를 위한 통합 전자상거래 프레임워크에 도입함으로써 의미론적 데이터 통합의 현실화가 가능하도록 하며 이를 이용한 정보서비스의 활성화를 촉진시키고자 한다. 또한 재현율을 적용시킨 성능 측정 알고리즘을 개발하여 이를 이용한 실험 및 분석이 시도되어야 할 것이다.

참 고 문 헌

[1] DAML+OIL, http://www.daml.org/2001/03/daml+oil-index.htm, March, 2001.

[2] RDF : Resource Description Framework by Josef Dietl & Ralph Swick, http://www.w3.org/Talks/1998/0417-WWW7-RDF/

[3] Yahoo, http://www.yahoo.com/

[4] MSN Search, http://search.msn.com/

[5] LookSmart, http://search.looksmart.com/

[6] Metacrawler, http://www.metacrawler.com/

[7] person_book.rdf, http://ec.cse.cau.ac.kr/okchoi/person_book.rdf

[8] Amit Sheth, Clemens Bertram, David Avant, Brian Hammond, Krysztof Kchut, Yashodhan Warke, "Managing Semantic Content for the Web", IEEE Internet Computing, Vol.6, No.4, pp.80~87 Jul., 2002.

[9] OQL, http://www.db.ucsd.edu/People/michalis/notes/02/OQL Tutorial.htm

[10] Lee Jae-ho and Yang Jeong-jin, "The Semantic Web: The Intelligent Technology of the Net Generation", TTA Journal, Serial No.81, Jun., 2002.
[11] Decker, S.; Mitra, P.; Melnik, S., "Framework for the semantic Web:an RDF tutorial", IEEE Internet Computing, Vol.4 Issue 6, pp.68~73, Nov.-Dec., 2000.
[12] Decker, S.; Melnik, S.; van Harmelen, F.; Fensel, D.; Klein, M.; Broekstra, J.; Erdmann, M.; Horrocks, I., "The Semantic Web:the roles of XML and RDF", IEEE Internet Computing, Vol.4 Issue 5, pp.63~73, Sep.,-Oct., 2000.
[13] The RDF Query Language (RQL), http://139.91.183.30:9090/RDF/RQL/.
[14] Kim Yeong-cheon et al, "Research on Enhancing Information Search by Reviewing Term Weight", Korea Fuzzy Logic and Intelligence Systems Society, Vol.11, No.9, pp.811~816, 2001.
[15] Google, http://www.google.com
[16] JTP: An Object-Oriented Modular Reasoning System, http://www.ksl.stanford.edu/software/JTP/
[17] 최호섭, 옥철영, "정보검색 시스템과 온톨로지", 정보과학회지, 제 22권 제 4호, pp.62~71, 2004.4.
[18] 김제민, 박영택, "사용자에 따라 검색 결과의 순위를 적용하는 DQL 검색 시스템", 정보과학회 2004년 춘계학술대회, Vol.31, No.1, pp.589~591, 2004.4.
[19] Okkyung Choi, Seokhyun Yoon, Myeongeun Oh, Sanygoung Han, "Semantic web Search Model for information retrieval of the semantic data", The Second HSI Conference, pp.588~593, Jun., 2003.

최 옥 경

e-mail : okchoi20@gmail.com
1996년 단국대학교 이과대학
1996년 삼성전자(주) 정보 통신 본부 근무
1999년~2000년 CJ드림소프트(주) F/E팀 근무
1999년 중앙대학교 대학원 컴퓨터소프트웨어학과(공학석사)
2000년~현재 중앙대학교 대학원 컴퓨터공학과 박사과정
관심분야 : Semantic Web Services System, e-Commerce and Auction System, Information Retrieval

한 상 용

e-mail : hansy@cau.ac.kr
1975년 서울대학교 공과대학(공학사)
1984년 Minnesota 공과대학(공학박사)
1984년~1995년 IBM 책임연구원
1995년~현재 중앙대학교 컴퓨터공학과 교수, 감정평가 학회 회장
관심분야 : Web Technologies, Web Services, Semantic Web, Information Retrieval and Multimedia