제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

SW-IQS : 의미론적 데이터 통합을 위한 시맨틱 웹 기반의 통합 정보 검색 시스템

최옥경*, 한상용*
*중앙대학교 컴퓨터공학과
e-mail : okchoi@archi.cse.cau.ac.kr

SW-IQS : Semantic Web based Information Query System for the integration of semantic data

Okkyung Choi*, Sangyoung Han*
*Dept. of Computer Engineering, Chung-Ang University

요 약

본 연구에서는 온톨로지를 이용한 SW-IQS(Semantic Web based Information Query System)를 제안한다. 제안한 시스템은 자동 분류 기술과 정보 검색 기법들을 이용하여 반구조(semi-structured) 문서 뿐만 아니라 비구조(unstructured) 문서의 처리를 극대화 시키고자 한다. 또한 상호 운용성 및 데이터 통합을 위해 RDF(S) 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.

1. 서론

시맨틱 웹은 현 웹의 확장으로 자원의 의미를 체계적으로 정의하여 기계와 사람의 협력적 운용을 유도하는 방법으로, 의미론적 검색, 자동화, 통합, 재 사용을 시도하는 새로운 접근이다. 이에 본 연구에서는 시맨틱 웹 기반 기술을 하위 구조로 설계한 후 여기에 에이전트와 자동 분류화 기술을 접목시킨 SW-IQS(Semantic Web based Information Query System)를 제안하여 현재 검색 시스템이 가지고 있는 정확률(precision)과 재현율(recall)이 떨어지는 문제점을 해결하고자 한다. 제안한 시스템은 RDF(Resource Description Language)와 온톨로지 기반의 E-engine Ontology Server 를 중심으로 하고, 의미론적 데이터의 해석 및 분석을 위한 *Semantic Management Module*, 검색의 효율성과 정확성을 높이기 위한 Interface Management Module, 온톨로지 정보를 바탕으로 메타 데이터를 RDF(S) 등의 유형으로 저장하는 Content DB Server 로 구성되어 기존 인간 중심의 웹 환경에서 기계 중심의 웹 환경으로의 전환을 추구하고자 한다. 또한 제안한 시스템의 효율성과 정확성을 입증하기 위하여 의미론적 데이터의 평가를 위한 새로운 순위 측정 알고리즘을 적용하여 성능 측정을 수행 한다.

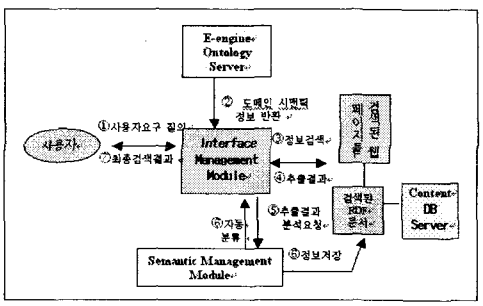
본 연구의 구성은 다음과 같다. 2 장에서는 본 연구에서 제안하는 SW-IQS 의 설계 기법, 구조, 모듈별 기능 및 특징에 대해 논하고, 3 장에서 기존 논문에서 제안한 성능 측정 기법을 보완한 새로운 순위 측정 알고리즘을 제안함과 동시에 이를 바탕으로 비교 및 분석, 평가를 실시하며, 마지막으로 결론 및 향후 연구 과제를 4 장에서 언급하였다.

2. SW-IQS 설계

본 장에서는 시스템의 전체 흐름도와 방법론을 제시하고 시스템의 전체 구조와 각 모듈별 기능 및 특징에 대해 서술한다

2.1 방법론 및 절차

본 절에서는 시스템의 전체 흐름도와 방법론을 제시한다. 제안한 방법론은 검색 결과의 재현율(recall)과 정확률(precision)을 높이기 위해 Interface Management Module 을 제공한다.



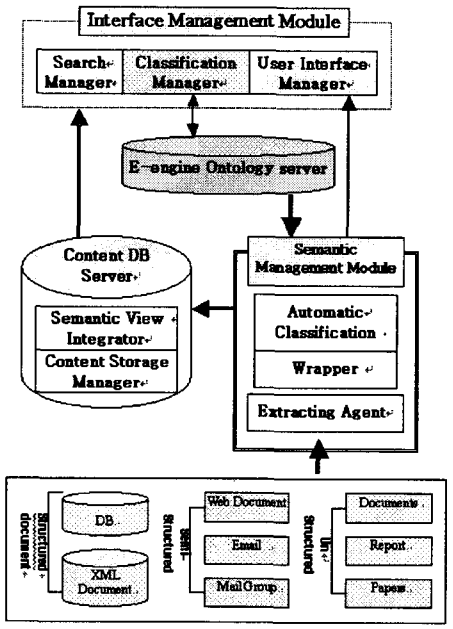
(그림 1) 시스템 전체 흐름도

(그림 1)은 시스템의 전체 흐름도로 사용자가 검색 질의를 하고 최종 검색 결과를 제공해주기까지의 과정을 나타낸 것이다. 최상위 어플리케이션인 Interface Management Module 은 사용자가 원하는 검색 정보를 입력하면 E-engine Ontology Server 에서 반환된 Domain Semantics 정보를 이용하여 웹에서 관련 페이지들을 검색한다. 이 때 Semantic Management Module 은 검색된 페이지들과 Content DB Server 의 RDF 문서 정보들을 가져와 자동 분류 및 순위를 제공한 후 결과를 Interface

Management Module 에게 넘긴 후 최종 결과를 사용자에게 제공한다.

2.2 시스템 구조 및 모듈별 기능

전체적인 시스템 구조는 (그림 2)에서 보는 바와 같이 Interface Management Module, E-engine Ontology Server, Content DB Server, Semantic Management Module 로 구성된다. 본 단원에서는 각 모듈별 상세 기능 및 특징을 살펴보고자 한다.



(그림 2) SW-IQS 시스템 구조

2.2.1 E-engine Ontology Server[3]

E-engine Ontology Server 는 World Map 이라고도 불리며 syntactic layer(XML), semantic layer(RDF)의 위에 존재하여 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.

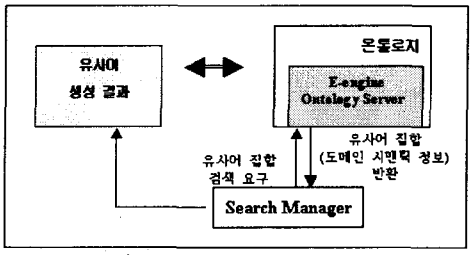
E-engine Ontology Server 는 Content Manger, Schema Manager, Thesaurus Manager 의 3 개의 층(layer)으로 구성된다. Content Manager 는 시맨틱 메타 데이터에 대한 정의, 의미론적 데이터 검색을 위한 분류 모델 정의, 상속, 대등 등을 이용하여 메타 데이터 간에 관계를 정의한다. Thesaurus Manger 는 전자상거래 국제 표준에 따라 식별, 속성 표준을 정의한 일종의 백과사전으로 스키마 통합이나 유사 용어들에 대한 통일 및 재구성의 역할을 한다. Schema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

2.2.2 Interface Management Module

Interface Management Module 은 검색의 정확성을 높이기 위해 Ontology Server 의 도메인 시맨틱 정보를 가져와 사용자에게 재질의를 통한 정확한 검색 결과를 유도하며 또한 Search Manager, Classification Manager, User Interface Manager 를 통해 검색의 효율성을 증진시키고자 한다. 본 단원에서는 SW-IQS 의 최상위 층(Layer)인 Interface Management Module 의 각 모듈별 기능 및 특징에 대해 서술한다.

(1) Search Manager

Search Manager 는 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. (그림 3)은 Search Manager 의 유사어 집합 생성 과정이다.



(그림 3) 유사어 집합 생성 과정

(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 Manager 의 주제별 검색 방식은 계층적 구조 방식의 기존 검색 기법과는 차별화하여(다르게) Content DB Server 가 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 Module 로부터 자동 분류 및 순위화한 결과값을 반환 받은 후 최종 결과를 나타내 주는 화면을 사용자에게 제공한다.

2.2.3 Semantic Management Module

Semantic Management Module 은 정보 추출 에이전트(Extracting Agent)를 이용하여 관련 웹 페이지들을 추출하고 Wrapper 를 이용하여 자료 중심의 XML 문서로 변환 시킨 후 Automatic Classification Module 을 이용하여 페이지를 자동 분류하고 그 결과를 Content DB Server 에 저장한다. 여기서 정보를 자동 분류하고 순위를 부여하기 위해선 관련 페이지들의 유사도를 측정하여야 하는데 이러한 유사도 측정을 위해 본 연구에서는 각 용어들(*i*)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 1 에 가깝고 유사성이 떨어질수록 9 에 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.

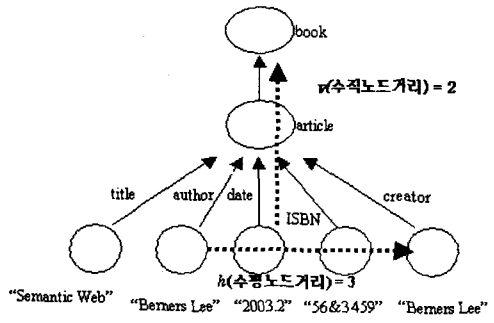
정의 1: ***term relationship*** (각 용어들간의 동의어 관계)

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

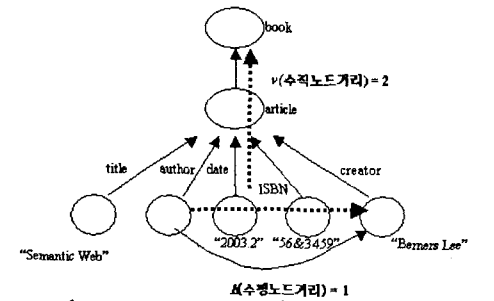
f_{ij}: 문서(i)에서 용어(j)의 발생수

t_r: 용어(j)들간의 유사도 측정 변수 $= level(t_r)$

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 4,5)는 XML 문서와 RDF 문서를 분리해서 ***Semantic Distance*** 값을 측정한 것을 비교 분석 한 예이다



(그림 4) XML 문서를 이용한 ***Semantic Distance***



(그림 5) 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, "Author 와 Publisher 가 모두 Berners Lee 인 book"을 찾는다고 했을 때 RDF 문서에선 author 와 creator 사이의 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 2: ***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

마지막으로 자동 분류 및 순위화를 위한 가중치 부여 비례 반영치(k_j)는 다음과 같다.

정의 3: 가중치 부여 비례 반영치(k_j)

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

R_j : 각 용어들(i)간의 동의어 관계를 측정한 ***term relationship*** 변수

D_j: 용어간에 관계성(relationship)을 측정한 ***Semantic Distance*** 변수

다음 장에서는 가중치 부여 비례 반영치(k_j) 를 이용하여 새로운 순위 측정 알고리즘을 제안한다.

3. 성능 평가

3 장에서는 기존 논문[3]에서 제안한 검색 모델의 문제점을 분석하고 이를 개선하기 위한해 새로운 순위 측정 알고리즘을 제시하여 이를 이용한 SW-IQS 의 성능 평가를 실시하고자 한다.

3.1 순위 측정 알고리즘

기존 방법[3]은 실제적으로 의미 있는 정보가 검색될 가능성이 있는 반구조적 문서나 비구조적 문서들을 제외시키고 시맨틱 정보(RDF 문서)의 유무만을 판단하여 최종 유사도 및 순위를 부여하기에 오히려 순위의 정확성을 떨어뜨리는 역효과를 가져 오게 된다. 이에 본 연구에서는 이러한 문제점을 해결하기 위해 문서들의 구분을 두지 않고, 가중치 부여 비례 반영치(k_j)와 벡터 모델의 코사인 유사도를 이용한 순위 측정 알고리즘(그림 6)을 통해 문서에 대한 우선 순위를 부여하고자 한다.

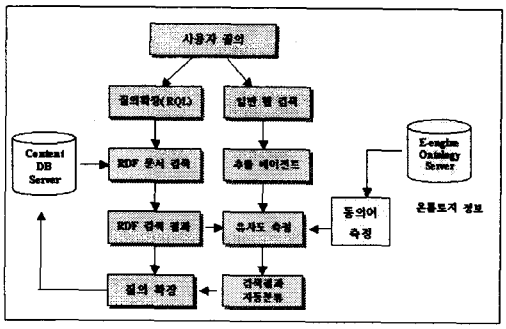
$$sim(d,q) = k_j \times \left(\frac{\vec{d} \cdot \vec{q}}{|\vec{d}| \times |\vec{q}|} \right)$$

(그림 6) 기존 벡터 모델을 개선한 순위 측정 알고리즘

3.2 비교 및 분석

본 절에서는 구글에서 'the book which the author is berners Lee' 이라는 검색어를 가지고 검색한 페이지 중

10 위 안에 있는 10 개의 문서와 2 개의 XML, RDF 문서를 가지고 성능 분석을 하였다. 검색한 웹페이지의 결과 문서들은 [4]와 같으며 각 번호는 문서 번호에 해당한다. 성능 비교는 새로운 단계별 검색 방법(그림 7)에 따라 진행되며 진행 순서를 상세히 살펴보면 다음과 같다.



(그림 7) 순위 측정 알고리즘을 적용한 단계별 검색 기법

1 단계 : 일반 검색엔진[Google]을 이용하여 상위 10 개의 문서를 검색한다. 검색된 문서들을 추출 에이전트를 이용하여 불필요한 웹 페이지를 추출한다. 여기서 문서 3, 9, 11 번의 경우 잘못된 링크 정보(broken Link)로 웹 페이지 리스트에서 제거된다.

2 단계 : RQL 을 이용하여 RDF 문서를 검색한다. 검색 결과 person_book.rdf 문서[5]가 검색되었다.

3 단계 : 추출 에이전트를 통해 얻어진 7 개의 웹 문서와 2 개의 XML ,RDF 문서를 바탕으로 각각의 벡터 기반 코사인 유사도 값을 계산한다.

4 단계 : 용어간의 관계성을 측정하기 위해 HtmltoXML Wrapper 를 이용하여 Html 문서를 XML 문서로 변환한다. XML 문서들과 RDF 문서를 새로운 순위 측정 알고리즘을 이용해 유사도 측정을 한다. 그 결과 <표 1>과 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.

<표 1> 유사도 측정 결과

문서 번호	용어 가중치				코사인 유사도	순위	k1	유사도 측정	순위
	book	author	berners Lee	book author berners Lee					
0	0	0	0	0	0	8	0	0	8
1	0	0.001174556	0.000335556	0.001510111	0.000755056	4	0.002064926	0.001409991	1
3	0.006958333	0.000468	0.002329778	0.00975611	0.004877555	0	0.001553185	0.00321537	0
5	0.000873444	0.000505	0.000188333	0.001367778	0.000683889	5	0.000399741	0.000541815	6
6	0.001533778	0.000898556	0.000898556	0.003336888	0.001665445	1	0.000994425	0.001329935	3
7	0.000541445	0.00008	0.000152222	0.000788778	0.000394889	5	0.001152593	0.000768741	5
8	0.001656111	0.000589889	0.000527778	0.002773778	0.001386889	3	0.000910704	0.001148796	4
9	0.000043	0.000040	0.000040	0.000122	0.000061	7	0.000053	0.000057	7
11	0.000375556	0.002239222	0.0000689	0.003285778	0.001494864	2	0.001205704	0.001350284	2

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 1>에서 보는 바와 같이 RDF 문서가 4 위로 일반 웹 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 알고리즘(그림 6)을 이용하여 유사도 측정을 한 결과 기존에 4 순위 였던 RDF 문서가 1 순위로 올라가고 기존 1 순위 였던 6 번 문서가 3 순위를 기록한 결과를 볼 수 있으며 보다 높은 정확률과 재현율로 사용자가 신뢰할 수 있는 결과를 가지게 된다.

4. 결론 및 향후 연구

본 연구에서는 검색의 효율성과 정확성을 증진시키기 위해 Semantic Management Module, Content DB Server, E-engine Ontology Server, Interface Management Module 로 구성된 SW-IQS(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.

또한 제안한 통합 검색 시스템의 성능 측정을 하기 위한 방안으로 기존에 제안한 방법을 보안한 새로운 유사도 측정 기법을 제시 하여 그 효율성과 정확성을 검증해 보였다. 기존 방법에서 제안한 RDF 의 의미론적 메타 정보를 이용한 검색 기법을 개선하기 위해, RQL 을 이용한 이진적 가중치를 부여하는 불리언 모델의 방식을 보완하여 비이진 가중치의 유사도 측정이 가능한 새로운 의미론적 벡터 모델을 제시하였다. 새로운 성능 측정 방법을 이용하여 본 시스템을 분석한 결과 웹상에서 추출된 문서와 Content DB Server 가 보유하고 있는 문서들의 유사도 측정과 순위부여 방법의 신뢰성과 정확성을 높여주는 결과를 가져 오게 되었다. 즉 RDF 검색 결과를 질의 확장에 이용하여 기존의 웹 상에서 보유하고 있는 HTML,XML 문서와의 구분을 두지 않고 통합 검색 엔진으로도 의미론적 검색이 가능하도록 하였다. 또한 추출된 웹 페이지들의 관련성을 증진시키기 위해 문서의 구조, 동의어, 문맥어의 형태를 이용한 Semantic Management Module 의 자동 분류 기법을 이용하였으며, E-engine Ontology Server 는 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 Server 를 보완하여 e-business 를 위한 통합 전자상거래 프레임워크에 도입함으로써 의미론적 데이터 통합의 현실화가 가능하도록 하며 이를 이용한 정보서비스의 활성화를 촉진시키고자 한다.

[참고문헌]

1. Gomez-Perez, A.; Corcho, O. "Ontology languages for the semantic web", IEEE Intelligent Systems , Volume: 17 Issue: 1 , Jan/Feb. 2002 p.54 -60

2. The RDF Query Language (RQL), http://139.91.183.30:9090/RDF/RQL/.

3. Okkyung Choi, Seokhyun Yoon, Myeongeun Oh. Sanygoung Han, "Semantic web Search Model for information retrieval of the semantic data", 2003

4. Test Web data, http://ec.cse.cau.ac.kr/test_webdata.html

5. RDF document, http://ec.cse.cau.ac.kr/okchoi/person_book.rdf

6. Tim Berners-Lee, work in progress, October 1998, Why RDF model is different from the XML model, http://www.ilrt.bris.ac.uk/discovery/rdf/resources/#sec-examples