의미 검색을 위한 시맨틱 저장소 설계

정연진° 구태완 이광모
한림대학교 컴퓨터공학과
{yjjung°, taewani, kmlee}@hallym.ac.kr

Design of Semantic Repository for Searching the Semantics

Yeonjin Jung° Taewan Gu, Kwangmo Lee
Dept. of Computer Engineering, Hallym University

요 약

시맨틱 웹은 네트워크에 존재하는 자원에 의미를 부여하고 이를 컴퓨터가 자동으로 처리 할 수 있도록 설계된 차세대 지능형 웹이라 할 수 있다. 그러나 수많은 양의 문서를 대상으로 각각의 의미를 정의하기 어려울 뿐만 아니라 이미 정의된 의미를 바꾸는 데에도 문제가 있다. 또한 여러 종류의 의미를 중첩적으로 갖는 것이 힘들며, 문서 검색을 하는데 있어 전체 네트워크를 대상으로 검색해야 한다는 점에서 소모적인 면을 가지고 있다. 본 논문에서는 문서의 의미 정의에 있어 내재된 문제점과 다의성 문제를 해결하기 위해, 문서와 대응할 수 있는 의미를 구조화된 형식에 따라 분리하여 이를 통합적으로 관리 할 수 있는 SR(Semantic Repository)을 설계 하고자 한다. 여기서 SR은 각 문서에 대응되는 의미와 각 문서들 사이의 독립된 의미관계만을 나타내므로 검색에 따른 부하 감소를 기대할 수 있다.

1. 서 론

팀 버너스리(Tim Berners-Lee)에 의해 1989년 처음 제안된 월드와이드웹(WWW)은 간단한 HTML(HyperText Markup Language)을 이용하여 사용자가 정보에 쉽게 접근할 수 있게 하였고, 이런 단순성으로 인해 급속한 정보의 증가를 가져왔다. 그러나 정보의 양이 급속도로 증가한 상황에서 이러한 웹의 단순성은 사용자가 원하는 문서에 접근하는데 상당한 걸림돌이 되기도 한다[1]. 이런 문제점이 발생하는 가장 주된 원인은 현재의 웹이 사람을 위한 것이고 단순히 결과를 브라우저에 보여주기 위한 결과용 웹이기 때문이다.

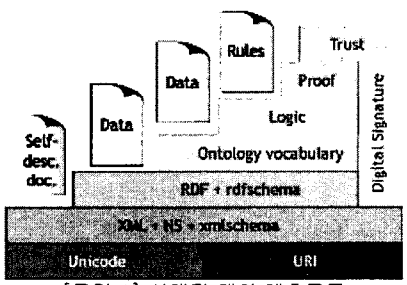
그러나 시맨틱 웹은 웹상에 존재하는 정보들을 사람뿐만 아니라 컴퓨터 프로그램 같은 기계들이 해독하고 작업하기 용이하게 표현하고자 하며, 정보간의 유기성까지 체계적으로 나타내고자 하였다. 자연어 위주의 기존 웹 문서와는 달리 컴퓨터가 해석하기 쉽도록 의미를 부여하는 계층을 갖도록 하는 것이다[2].

하지만 시맨틱 웹에 관한 연구는 태그(tag)를 이용하여 문서화 시키고 그것을 근거로 검색하는 점에 있어 기존의 그것과 크게 달라진 점이 없다. 또한 시맨틱 태그들이 문서에 묻혀 있기 때문에 다른 형식의 의미를 부여하거나, 수정이 용이하지 못하다는 단점이 존재한다. 그리고 문서를 검색할 경우 해당 문서의 전체를 검토하고 문서에 대한 인덱스를 중앙에서 관리하기 때문에 인덱스 생성에 따른 부하가 매우 크게 된다.

본 논문에서는 이러한 문제들을 해결하기 위해 문서에 1:1로 대응되는 의미를 문서에서 분리하여 해당 의미를 RDF(Resource Description Framework)를 이용하여 구조화 하고, 문서와 문서간의 의미관계를 SR(Semantic Repository)에 저장한다. 이것은 문서의 의미에 대한 추론 규정을 정의하고, 문서에 대한 메타 정보를 표현하여 문서 본문 전체를 검토하지 않아도 되는 장점이 있으며, 이들을 SR에 저장하기 위해 문서와 문서간의 관계를 구조적으로 정의함으로써 웹에서 기계의 인식력을 향상시켜 서로 다른 형식의 의미 부여가 가능해지도록 한다.

2. 관련연구

2.1 시맨틱 웹(Semantic Web)



[그림 1] 시맨틱 웹의 계층구조

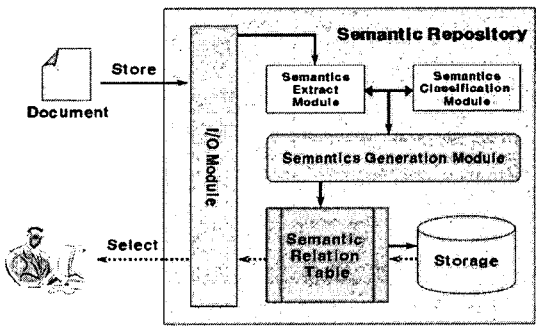
시맨틱 웹은 RDF와 온톨로지(ontology)에 기반하여 W3C(World Wide Web Consortium)이 제정한 데이터의 추상적 표현법이다. 기본적으로는 다음 [그림 1]과 같이 URI(Uniform Resource Identifier)를 기본으로 하는 계층적인 구성 요소로 이루어 진다. [그림 1]은 임의의 개념을 모듈 방식으로 정의할 수 있는 XML(eXtended Markup Language)과 이름공간(Name Space), 자원을 기술 할 수 있는 문법적 기반을 제공하는 RDF와 RDF 스키마, 특정 도메인의 지식을 표현하기 위한 개념들을 표현할 수 있는 온톨로지가 있으며 그 상위에는 질의와 추론이 가능 하도록 하는 규칙(Rule)와 논리(Logic)가 존재한다.

2.2. RDF(Resource Description Framework)

RDF는 W3C의 가장 기본적인 시맨틱 웹 언어로서 웹에 있는 자원에 대한 메타 정보를 표현하기 위한 언어이다[3]. RDF는 구조적인 메타 데이터를 인코딩, 교환, 재사용할 수 있도록 하는 하부 구조이며, 메타 데이터를 처리하기 위한 근거 즉, 웹에서 기계가 이해할 수 있는 정보를 교환하는 응용 프로그램들 사이에 상호작용을 제공한다. RDF는 기본적으로 3개의 정보를 지닌 쌍들을 정의하는데 주어(Subject), 목적어(Object), 서술어(Predicate)의 상태로 표현되며 각각은 URI로 지정할 수 있다. 따라서 사용자는 새로운 개념이나 동사를 URI를 써서 쉽게 정의할 수 있다. 그러나 RDF 자체는 개별 응용프로그램이 사용한 사물의 종류나 성질을 정의할 수 있는 방법을 제공하지 않고 있기 때문에 RDF 어휘 형태로 지정된 RDF를 확장한 RDF 스키마를 사용한다.

3. 시맨틱 저장소(Semantic Repository) 설계

시맨틱 저장소(Semantic Repository)는 의미단위의 문서 헤더(의미 헤더, Semantic Header)를 저장하여 검색이 필요할 경우 해당 헤더만을 참조함으로써 실제 검색에서 발생하는 인덱스 생성에 따른 부하를 감소시키고, 다른 형식의 의미를 부여하여 문서가 갖는 다의성 문제를 해결하는 것을 목표로 한다. 뿐만 아니라 문서와 문서간의 의미 관계(Semantic Relation)를 정의할 수 있도록 하여 문서 검색의 효율성을 높이게 된다. 다음의 [그림 2]는 시맨틱 저장소의 구조를 나타낸다.



[그림 2] Semantic Repository 구조

3.1. 의미 추출(Semantics Extract Module)

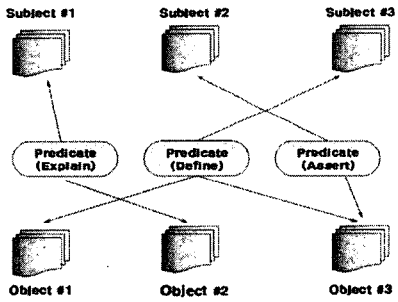
네트워크 상에 존재하는 많은 양의 언어 자료는 taxonomies라고 하는 토픽 계층으로 분류할 수 있다. 이것은 관련 있는 자료들을 색인으로 연결함으로써 문서의 검색과 접근을 더 효율적으로 할 수 있다. 그러나 문서 양이 지나치게 방대해짐에 따라 taxonomies를 유지하는 것은 거의 불가능하게 된다. 따라서 taxonomies에서 의미가 유사한 것을 추출하는 시스템을 구축하는 일은 중요한 작업이 될 것이다.

그러므로 taxonomy에서 불용어(stopword)와 무의어(noise)를 제거하고 문서의 의미를 대표할 수 있는 단어를 찾아내는데, 이때 단어들 중에서 자주 등장하는 명사와 동사만을 추출하고 그 빈도수와 함께 각각에 대한 매트릭스를 구성한다[4]. 전자를 M_n, 후자를 M_v라고 정의 한다.

각 매트릭스는 단어와 빈도수로 구성되며 빈도수를 중심으로 정렬하고 고유한 ID를 부여함으로써 문서 내에서 가장 자주 쓰인 단어들을 빠르고 쉽게 찾아낼 수 있도록 한다.

3.2. 의미 분류 및 의미헤더 생성(Semantics Classification Module and Semantic Header Generation)

의미헤더 부분을 M_n과 M_v로 구분한 것은 보다 강력하고 융통성 있는 추론이 가능하도록 하기 위함은 물론 다른 문서와의 의미 관계를 규정하고자 할 때 보다 효율적으로 검색할 수 있도록 하기 위함이다. M_v를 통해 규정할 수 있는 문서의 성격은 크게 " 설명(explain), 정의(define), 주장(assert)" 의 세 가지로 구분할 수 있다. ' 설명(explain)' 은 어떤 일의 내용이나 이유, 의의 따위를 알기 쉽게 밝혀서 풀어 쓴 문서를 의미하며, ' 정의' 는 어떤 개념의 내용이나 용어의 뜻을 다른 것과 구별할 수 있도록 명확히 한정하는 일이나 그 개념을 나타낸 문서이고, ' 주장' 은 자기의 학설이나 의견 따위를 굳이 내세우는 문서를 의미한다. 따라서 모든 문서는 이 세 가지 중에 하나로 성격 지어 짐으로써 SR 내에서 유사한 성격의 문서들을 통합하여 관리할 수 있도록 해준다. 다음의 [그림 3]은 의미헤더 내에서 의미 추출 결과를 분류 한 것이다. 이것은 문서 내에서 의미가 다양하게 조합될 수 있기에 의미의 다양성을 표현하기에 적합한 구조를 가지게 된다.



[그림 3] 의미분류 구조

3.3. 의미 관계(Semantics Relation Table)

RDF의 형식을 이용하는 의미헤더들은 각각이 술어 논리로 구성되며, 이것은 다시 의미공간(Semantic Space)상의 벡터형태로 추출이 가능하다. 그리하여 일반화된 벡터 공간모델(General Vector Space Model)[7]을 이용하여 각 의미헤더간의 유사성을 분석하여 제안된 의미 관계 테이블을 구성한다. 이때 사용되는 벡터 d_1와 d_2의 유사도는 일반적으로 다음과 같이 표현할 수 있다[6, 8].

$$sim(d_1, d_2) = (P^T d_1)^T \bullet (P^T d_2) = d_1^T P P^T d_2 \quad (1)$$

행렬 P는 기본 벡터공간에서의 의미헤더를 특정한 다른 자질 공간으로 매핑시키기 위한 변환행렬로서 $P = I_M$ (M은 의미헤더 집합 내 어휘 개수이고 I_M은 $M \times M$ 항등행렬) 일 때는 원래 용어 벡터 공간에서의 두 의미헤더간 유사도를 의미한다[8].

우선 의미헤더에 나타나는 용어들을 이용하여 아래와 같은 벡터를 구성한다.

$$[k_1, k_{2,K}, k_t] \ (t = 1,2,K,n)$$

의미헤더들을 용어들의 출현 패턴에 따라 0 과 1 로 표시한 minterm m_{2^t}을 구성하고, 각각의 minterm m_i로 쌍방향 직교 벡터 $\overleftrightarrow{m}_i$를 구성한다[7]. 여기서 m_1는 용어들을 어느 것도

포함하지 않는 의미헤더를 말하며, m_2는 용어 k_1 만을 포함하고 있는 의미헤더를 말한다[7].

$$m_1=(0,0,...,0)\ \ m_2=(1,0,...,0)\ \ ...\ \ m_{2^t}=(1,1,...,1)$$
$$\vec{m}_1=(1,0,..,0,0)\ \ \vec{m}_2=(0,1,..,0,0)\ \ \vec{m}_{2^t}=(0,0,..,0,1)$$

용어들의 출현 패턴이 minterm m_r과 일치하는 용어 가중치 $w_{i,j}$ 을 합산한 상관계수 c_{ir}는 식(2)처럼 계산된다[7].

$$c_{ir}=\sum_{d_j|g_l(\vec{d}_j)=g_l(m_r)\ for\ all\ l} w_{ij} \quad : w_{i,j}\rightarrow[k_i,d_j]\text{의 조합} \quad (2)$$

용어 k_i가 포함되어 있는 모든 minterm m_r의 벡터를 합산하여 정규화한다[7].

$$\vec{k}_i=\frac{\sum_{\forall r,g_i(m_r)=1} c_{ir}\vec{m}_r}{\sqrt{\sum_{\forall r,g_i(m_r)=1} c_{ir}^2}} \quad (3)$$

즉, $\vec{k}_i$들로 행렬 A가 구성된다. 또한, 행렬 A는 질의 q를 $q'=A^Tq$ 처럼 이차원 공간으로 변환시킨다. 의미헤더도 마찬가지로 변환되며, 표준 코사인 측정을 사용하여 변환된 질의와 비교 한다[8].

$$sim(d,q)=\cos(A^Td,A^Tq)=\frac{d^TAA^Tq}{\left|A^Td\right|\left|A^Tq\right|} \quad (4)$$

변환 $d'=A^Td$는, 의미헤더들이 행렬 A로 길이-정규화된다면, d와 모든 의미헤더들 A_i 사이의 유사점의 벡터가 산출된다[8].

$$A^Td=\left[A_1^Td,A_2^Td,\mathrm{K},A_n^Td\right] \quad (5)$$

본 논문에서는 위의 식들을 이용하여 행렬 A를 만들고, 식(4)을 이용하여 서로 다른 의미헤더간의 유사도를 측정하여 문서간의 관계 테이블을 정의하게 된다.

4. 결론 및 향후 과제

시맨틱 웹에 관한 연구에 있어 기존의 큰 틀을 벗어나지 못하는데 따르는 문제점은 문서가 갖는 다양한 의미를 반영하지 못하고 문서의 수정이 어려우며, 이를 관리하는 데에 따르는 부하가 상당하다는 것이다. 본 논문은 의미저장소(Semantic Repository)를 이용하여 그런 문제점들을 해결하고자 하는 시도라고 볼 수 있다.

네트워크에 존재하는 문서에 1:1로 대응되는 의미를 문서에서 분리하고 해당 의미를 RDF(Resource Description Framework)를 이용하여 구조화 한 다음, 문서와 문서간의 의미관계를 SR(Semantic Repository)에 저장하는 구조는 문서간의 다양한 의미를 표현 가능하게 할 뿐만 아니라 기존의 문서 전체에 대한 인덱싱에 기인하는 부하 문제도 작은 크기의 의미헤더(Semantic Header)만을 검토하게 되므로 부하의 조정이 가능하게 된다. 마지막으로 향후, 의미 기반의 검색에서도 문서간의 유사성에 근거한 관계 테이블을 이용한다면, 보다 나은 검색결과를 반환하게 될 것이다.

앞으로 의미헤더 부분의 수정과 관리에 대한 구체적인 연구가 진행되어야 하며, SR의 성능에 대한 객관적 분석을 통해 기존에 시맨틱 웹에서 검색한 결과와 의 비교 고찰이 필요할 것이다.

또 각 SR간의 네트워크를 구성함으로써 검색의 범위를 확대하고, SR 간의 계층 구조를 정의하여 검색의 속도를 향상시킬 수 있도록 한다. 또한 SR을 기반으로 한 시맨틱 웹 브라우저를 설계하는 것을 향후 과제로 하고자 한다.

참고문헌

[1] 최중민, " 시맨틱 웹의 개요와 연구 동향" 정보과학회지 제21권 제3호, 2003년 3월

[2] Berners-Lee, T., Hendler, J. & Lassila, " The semantic web" , Scientific American, May. 2001.

[3] RDF Primer, W3C Recommendation, 10. Feb. 2004. http://www.w3c.org/TR/rdf-primer

[4] Yugyung Lee, Changgyu Oh, Eun Kyo Park, Intelligent knowledge discovery in peer-to-peer file sharing, CIKM pp.308-315, 2002

[5] Cristianini, N., Shawe-Taylor, J., and Lodhi, H., Latent semantic kernels, Journal of Intelligent Information System, vol.18, no.2/3, pp.127-152, 2002.

[6] 장정호, 김유섭, 장병탁, " 헬름홀츠머신 학습 기반의 의미커널을 이용한 문서 유사도 측정" , 추계학술대회, 한국정보과학회, 2003.04

[7] Wong, S. K. M., Ziarko, W., and Wong, P. C. N., Generlized vector space model in information retrieval, ACM SIGIR Conference on Research and Development in Information Retrieval, pp.18-25, 1985.

[8] I. M. Soboroff and C. Nicholas. Collaborative Filtering and the Generalized Vector Space Model. In Proceedings of the 23rd Annual International Conference on Researech and Development in Information Retrieval (SIGIR), 2000.