온톨로지 기반의 태그 정보 검색

한기동*, 이창훈*
*건국대학교 컴퓨터공학과
e-mail:myhanki@konkuk.ac.kr

Tag Information Search based on Ontoloty

Ki-Dong Han*, Chang-Hun Lee*
*Dept of Computer Engineering, Kun-Kuk University

요 약

기존의 웹 서비스가 수동적이고, 단방향 통신을 축으로 뒀다면 현재의 웹 서비스는 점차 능동적이고 변화되었으며, 양방향 통신 환경을 지향하게 되었다. 이러한 웹 서비스 변화의 흐름을 일컬어 웹 2.0 이라 한다. 웹 2.0 세대를 살아가는 사용자들은 기존과 다른 다양한 정보의 홍수에 노출되게 되었다. 이들은 일방적이고, 제한적인 정보를 얻는 기존 환경에서 탈피, 스스로 가치 있는 정보를 생산해 내기 시작했고, 이렇게 생산된 정보는 인터넷을 통해 다른 사용자와 교류하며 더욱 가치 있는 정보를 창출해 나가고 있다. 이런 발전 과정에서 지속적으로 더욱 더 커져가는 정보를 더 빠르고 정확하게 공유하는 기술이 필요하게 되었고, 현재 이런 필요성을 충족시키는데 유용한 기술의 한 갈래로 나온 것이 태그와 시맨틱 웹으로 대표되는 온톨로지 이다. 태그는 정보의 주제나 표제를 나타내는 단어를 해당 컨텐츠 정보를 제공하는 사이트에서 정보 분류 단위로 사용, 이를 통한 더 빠른 정보 공유를 할 수 있게 되었다. 시맨틱 웹은 현재의 인터넷과 같은 다양한 리소스에 대한 정보와 자원 사이의 관계-의미 정보를 기계(컴퓨터)가 처리할 수 있는 온톨로지 형태로 표현하고, 이를 자동화된 기계(컴퓨터)가 처리하도록 하는 기술이다. 이 논문에서는 웹 2.0의 대표기술이라 할 수 있는 온톨로지 기법을 이용, 기존 태그의 정보 분류 효율을 높이기 위한 태그와 태그의 의미관계 형성을 제안하였다.

1. 서론

최근 인터넷의 급속한 발달과 더불어 인터넷 사용자의 증가와 발달은 웹 서비스 환경을 다양한 형태로 발전시켰다. 기존의 웹 서비스가 수동적이고, 단방향 통신을 축으로 뒀다면 현재의 웹 서비스는 점차 동적이고 능동적이며, 양방향 통신 환경을 지향하게 되었다.

이러한 웹 서비스 변화의 흐름을 일컬어 웹 2.0 이라 한다. 웹 2.0의 핵심은 기존의 단방향에서 벗어나 사용자가 능동적으로 참여하는 양방향 환경이다.

네티즌 세대라 불리 울만큼 인터넷 사용에 익숙한 사용자들은 기존과 다른 다양한 정보의 홍수에 노출되게 되었다. 이들은 일방적이고, 제한적인 정보를 얻는 기존 환경에서 탈피, 스스로 가치 있는 정보를 생산해 내기 시작했다. 이렇게 생산된 정보는 인터넷을 통해 다른 사용자와 공유하고, 교류를 통해 더욱 가치 있는 정보를 창출해 나가고 있다.

이런 발전 과정에서 지속적으로 더욱 더 커져가는 정보를 더 빠르고 정확하게 공유하는 기술이 필요하게 되었고, 현재 이런 필요성을 충족시키는데 유용한 기술로 대두되고 있는 것이 태그와 시맨틱 웹으로 대표되는 온톨로지 이다.

태그(Tag)는 정보의 주제나 표제를 나타내는 단어를 해당 컨텐츠 정보를 제공하는 사이트에서 정보 분류 단위로 사용하고 있다. 이 태그의 사용으로 많은 정보를 각 특성에 맞게 분류가 가능해 졌고, 이를 통해 좀 더 빠른 정보를 공유할 수 있게 되었다.

시맨틱 웹(Semantic Web)은 현재의 인터넷과 같은 분산 환경에서 리소스(웹 문서, 각종 화일, 서비스 등)에 대한 정보와 자원 사이의 관계-의미 정보(Semantics)를 기계(컴퓨터)가 처리할 수 있는 온톨로지(Ontology)형태로 표현하고, 이를 자동화된 기계(컴퓨터)가 처리하도록 하는 프레임워크이자 기술이다.

이 논문에서는 웹 2.0의 대표기술이라 할 수 있는 온톨로지 기법을 이용, 기존 태그의 정보 분류 효율을 높이기 위한 태그와 태그의 의미관계 형성을 제안하였다.

2. 관련 연구

2.1 웹 2.0

웹 2.0(Web 2.0)은 단순한 Text 웹사이트의 집합체를 웹 1.0으로 보고, 다양한 리소스를 제공하는 하나의 완전한 플랫폼으로의 발전을 웹 2.0이라고 한다 이 용어는'O'Reilly Media'에서 2003년부터 사용하기 시작하면서 대중화 되었는데, 웹 2.0이 데스크톱 컴퓨터의 응용 프로그램을 대체할 것으로 예견되어지고 있다. 기존의 웹의 개념은 생산자가 이따금 갱신하는 정적 HTML 페이지들의 집합일 뿐이었다. 웹 2.0은 이와 다르게 블로그의 트랙백이나 위키와

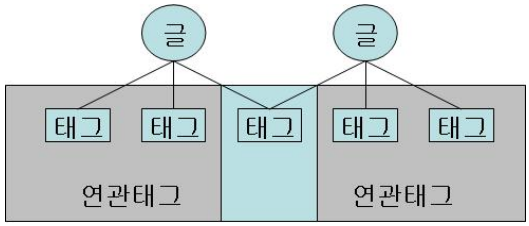
같이 각 주체가 생산자이면서 동시에 소비자가 되는 상호작용을 통해 컨텐츠를 재생산하는 사회적 양방향 네트워크를 형성해 나가는 것이다.

2.2 온톨로지

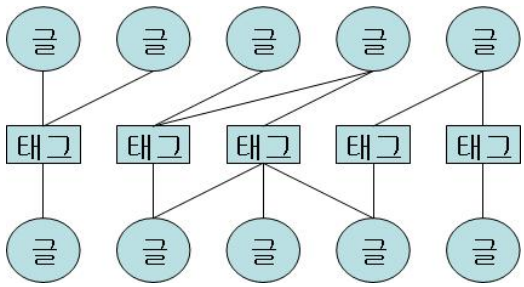
온톨로지(Ontology)란 사람들이 사물에 대해 생각하는 바를 추상화하고 공유한 모델로, 정형화되어 있고 개념의 타입이나 사용상의 제약 조건들이 명시적으로 정의된 기술을 말한다. 또한 프로그램이 이해할 수 있어야 하므로 여러 가지 정형화가 존재한다. 이는 전산학과 정보 과학에서, 특정한 영역을 표현하는 데이터 모델로서 특정한 영역(Domain)에 속하는 개념과, 개념 사이의 관계를 기술하는 정형(Formal) 어휘의 집합으로 정의된다. 웹의 등장은 전통적인 정보검색을 비롯하여 지식관리와 일반 상거래 등 사회 전 분야의 변혁을 초래하였다. 특히 웹 정보 검색은 소장 자료를 대상으로 하는 제한된 검색에서 웹을 통해 접근할 수 있는 전자자원을 대상으로 하는 검색을 가능하게 하였다. 웹의 급속한 발달로 인해 검색 대상 범위의 확대는 보다 정교한 검색을 필요로 하게 되었으며, 지능화된 정보 검색 시스템 개발을 촉진하는 계기가 되었다. 이런 계기를 바탕으로 웹 자원을 효과적으로 관리할 수 있는 정보 검색의 새로운 도구의 필요성이 대두되었다. 온톨로지는 시맨틱 웹을 구현할 수 있는 도구로써 지식개념을 의미적으로 연결할 수 있는 도구이다. 온톨로지는 자연어의 기계 번역과 인공지능 분야에서 활용되며, 최근에는 특정 분야의 인터넷 자원과 그 사이의 관계를 기술하는 온톨로지를 사용하는 시맨틱 웹과 이것에서 파생된 시맨틱 웹 서비스 등의 핵심 요소로서 주목받고 있다.

2.3 태그

태그는 일반적으로 이름표, 수확물의 표지, 제품의 상표 등을 뜻하는데, 웹에서도 태그는 어떤 글이나 자료에 붙여 놓은 추가 정보를 뜻한다. 태그는 특정한 규칙 없이 사용자가 임의 데로 부여할 수 있고, 이렇게 부여된 태그를 토대로 다양한 정보 간 관계를 형성하는 수단이 된다. 태그가 웹 2.0에서 주목받는 이유는 누구나 손쉽게 쓸 수 있고, 태그와 태그 사이의 연관 관계를 맺을 수 있다는 점이다. 그림 1과 같이 글과 글 사이에는 연관성이 없지만 같은 태그를 입력 하였을 경우 태그를 통해서 글과 글 사이의 연관 관계를 맺을 수 있다. 그림 2와 같이 하나의 글에 여러 개의 태그를 입력하는 경우, 하나의 글에 입력된 여러 개의 태그는 서로 연관성을 맺게 된다. 이런 태그들 중 중복된 태그의 개수가 많을수록 정보의 연관도도 높아진다고 볼 수 있다. 그리고 연관 태그 중에서 가장 높은 연관성을 가지고 있는 것이 대표 태그가 될 수 있다.



(그림 1) 태그를 이용한 글과 글 사이의 관계



(그림 2) 글과 글 사이의 연관 태그

3. 태그 간 온톨로지 관계 형성 기법 적용

시맨틱 웹 이란, 사람의 머릿속에 있는 언어에 대한 이해를 컴퓨터 언어로 표현하고 이것을 컴퓨터가 사용할 수 있게 만드는 것인데, 특별한 분산 환경을 갖춘 웹에 구현하자는 것이다. 이것은 기계가 정보검색과 같은 사람의 요구를 더 잘 이해하고 적절하게 반응하기 위해서이다. 사람과 기계 사이에 진정한 커뮤니케이션이 가능하기 위해서는 사람이 이해하는 수준으로 기계도 언어를 이해할 수 있어야 한다. 사람들이 언어를 이해하는 방식을 보통 개념화라고 하는데, 즉 사람들은 세상에 있는 각각의 사물이나 사건들을 경험하면서 이 들 속에 들어있는 특징을 파악해서 언어로 개념화한다. 이렇게 컴퓨터에서도 사람이 갖고 있는 개념과 같은 것을 일종의 데이터베이스 형태로 만드는 기술을 온톨로지 기술이라고 부른다. 이는 현재의 인터넷과 같은 분산 환경에서 리소스에 대한 정보와 자원 사이의 관계-의미 정보를 기계(컴퓨터)가 처리할 수 있는 형태로 표현하고, 이를 자동화된 기계(컴퓨터)가 처리하도록 하는 프레임워크이자 기술이다. 기존의 HTML로 작성된 문서는 컴퓨터가 의미정보를 해석할 수 있는 메타 데이터보다는 사람의 눈으로 보기에 용이한 시각정보에 대한 메타데이터와 자연어로 기술된 문장으로 가득 차 있다. 예를 들어 수박은 초록색이다.

라는 예에서 볼 수 있듯 이라는 태그는 단지 수박과 초록색이라는 단어를 강조하기 위해 사용된다. 이 HTML을 받아서 처리하는 기계(컴퓨터)는 수박 이라는 개념과 초록색 이라는 개념이 어떤 관계를 가지는지 해석할 수 없다. 단지 태그로 둘러싸인 구절을 다르게 표시하여 시각적으로 강조를 할 뿐이다. 게다가 수박이 초록색이라는 것을 서술하는 예의 문장은 자연어로 작성되었으며 기계는 단순한 문자열로 해석하여 화면에 표시한다. 시멘틱 웹은 XML에 기반한 시멘틱 마크업 언어를 기반으로 한다. 가장 단순한 형태인 RDF는 <Subject, Predicate, Object>의 트리플 형태로 개념을 표현한다. 위의 예를 트리플로 표현하면 <urn:수박, urn:색, urn:초록>과 같이 표현할 수 있다. 이렇게 표현된 트리플을 컴퓨터가 해석하여 urn:수박 이라는 개념은 urn:초록 이라는 urn:색 을 가지고 있다는 개념을 해석하고 처리할 수 있게 된다. 보다 구체적인 예로 네이버가 NHN 소유임을 나타내는 트리플은 <http://naver.com, urn:wikipedia-ko:소유, http://nhnco-rp.com> 과 같이 된다. 이러한 트리플 구조에 기반 하여 그래프 형태로 의미정보인 온톨로지를 표현한다.

포탈 뉴스나 블로그 사이트에서 사용자에 의해 글이 작성되고, 이글에는 기본적인 태그가 형성되어진다. 형성된 글의 특성에 맞게 표현 된 다양한 태그들을 어떤 식으로 분류하는가에 따라 검색의 효율에 큰 차이를 보인다. 이 논문에서 제안하는 기법은 서로 다른 영역의 태그 소스로부터 온톨로지의 개념과 개념관계에 따라 정보를 인스턴스화 하여 웹 사이트의 데이터 테이블에 저장 시킨다. 예를 들어 블로그에 '세계의 다양한 격투기에 대한 고찰' 이란 글을 작성한 후, 태그를 다음과 같이 단다. '태권도', '유슬', '가라데', '유도', '레슬링', '수박', 등... 이런 형태의 태그를 단다면, '수박'이란 태그 때문에 검색에서 과일이란 검색을 하면, 주제와 무관하게 위의 글이 분류되어 보여지게 된다. 하지만, 온톨로지 기법을 통해 글의 태그들에 '격투기'란 의미관계를 부여한다면, 과일을 검색 시 위의 글이 검색 되지는 않는다. 또한 격투기로써의 '수박'을 검색할 때 도 과일이 검색 되지 않기에 더욱 효율적인 검색이 이루어진다. 이렇게 각 글들을 작성하고, 이 글들의 태그를 온톨로지 기법을 이용한 태그 그룹화를 통해 더욱 나은 검색을 실현할 수 있게 된다.

4. 실험결과 및 분석

태그 검색 속도와 정확성을 측정하기 위해 동일한 데이터베이스 환경에서 일반적인 태그 검색 기법과 본 논문이 제시한 온톨로지 관계 검색 기법을 비교 하였다. 테스트 데이터베이스에 사용자가 하나의 글에 1 ~ 5개 정도의 태그를 입력한다고 가정하고 랜덤으로 입력된 태그를 기존 태그 방식과 제안 태그 방식으로 나누어서 입력하였다. 태그는 한글 3자리로 구성되었으며, 제시한 데이터베이스 에는 온톨로지 스키마를 추가 태그 간 관계 입력하였다.

본 논문에서 제안한 기법의 속도, 정확도와 일반적 태그 기법의 차이는 다음 표 1과 같다.

데이터 3,000건을 비교하였을 경우, 최소에서 최대 배 검색 속도가 향상 되었고, 정확도는 의 차이가 났다.

<표 1> 3,000개의 글에 대해서 생성된 태그 검색 시간과 정확성

데이터	일반 태그검색		제안된 태그검색	
	속도	오차빈도	속도	오차빈도
100	23.1ms	1%	17.3ms	0%
500	29.0ms	2%	18.1ms	0.1%
1,000	34.7ms	3%	21.4ms	0.1%
1,500	41.6ms	3%	23.8ms	0.8%
3,000	52.3ms	4.5%	31.5ms	1.1%

5. 결론

본 논문에서는 웹 2.0과 함께 등장한 시맨틱웹의 온톨로지 기법을 태그와 연계하여 더욱 정확하고 빠른 처리를 가능케 하는 태그 관리 기법을 제시하였고, 기존 태그 기법과 비교하여 우수한 성능을 보여주었다. 이 기법을 활용하여 실제 검색에서 더욱 많은 활용 가능성을 제시해 본다. 향후 과제는 실제 웹과 연동하는 일이 남았으며, 태그와 태그 사이에 온톨로지 관계를 더욱 확장 시키는 일이 남았다.

참고문헌

[1] Wallace A. P., and Ana M. C., "An Ontoloty Based-Approach for Semantic Search in Portals," In Proceedings of the 15th International Workshop on Database and Expert Systems Applications, pp. 127-131, 2004

[2] Atom Publishing Format and Protocol, http://틔.coverpages.org/atom.html, 2007

[3] R. Sinha, "A congnitive Analysis of Tagging," http://www.rashmisinha.com/, 2005

[4] 경북대학교 온톨로지 검증 및 온톨로지 기반 인스턴스 생성에 관한 연구, 최종 보고서, pp. 52-65, 2006

[5] 최호찬, " 인터넷의 새로운 문화 블로그", 경향잡지, pp. 107-109, 2004년 3월호

[6] 오량, "무한으로 확장하는 웹 2.0 세계", 월간 말, pp. 224-225, 2006년 9월호

[7] 강필구, 김남중, 이예슬, 채진석, "웹 2.0을 위한 효율적인 태그 관리 시스템의 설계 및 구축," 한국정보과학회 2006 가을 학술발표 논문집, 제 33권 제2호(D), pp. 170-173, 2006.