정보 추출과 완벽한 재사용 보장

한국정보검색위원회에서는 위원간의 의욕 고취와 새로운 검색
및 데이터베이스 관련기술 보급을 위해 매월 연구발표회를 개최하고 있다.
본 코너에서는 주제논문을 게재함으로써 정보 검색과 관련된 정보를
제공하기 위해 마련된 것이다. 〈편집자〉

조영환/(주)언어기술
이상기/(주)K4M

서론

인터넷 혹은 웹(Web)은 전 세계의 모든 컴퓨터를 거대한 정보망의 형태로 묶어 냄으로써 정보의 전달과 공유에 대한 새로운 관점을 생성시켰다. 이는 기존의 모든 관념을 일시에 무너뜨리고, 거기에 새로운 관념으로 정복하고도 남음이 있는 규모의 혁신이다. 인터넷 이전의 컴퓨터는 단지 호기심을 채워주거나 주어진 업무를 처리하는 도구의 위치에 있었다.

그러나 컴퓨터와 컴퓨터가 서로 정보를 주고 받을 수 있다는 약간의 확장된 시각과 각각의 컴퓨터에서 공개한 정보가 다른 컴퓨터의 정보와 쉽고 자연스럽게 합쳐질 수 있다는 이해하기는 쉽지 않지만 상당히 유용한 기능이 실로 엄청난 결과를 만들어 내고 말았다.

여기에는 현재는 너무 알려진 HTML (HyperText Markup Language)이라는 언어가 중심에 있었다. 실로 대단한 언어이다. 사이버 공간으로 이름 지어진 가상의 공간에서 정보를 교류하고 전달하는 새로운 정보유통 구조를 만들어 내게 된 것이다. 이제 변모될 웹에 대하여 생각하여 보아야겠다. 지금의 웹은 사람이 보기 위한 것이다. 참 좋은 점도 있지만, 속도가 늦다거나 보아야 할 정보가 많다거나 잠시 자리를 비운다거나 해서 느끼는 불편이 없어지면 어떠한 일이 일어날까? 누구인가가 대신 해주어야나 해결될 일이다.

누구인가가 나를 대신하여 작업하도록 하기 위해서는 웹상의 문서나 정보를 이해할 수 있는 누구여야 할 것인데, HTML로 작성된 웹의 정보는 사람이나 보기 위한 것이어서 누구를 사람으로 밖에 생각할 수 없다. 그러한 역할을 "컴퓨터 프로그램"이 담당하게 하려면 어떻게 하여야 할까 고민해보자. HTML은 이해할 수 없기 때문에 곤란하다. 이 때문에 XML(eXtensible Markup Language)이라는 언어가 등장하게 되었다. 등장이라기 보다는 고안되었다고 보아야 합당할 것이다.

본 글은 XML언어에 대한 설명은 생략하고 XML의 아이디어와 구조문서 정보검색에 관하여 다루고자 한다. 인터넷 혹은 웹의 막대한 파급력의 원천이었던 HTML의 힌트만큼의 XML 힌트를 전달해서 앞으로의 새로운 웹을 상상하는 계기가 되었으면 한다. 먼저 생각하는 것이 먼저 도착해 있는 것과 유사한 의미를 갖는 시대이기 때문이다. 더불어 구조문서에 대한 정보검색이 기존의 정보검색과 어떻게 다른 모습을 가지고 있는지를 언급하겠다. 이 또한 새로운 웹을 지탱할 중요한 기능이기 때문이다.

XML의 개요

누구나 손쉽게 다양한 정보를 접하게 됨으로써 습득하고 이해되고 나아가서는 관리해야 할 모든 정보가 일순간에 폭발적으로 증가하게 되었다. 정보의 폭발적인 증가현상에 대응하기 위해서는 필연적으로 정보의 효율성 측면을 고려하여야 한다. 즉 한번 생성되고 전달되면 그 생명을 다하는 1회성의 정보유통이 아닌 정보생산과 유통, 그리고 재생산의 모든 과정에 대한 좀 더 효율적인 정보인프라를 요구하게 되었다.

HTML은 정보의 제시만을 목적으로 하는 언어(presentation only)이기 때문에 HTML 언어로 저작된 정보는 일회성을 갖는다. 즉 정보사용자가 포맷팅된 문서를 보면서 정보를 이해해야 하기 때문에 정보를 재사용하기 위해서는 자신이 원하는 포맷과 형태로 다시 작성하여야 하는 번거로움을 수반한다.

비용의 관점에서는 정보의 생성과 유통비용 외에도 정보를 재생산하기 위한 추가적인 비용이 필요한 것이다. 이는 일회용 도시락처럼 정보를 담는 그릇과 그 내용을 분리하지 않고 하나로 녹여 놓음으로 인해 생기는 비용의 추가 발생이다.

종이 문서에 기반한 정보의 생성, 공유, 보관, 교환은 대규모의 정보가 빈번히 갱신되는 환경에서는 관리가 불가능하므로, 전자 정보에 의한 문서의 필요성 때문에, 문서 정보 구조화 언어인 SGML(Standard Generalized Markup Language)이 1986년에 국제표준(ISO 8879)으로 제정된 후 전자 정보의 저장과 교환을 위한 표준으로 채택되고 있다.

그러나 SGML 언어는 너무 복잡하기 때문에 이를 처리하는 프로그램 또한 규모가 커야만 했다. 모든 것을 표현하여야 한다는 목적 때문에 이것저것 정의하여야 하는 것이 늘어난 때문이다. 반면에 웹에서는 웹-브라우저라는 프로그램으로 문서를 처리해야 하기 때문에, XML은

HTML 문서	XML 문서
〈HTML〉〈BODY〉 〈TABLE〉 〈TR〉 〈TD〉客地〈/TD〉 〈TD〉황석영〈/TD〉 〈TD〉창작과 비평사〈/TD〉 〈TD〉1974 〈/TD〉 〈/TR〉 〈/TABLE〉 〈/BODY〉〈/HTML〉	〈?XML version = 1.0 ?〉 〈단행본〉 〈도서명〉客地 〈/도서명〉 〈저자〉황석영 〈/저자〉 〈출판사〉창작과 비평사 〈/출판사〉 〈출판년도〉1974 〈/출판년도〉 〈/단행본〉

〈그림〉 HTML문서와 XML 문서

SGML을 웹 환경에 맞도록 변형한 언어로 처리의 용이성을 추가한 언어라고 이해하는 것이 좋겠다.

XML이 1998년 2월에 권고안으로 제정되어 SGML문서를 웹 환경에서 쉽게 저작, 관리, 전송, 공유하기 위해 사용되기 시작되었다. 그러므로 SGML을 잠시 잊고 생각해 보면, 인터넷 혹은 웹상에서 정보의 생성단계에서부터 정보의 재사용을 고려하여 새로이 나타나게 된 언어이다. SGML과 XML은 모두 정보를 표현하는 외형(physical structure or formatting information)과 그 내용(logical structure, content information)을 분리하여 정보를 표현하기 때문에 정보의 전달, 공유 그리고 재사용을 극대화하기 위해 제안된 언어이다.

XML을 통해 구조화된 정보는 외형적인 형태를 배제하여 존재하므로 추가적인 개입 없이 원하는 정보만을 추출하고 전달하며 완벽한 재사용을 보장하게 된다. 〈그림〉에서 같은 정보를 표현하기 위해 사용된 HTML과 XML의 문서를 비교하면 그 차이가 확연하게 드러난다.

〈그림〉에서 HTML문서와 XML문서는 정확히 같은 내용을 정보사용자에게 보여준다. 그러나 정보를 이해하는 입장에서 보면 XML로 되어 있는 문서를 통해 客地라는 것이 책의 이름을 의미하고, 저자가 황석영이라는 정보의 이해를 훨씬 쉽게 할 수 있다.

또한 위와 같은 문서들에서 도서명과 저자만을 따로 모아 리스트를 작성하는 것을 가정하자. HTML문서로 시작한 경우에는 추가적인 수작업이 항상 필요하다. 반면에 XML문서로 시작한 경우에는 〈도서명〉으로 시작해서 〈/도서명〉로 끝나는 부분과 〈저자〉로 시작해서 〈/저자〉로 끝이나는 부분을 추출해서 저장하는 일을 프로그램이 자동으로 수행할 수 있다. 이렇게 구조화된 정보는 한번 생성된 정보를 추가적 작업 없이 정보추출과 전달 그리고 재생산 할 수 있게 한다. 즉 일회용 정보가 아닌 끊임없이 재사용될 수 있는 정보를 만들어 낼 수 있는 것이다.

〈그림〉의 문서가 요즘 인기가 있다는 인터넷 서점들의 문서라고 가정해 보자. 여러 인터넷 서점의 책 가격을 비교한다고 하면 일부러 사람이 그러한 일을 할 이유가 있을까? 그리고 새로운 책을 구입한 인터넷 서점에서 일일이 책에 대한 정보를 입력할 필요가 있을까? 책에 대한 정보를 다양하게 해서 간단한 안내와 자세한 안내로 각각 다른 문서를 만들 필요가 있을까? XML은 위와 같은 질문에 답변으로 제시할 수 있는 대안이다. 사람과 더불어 컴퓨터 프로그램이 인터넷 정보를 읽을 수 있고 이해할 수 있고 생산할 수 있는 웹이 바로 XML로 열어가려는 새로운 웹이다.

구조 문서와 정보 검색

개인과 기업 그리고 어떠한 조직에서도 고유의 목적을 달성하기 위해 습득되고 생산되는 정보가 홍수처럼 증가하고 있다. 정보과다는 정보비용을 증가시키고 비효율을 야기해 조직의 경쟁력을 저하시키게 된다. 정보과다상황에서 정보고립을 배제하기 위해서는 새로운 정보인프라를 구축해야 한다. 인터넷을 정보 환경으로 선택한다면 그 첫 단계가 XML을 이용한 정보의 구조화이다.

XML을 통한 정보구조화는 정보의 내용과 정보가 보여지는 외형을 분리하게 된다. 〈그림〉의 XML문서의 예에서도 보듯이 특정문서에 내재된 정보를 자동으로 분리가능 하여 새로운 정보로 재생산 할 수 있다. 표현되어 전달된 정보 자체가 유기적인 정보구조 차체를 포함하므로 단순한 재입력부터 특정정보의 통합까지를 포함한 정보재사용을 보장하게 된다.

정보의 구조화는 정보를 분자와도 같이 가장 작은 단위로 분할하고 이를 구조적으로 통합하는 방식으로 각 단위정보들이 서로 분리되어 통합정보가 달라지더라도 단위정보(엘리먼트)자체에 기술된 정보는 그대로 유지되는 것을 의미한다. 정보의 관리의 목표는 정보의 적절한 통제와 저장/변경 그리고 원활한 정보유통을 위한 인프라를 제공하는 것이다. 정보의 통제와 저장을 위해서는 정보가 되도록 큰 단위로 존재하는 것이 유리하다. 즉 통제하고 저장할 단위의 수가 적은 것이 시스템의 부하와 처리량을 줄일 수 있게 된다.

그러나 정보유통과 변경을 위해서는 정보는 세분화 될 수 있는 가능한 작은 단위로 존재하는 것이 유리하다. 정보가 너무 큰 단위로 존재하면 정보의 변경과 유통을 위해 한번에 움직이고 처리해야 할 정보의 양이 너무 많아지게 된다.

XML은 웹의 정보자체를 구조화하므로 정보자체가 작은 단위정보로 원하는 단위로 나뉠 수 있고 이를 원하는 단위로 통합할 수 있다. 따라서 사용자가 원하는 단위로 모든 관리가 이루어질 수 있어 응용에 대한 요구사항이 바뀌어도 유연하게 대처할 수 있다. 이러한 장점들은 공동저작(collaborative authoring)을 통한 정보생성의 효율화와 정보통제와 유통의 효율화를 통하여 새로운 정보관리 환경을 가능케 한다.

정보의 생성은 제품을 생산하는 생산라인에 비유할 수 있다. 즉 다듬어 지지 않은 생각을 가진 다수의 생성자가 서로가 맡은 부분에 대한 생각을 정리하여 작성하는 과정이다. 레고블럭과 같이 단위정보를 바탕으로 하여 논리적 분할 하에 각자의 정보를 생성하고 이러한 정보를 통합함으로써 정보를 생성할 수 있도록 해 주는 것이 정보의 구조화이다. 대량의 정보를 생성하는 환경에서 전체와 부분의 통합을 보장하여 주는 정보인프라가 구축된다면 정보생성에서의 비효율은 산업혁명과도 같은 정보혁명을 촉진시켜 정보생성비용을 현저히 감소시킬 수 있다.

구조화된 정보라면 통제되는 단위와 유통되는 단위를 모두 가변적으로 유지할 수 있다. 즉 행정단위를 시,도-읍,면-동,리와 같이 원하는 크기로 세분하여 사용하듯이 XML로 구조화된 정보는 사용자가 원하는 단위로 정보를 세분화할 수 있고 이들의 포함관계를 다룰 수 있으므로, 통제와 유통에 있어서의 효율성을 모두 극대화 할 수 있다.

정보검색시스템은 정보를 색인하여 색인구조로 저장하고 이를 바탕으로 정보의 내용에 대한 질의를 통해 정보를 찾도록 도와주는 시스템이

다. 정보검색시스템 측면에서 구조화된 정보와 그렇지 않은 정보를 다루는 가장 큰 차이는 정보의 분할을 정보색인과 검색에서도 고려하였는가 하는 점이다. 이는 기존의 정보검색의 단순확장이 아닌 구조화의 패러다임을 그대로 유지시킬 수 있는 새로운 정보검색시스템을 의미한다.

엘리먼트 단위로 세분된 정보를 분할하여 색인 검색할 수 있고 이러한 엘리먼트 간의 논리적 관계도 검색에서 사용할 수 있는 검색시스템 기능이 제공되어야 한다. 정보검색은 대량의 정보 속에서 정보사용자가 원하는 정보를 빠르게 찾아낼 수 있도록 도와주는 중요한 정보인프라 중의 하나이다.

정보검색시스템의 성능은 사용자가 원하는 정보를 얼마만큼 효율적으로 찾아내느냐에 달려있다. 이러한 정보검색시스템의 성능은 두 가지의 구성요소를 갖고 있다. 정보를 얼마나 빠르게 찾아내느냐에 관한 정보탐색(searching) 요소와 더불어 정보검색시스템 자체의 정보표현 능력이 어느 정도인가에 관한 정보표현(representation or indexing) 요소가 있다. 도서관시스템을 예로 들면, 아무리 대량의 도서를 빠르게 찾아내더라도 책의 제목으로만 찾아낼 수 있다면 정작 사용자는 자신이 원하는 정보를 빠르게 찾아 낼 수 없고, 많은 양의 검색결과리스트를 헤매다가 검색을 포기하게 될 것이다. 만약 저자정보 혹은 출판사, 요약정보로도 검색이 가능하다면 사용자에게는 자신이 질의할 수 있는 영역자체가 확장되므로 좀 더 정확한 정보를 빠르게 찾아낼 수가 있다.

구조화 정보검색 시스템은 상당한 분량의 전문을 논리적으로 분할하고, 각 분할된 다단계 논리단위(엘리먼트)를 이용하여 검색함으로써 아무리 큰 분량의 문서라도 사용자는 불편함 없이 검색을 행할 수 있게 해 준다. XML을 통해 구조화된 정보의 가장 큰 특징은 정보가 계층적인 포함관계를 가진 형태로 분할되어 있고, 문서(instance)내에 기술된 논리적 구조정보가 이러한 분할된 정보들을 통합하고 있다는 것이다.

기존의 정보검색엔진에서 대상으로 했던 단일 단위화 된 정보와는 표현된 정보자체가 다르게 구성되어 있다. 구조화되지 않은 정보는 정보자체를 분할하기 위해서 단일 단위화 된 정보에서 필요한 정보를 사용자가 수동으로 추출하는 작업을 거쳐야 한다. 그러나 대량의 자료를 일일이 수동으로 처리하는 것은 거의 불가능한 작업이므로, 전체 정보를 하나로 다룰 수밖에 없게 된다. 따라서 하나의 문서 내에 포함된 정보가 제한될 수밖에 없어 메타데이타(metadata)형태로 부가 정보를 처리하게 된다.

전자화 된 문서의 증가는 이러한 부가정보의 수동부착 등의 작업을 불가능하게 하여 서비스의 어려움을 야기했고, 사용자들에게는 정보전체를 기술한 전문(fulltext)을 요구하게 하였다. 즉 도서, 논문 또는 특허정보 등에 대한 전문이 전자화되면서 상당한 분량의 전문 자체에서 특정부분만을 검색하고자 하는 사용자의 요구가 증가되었다. 그러나 기존의 검색 시스템에서 전문을 모두 색인하여 검색서비스를 제공하게 되면 문서와 문서간의 변별력을 상실하게 되어 검색시스템 자체가 무용하게 된다.

앞서 열거한 정보구조화에서 오는 장점을 모두 수용하고 이를 사용자에게 제시하기 위해서는 기존의 검색시스템보다는 진일보한 새로운 검색시스템이 필요하다. 다음은 구조화 정보검색시스템이 가져야 할 기능적인 특성이다.

▲ **기존 정보검색 시스템의 기능을 기본으로 제공** : 구조화 정보검색시스템이 가장 최소한으로 제공해야 할 기능은 당연히 기존의 정보검색시스템이 갖는 모두 기능을 기본적으로 제공해야 한다는 것이다. 이러한 기존 정보검색시스템의 기본 기능은 자동색인기능, 빠른 탐색기능, 문서순위화 기능 등을 포괄한다.

▲ **엘리먼트별 분할 검색 기능**(Element Content Search) : 정보구조화를 통해 논리적으로 구분된 계층적인 정보단위에 대한 개별 검색이 가능해야 한다. 이는 검색결과가 문서단위가 아닌 엘리먼트별도 될 수 있음을 의미한다. 책 한 권이 구조화되어 XML로 표현되어 있다면 책 단위의 검색은 물론 장(chapter), 절(section)단위로도 개별 순위화가 가능하여 사용자가 분할된 정보단위로 검색결과를 제공받을 수 있어야 한다.

▲ **엘리먼트 구조검색 기능**(Structure Constrained Content Search) : 계층분할된 엘리먼트들은 서로 포함관계와 상관관계를 갖는다. 즉 장(chapter)은 절(section)을 포함하고 절은 다수 개의 단락(paragraph)을 포함하는 일종의 트리 구조를 갖는다. 이러한 포함상관관계가 검색에서 사용 가능해야 한다. 즉 어떤 엘리먼트가 다른 엘리먼트의 하위 엘리먼트인지의 여부를 통해 사용자가 검색의 영역을 제한할 수 있어야 한다. 정보의 분할과 이의 통합정보가 색인과 검색에서 모두 고려되어야 하며 이는 정보검색 시스템의 색인저장구조를 완벽히 재구성해야 함을 의미한다.

▲ **애트리뷰트 검색기능** : 애트리뷰트는 XML에서 분할된 정보블럭에 레이블을 다는 것과 같은 강력한 기능을 한다. 이러한 애트리뷰트 검색기능은 엘리먼트별로 관리될 수 이어야 분할된 정보를 제대로 골라낼 수 있다. 즉 같은 이름의 엘리먼트라도 애트리뷰트를 달리하여 정보를 구성할 수 있으므로 애트리뷰트를 통한 정보분할의 인식 및 이를 통한 검색이 가능해야 한다.

▲ **혼합검색** : 내용검색, 구조검색, 애트리뷰트 검색이 혼합되어 동시에 사용자가 조합하여 사용할 수 있는 기능으로 정보의 구조화된 특징을 모두 사용하여 사용자가 원하는 정보를 가장 빨리 찾을 수 있는 기능을 제공해야 한다.

▲ **분할단위의 사용자 지정 기능** : XML로 표현된 정보는 분할 엘리먼트가 기하급수적으로 증가한다. 문서의 크기가 별로 크지 않더라도 내부적으로 수십만개의 엘리먼트를 가질 수 있다. 그러나 이러한 모든 분할엘리먼트가 사용자에게 의미가 있는 것은 아니다. 따라서 검색속도의 증가와 색인저장의 효율성을 위해 사용자(DTD 작성자)가 분할 단위를 지정할 수 있는 기능이 제공되어야 한다.

▲ **대용량 문서 및 다사용자에 대한 대처 기능** : XML문서는 정보의 조직화와 재생산을 목적으로 생성되는 경우가 대부분이므로 그 대상 문서의 개수가 일반적 규모의 정보검색 시스템보다 훨씬 많게 된다. 또한 특정 조직내의 정보유통을 목표로 하므로 다수의 사용자가 동시에 검색서비스를 사용하는 환경이 대부분이다. 따라서 검색 시스템 자체적으로 대용량 문서(1천만건 이상, 테라바이트급)에 대한 처리가 필수적으로 제공되어야 하며, 다사용자 접근시의 성능저하를 방지하는 기능이 필연적으로 요구된다.

▲ **repository와의 연동 API 제공** : XML문서의 저장과 관리를 담당하는 repository는 정보를 저장하고 유지하는 역할을 수행한다. 따라서 검색시스템은 검색을 지원하는 generic한 부분과 특정 응용에 맞도록 설계된 repository와 손쉽게 연동하여 통합할 수 있는 연동API를 가져야 한다. 대부분의 XML repository가 데이터베이스를 하부

구조로 사용하므로 검색시스템도 이러한 데이터베이스와 연동할 수 있는 API가 제공되어 원활하게 밀결합할 수 있는 기능이 제공되어야 한다.

▲ **정보 무결성 보장을 위한 색인자동 복구 기능** : 대부분의 repository는 정보의 무결성을 보장하기 위해 데이터베이스에서 제공하는 locking 기능과 롤백(rollback)기능을 사용한다. 따라서 repository는 네트워크의 불안정이나 클라이언트의 작업 취소로 인해 작업이 반영되기 전의 상태로 다시 자료를 되돌리게 된다. 이러한 상황에서 검색시스템 또한 자동복구 기능을 이용하여 항상 repository에서 저장된 정보와 동일한 무결성 상태의 정보를 동시에 유지하고 있어야 한다. 만약 검색시스템에서 색인자동복구 기능이 제공되지 않는다면 repository와 검색시스템과의 정보괴리 현상이 발생하여 정보의 무결성이 깨지고 정보의 신뢰도를 저하시키게 되므로, XML을 대상으로 한 구조화 정보검색 시스템이라면 자동복구기능은 반드시 요구되는 기본기능이다.

XML을 통해 구조화된 정보를 최대한 이용할 수 있는 구조화 정보검색엔진은 다음과 같은 장점을 정보사용자에게 제공한다.

▲ **정보접근(document access point)의 다양성** : 기존의 단순 색인어만으로 문서에 접근하는 기능 외에 문서의 계층적 구조정보와 내용정보를 혼합하여, 다양한 각도로 문서를 검색할 수 있다. 예를 들어 특정한 키워드가 제목에 포함되거나, 저자 소속, 요약 등의 문서의 특정 부분을 지칭하여 검색의 범위를 한정할 수 있으며, 6개의 장으로 구성된 문서, 세명이상의 저자가 있는 문서 등 문서의 구조정보를 이용하여 검색이 가능하다. 이러한 검색은 사용자가 생각하는 모든 조합이 가능하다.

▲ **동적인 문서 제시(dynamic presentation) 가능** : XML은 문서의 일부만을 골라내어 자유롭게 조합하여 사용자에게 제시할 수 있다. 즉 문서전체를 보낼 필요없이 문서의 요약, 저자부분, 참고문헌만을 독립적으로 분리하여 볼 수 있으며, 스타일정보를 외적으로 부가하여 여러 형태로 보여줄 수 있다. 따라서 불필요한 네트워크의 과부하와 시스템의 과부하를 감소시킬 수 있다. 예를 들어 1장만 분리하여 보여줄 수 있으며 사용자가 원하는 형태로 변경이 가능하다. 또한 문서내에 포함된 그림만을 조합한 새로운 형태로 제시가 가능하여 제시된 문서의 재사용 또한 보장된다.

▲ **검색정보의 일관된 관리** : 일반적으로 문서에서 특정 부분을 자동 인식 (예를 들어 제목부분)하면 오인식이 항상 개입되게 된다. 따라서 제목이 아닌 엉뚱한 부분이 제목으로 인식되어 검색될 수 있다. 그러나 XML은 구조에 대한 유효성(validation)을 마친 상태이므로 항상 일관된 정보를 바탕으로 검색에 사용될 수 있다.

▲ **다양한 부가 정보 제공** : 문서의 여러 부분이 계층적으로 구분되어 있으므로 검색 외의 여러 부가 정보를 얻을 수 있다. 예를 들어 가중치계산을 위한 용어 분포 정보, 문서 요약, 분류를 위한 용어분포정보 등을 부가적으로 얻어 활용 할 수 있다. 또한 문서의 브라우징을 위한 다양한 정보를 제공한다. 따라서 참고문헌에 있는 논문을 이용하여, 본 논문을 참고하고 있는 논문에 대한 브라우징 및 본 논문의 저자가 쓴 다른 논문 검색 등을 가능케 해준다.

구조화된 정보를 대상으로 한 검색 시스템은 일반적으로 사용되어 온 기존의 정보검색시스템이 요구하는 요구사항과 더불어 정보구조화 측면에서 많은 요구사항을 추가로 요구한다. 기존의 검색 시스템과 구조화 검색시스템이 대상문서가 XML이라는 단순한 비교점 외에도 정보의 구조화라는 특성을 모두 활용할 수 있도록 해야 한다는 큰 차이를 갖음을 의미한다.

XML이 기반이 되는 새로운 웹의 방향

기존의 HTML 언어가 주축이 되는 웹에 XML언어가 가세하면서 어떠한 일이 벌어지게 될 것인가를 예측하는 일은 그다지 어렵지 않다. 다만 언제가 그 시점으로 적합할 것인가가 문제일 것이다. 본 글에서는 시간을 한참의 앞으로 이동시켜 몇 가지의 중요한 웹의 응용 분야가 가능할 것인가를 논의하도록 한다.

▲ **인터넷 기반의 출판** : 현재의 HTML언어로는 종이 형태의 출판을 인터넷에 구현하기가 쉽지 않다. 페이지에 대한 개념이나 다단, 글자와 그림의 자연스러운 어울어짐 등이 HTML 언어에 존재하지 않기 때문이다. 그러나 XML로 출판을 하게 된다면 외형은 신경을 쓰지 않아도 좋다. 언제인가는 출판에 필요한 여러 가지의 조판기능이 XML로 작성된 문서를 아름답게 꾸며줄 것이기 때문이다. 이와 더불어 멀티미디어 기능이 어울어져 동적으로 반응하는 출판물을 기대할 수 있겠다.

▲ **자동화된 정보교환** : 둘 이상의 기관이 같은 형태의 XML 문서로 정보를 작성하고 교환할 수 있다면 그러한 기관들에 있는 모든 정보는 쉽게 공유될 수 있다. 병원의 경우를 예로 들어보면 A라는 병원의 산부인과에서 출생한 어린이를 B라는 병원의 소아과에서 진료하여야 하는 경우에 의사는 A병원의 진료기록을 자신의 병원 환자 관리 데이터베이스에 끌어 놓기(Drag and Drop) 방식으로 옮겨 놓음으로 해서 이전의 환자에 대한 이력을 모두 얻을 수 있다. 대상을 바꾸고 행동 양식을 자동화해서 다시 고찰해 보면 직장인들이 매년 작성해야 하는 세금 공제 신청서를 사람이 작성할 필요가 없다는 것이다. 어찌 보면 참으로 행복하지만 정보화 사회가 그렇듯이 어찌 보면 위태하기도 하다.

▲ **지능형 웹 에이전트** : 프로그램이 정보를 이해하자는 것이 중요한 시작점이라고 언급한 것을 상기하자. 그러한 프로그램을 보통 에이전트라고 한다. 여기에 지능형이라는 것은 추론 정도의 능력을 가지고 있다고 보면 되겠다. 보고 싶은 케이블 TV의 프로그램을 찾아내려고 고생할 필요 없이 에이전트에게 명령하면 된다. 그리고 그러한 프로그램을 좋아하는 사람들이 역시 자주 보는 프로그램도 덤으로 추천해 달라고 한다면 아주 좋겠다.

▲ **자동 언어 번역** : HTML 문서에서는 컴퓨터 프로그램이 자동으로 어떠한 내용이 문장의 일부인지 아니면 제목인지, 어떠한 분야의 문서인지를 판단하기가 아주 어려웠다. 그러나 XML문서는 이미 고정되어 있는 틀을 가지고 있으므로 좀더 정확한 번역이 가능할 것이다. 그러므로 비로소 전세계의 인터넷 정보가 하나로 묶이게 될 것이다.

위에서 기술한 것들은 새로운 웹의 영향력을 시사하고는 있지만 전체의 모습을 표현하기에는 부족함이 있다. 다만 힌트가 되어서 XML언어가 매개가 되는 새로운 웹에 대한 기대를 맘껏 부풀리고 기다려 보아야겠다.