차세대 웹에서 XML과 RDF 문서를 처리하는 XML&RDF 검색 에이전트

한기덕° 권혁철
부산대학교 컴퓨터공학과
{templer°, hckwon}@pusan.ac.kr

XML&RDF Retrieve Agent Processing XML and RDF Document in The Next Generation Web

Gi-deok Han°, Hyuk-chul Kwon
Department of Computer Science and Engineering, Pusan National University

요 약

차세대 웹을 표현하는 단어로 XML(extensible markup language)과 시맨틱 웹(Semantic Web)을 꼽을 수 있다. XML은 1996년 W3C (World Wide Consortium)에서 제안한 데이터 표현 능력이 높은 언어이며, 시맨틱 웹은 사람이 읽고 해석하기에 편한 현재의 웹 대신에 컴퓨터가 이해할 수 있는 형태의 새로운 언어로 표현해 기계들끼리 서로 의사소통을 할 수 있는 지능형 웹을 말하는 것으로써, 현재 XML을 기반으로 하는 RDF(Resource Description Framework)나 온톨로지 기술을 통해 시맨틱 웹 구축방안에 관한 연구가 활발히 진행되고 있다. 본 논문에서는 차세대 웹에서의 정보 공유를 위한 검색 에이전트의 역할 및 에이전트 간의 구조에 관한 설명, XML&RDF 검색 에이전트의 설계 모델 및 현재까지 구현된 시스템의 개요를 보여준다.

1. 서 론

차세대 웹을 표현하는 단어로 XML(extensible markup language)과 시맨틱 웹(Semantic Web)을 꼽을 수 있다. XML은 1996년 W3C (World Wide Consortium)에서 제안한 언어로써, HTML보다 홈페이지 구축 기능, 검색 기능, 클라이언트 시스템의 복잡한 데이터 처리를 용이하게 하는 등 여러 가지 장점을 가지고 있으며, 시맨틱 웹은 사람이 읽고 해석하기에 편한 현재의 웹 대신에 컴퓨터가 이해할 수 있는 형태의 새로운 언어로 표현해 기계들끼리 서로 의사소통을 할 수 있는 지능형 웹을 말하는 것으로써, 현재 XML을 기반으로 하는 RDF(Resource Description Framework)나 온톨로지 기술을 통해 시맨틱 웹 구축방안에 관한 연구가 활발히 진행되고 있다.

논문 제목에서 에이전트라 칭한 것은 논문에서 제안하는 시스템이 웹에서의 정보 공유를 위해 분산 배치된 에이전트의 하나로 동작하여 더욱 고도화된 서비스를 제공하려는 목표를 가지고 있기 때문이며, 본 논문에서는 차세대 웹에서의 정보 공유를 위한 검색 에이전트의 역할 및 에이전트 간의 구조에 관한 설명, XML&RDF 검색 에이전트의 설계 모델 및 현재까지 구현된 시스템의 개요를 보여준다.

본 논문의 단락은 1. 서론, 2. 관련 연구, 3. 차세대 웹에서의 에이전트들을 이용한 검색, 4. XML&RDF 검색 에이전트의 모델, 5. 구현 및 성능 평가, 6. 결론 및 향후과제, 7. 참고 문헌으로 구성되어있다.

이 논문은 교육인적자원부 지방연구중심대학육성사업(차세대물류IT기술연구사업단)의 지원에 의하여 연구되었음

2. 관련 연구

2.1 XML and RDF

데이터 교환과 공유에 유용한 XML 문서를 저장하고 검색하기 위한 방법은 2000년을 전후하여 매우 많은 연구가 진행되었으며, 국내의 경우 RDBMS, ORDBMS를 이용한 연구, XML Parser의 성능 및 이용에 관한 연구, XML 정보의 처리 방법에 관한 연구 등 수 많은 연구가 이루어졌다.

RDF(Resource Description Framework)는 XML을 기반으로 하여 확장된 언어로써, XML보다 의미 표현 능력이 높은 가장 기본적인 시맨틱 웹 언어이다.

2.2 Semantic Web

현재 Web의 확장 형태인 Semantic Web은 정보를 잘 정의된 의미 있는 형태로 표현해서 가지고 있으며, 기계와 사람이 작업을 더욱 수월하게 할 수 있도록 지원하고,[1] application, enterprise, community boundary간의 정보의 공유, 정보의 재사용을 제공하는 공통된 framework를 제공한다.[5]

3. 차세대 웹에서의 에이전트들을 이용한 검색

차세대 웹에서의 검색은 분산된 정보들을 기계가 이해하고 분석하여 정보를 가져오는 지능형 웹에서의 검색으로, 이것이 가능하려면 기계가 정보자원의 의미를 해석하고, 기계들끼리 서로 정보를 주고받으면서 자체적으로 필요한 일을 처리해야만 가능하다.

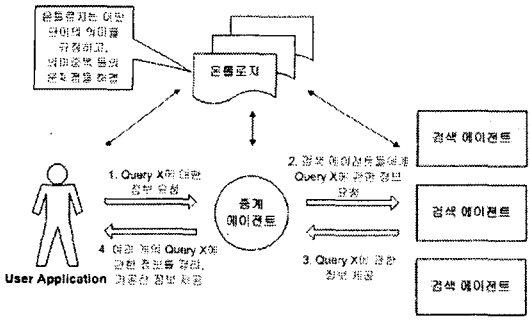

그림 1 차세대 웹에서 에이전트를 이용한 정보 검색

그림 1은 차세대 웹에서 사용자가 에이전트를 이용해 정보를 얻는 과정을 묘사한 그림이다.

사용자는 웹에서 어떤 정보를 얻기 위한 의미를 표현하는 단어(어떤 의미를 가지는지 온톨로지를 통해 분명하게 구별될 수 있는 중의성이 없는 단어)를 온톨로지를 통해 선택하고, 중계자의 역할을 수행하는 에이전트를 통해 많은 검색 에이전트에게 해당 단어와 관련된 정보를 수집한다.

이런 식으로 수집된 정보는 XML, RDF로 표현되어 있으며, 수집된 정보는 중계 에이전트를 통해 분석, 정리되어 사용자에게 제공된다.

3.1 에이전트들의 역할

표1 에이전트의 역할

종류	역할
중계 에이전트	사용자가 접근하는 에이전트로 사용자가 요청한 Query에 관한 정보를 가지고 있을 것으로 예상되는 검색 에이전트에 Query를 보내어 정보를 수집, 가공한 후, 사용자에게 제공하는 에이전트
검색 에이전트	실제로 데이터를 가진 에이전트로 요청된 Query에 관한 정보를 중계 에이전트로 전송한다. 어떤 Query나 정보에 대해 전문 에이전트도 존재 가능하다.

표1은 에이전트의 역할을 설명한 표이다. User Application과 에이전트들은 의미를 파악하기 위해 웹의 어느 곳에 위치한 온톨로지를 URI에 의한 접근에 의해 이용하며, 수집된 정보나 검색된 여러 개의 정보를 어떤 식으로 정리하고, 각각의 정보에 대해 어떤 식으로 Ranking을 매길 것인가와 관련된 작업은 User Application이나 2개의 에이전트 어느 곳에서나 수행이 가능하다.

3.2 검색 에이전트의 요구 사항

에이전트 간의 정보 요청 및 교환은 XML, RDF 등과 같이 의미 표현 능력이 있는 언어로 이루어지며, 관련기술은 Soap 웹서비스나 시맨틱 Soap 웹서비스 등이 있다. 정보 교환이 이루어지면 수집된 정보를 분석하기 위해 전송되는 정보도 역시 의미적으로 표현되어야만 한다.

검색 에이전트가 제공하는 정보는 다양할 것이나, 상당 부분은 어떤 Query와 관련된 연관 정보일 것이며, 이런 연관 정보는 XML, RDF 문서로 표현이 가능하다.

기능의 호출 측면에서 저장/삭제보다는 검색이 빈번하게 일어나며, 따라서 저장/삭제 속도보다는 검색 속도의 향상에 초점을 맞추어야 한다.

4. XML&RDF 검색 에이전트의 모델

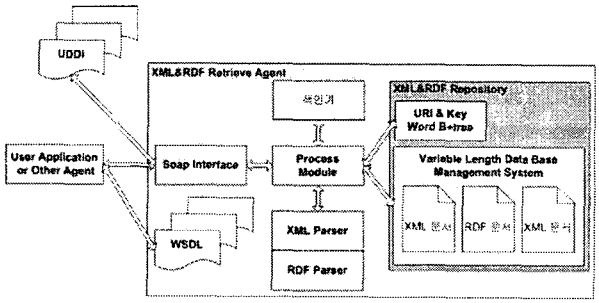

그림 2 XML&RDF 검색 에이전트의 모델

그림 2는 XML&RDF 검색 에이전트의 모델을 묘사한 그림이다. 사용자 응용 프로그램이나 다른 에이전트와의 연결은 Soap 웹 서비스를 통해 이루어지며, 실제 처리는 Process Module에서 수행된다. 단어를 색인하기 위한 색인기, XML과 RDF 문서를 파싱하기 위한 Parser가 존재하며 검색 시에 사용되는 URI나 Keyword는 B+tree에 저장되며, XML과 RDF 문서는 온전하게 Variable Length Data Base Management System(= VLDBMS)에 저장된다.

4.1 XML&RDF 문서 간의 Link

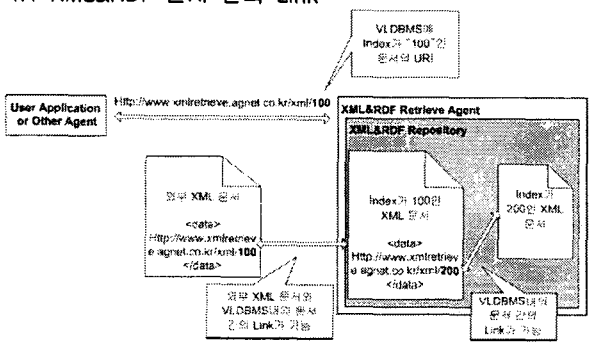

그림 3 XML 문서 간의 Link

XML&RDF Repository는 "URI&Keyword B+tree"와 "VLDBMS"를 합친 부분을 말한다. XML&RDF Repository에 XML 문서 그 자체를 저장함으로써, XML 문서 간의 Link를 통해 Indexing이 가능하며, 데이터 저장 형태를 쉽게 변경할 수 있으며, Index의 추가/삭제가 용이하다.

B+tree의 Key는 단어뿐만이 아니라, URI도 사용된다. VLDBMS에 저장된 모든 문서는 유일한 URI를 가지고 있어, 각 문서간의 Link 기능을 더욱 효율적으로 지원한다.

또한, 이런 구조를 통해 어떤 URI나 단어와 관련된 XML 문서 전체를 검색하는 속도는 매우 빠르며, XML 문서 내의 특정 정보를 추출하는 것은 Parser를 거쳐야

하므로, 검색 속도가 느리다. 본 논문에서의 검색 에이전트는 Query와 관련된 XML과 RDF 문서를 검색하여 제공하는 것을 가장 중요시하였으며, 모델을 만들 때에도 빠른 검색 속도와 간단한 구조를 염두에 두었다.

5. 구현 및 성능 평가

5.1 구현

시스템은 C++ Language를 이용하여 구현하였다. 앞에서 보여준 검색 에이전트의 모델을 완벽하게 구현한 것은 아니나, EJB, .Net의 Soap 웹서비스 기능을 테스트하였고, C++ Language를 이용하여 Soap 웹서비스 기능의 구현, DOM을 이용한 XML Parser, 색인기 모듈, B+tree 모듈, VLDBMS, Process 모듈 등을 구현하였으며, 실험은 Process 모듈, B+tree 모듈, VLDBMS, XML Parser를 연결하여 성능 평가를 수행하였다.

5.2 성능 평가

▷ 테스트 환경

표2 테스트 환경

품명	규격
CPU	Intel(R) Xeon(TM) CPU 3.06GHz
Memory	3.75GB RAM
OS	Microsoft Windows Server 2003

▷ 테스트 결과

- 각 동작을 1회에 100,000번씩 4회 실행

표3 XML 저장, 검색, 삭제의 성능 평가

동작	1회	2회	3회	4회	평균
1.XML 문서의 저장	3.9 sec	12.6 sec	6.5 sec	3.8 sec	6.7 sec
2.XML 문서에 정보 추가 후, XML Repository에 추가 (Tag가 미리 포함되어있음)	4752.1 sec	4774.2 sec	4771.2 sec	4682.9 sec	4745.1 sec
3.XML 문서에 Tag 및 정보 추가 후, XML Repository에 추가	4592.7 sec	4693.7 sec	5112.1 sec	4961.3 sec	4839.95 sec
4.XML 문서 자체를 검색	4.2 sec	2.1 sec	2.0 sec	2.0 sec	2.575 sec
5.XML 문서에서 Tag를 찾아 Tag의 정보를 검색 후 추출	2356.3 sec	2318.9 sec	2339.3 sec	2319.4 sec	2333.475 sec
6.XML 문서 자체를 삭제	3.8 sec	3.0 sec	2.8 sec	2.8 sec	3.1 sec
7.XML 문서에서 Tag와 Tag의 정보를 삭제	4259.3 sec	4727.8 sec	4598.4 sec	4873.8 sec	4614.825 sec

표4 XML 저장, 검색, 삭제의 성능 평가2

동작	동작 100,000번 수행 시, 걸린 평균 시간 ($=A_t$)	1개의 동작이 실행되는데 걸린 시간 ($=A_t/100,000$)
1.XML 문서의 저장	6.7 sec	0.000067 sec
2.XML 문서에 정보 추가 후, XML Repository에 추가 (Tag가 미리 포함되어 있음)	4745.1 sec	0.047451 sec
3.XML 문서에 Tag 및 정보 추가 후, XML Repository에 추가	4839.95 sec	0.0483995 sec
4.XML 문서 자체를 검색	2.575 sec	0.00002575 sec
5.XML 문서에서 Tag를 찾아 Tag의 정보를 검색 후 추출	2333.475 sec	0.02333475 sec
6.XML 문서 자체를 삭제	3.1 sec	0.000031 sec
7.XML 문서에서 Tag와 Tag의 정보를 삭제	4614.825 sec	0.04614825 sec

표3과 표4는 XML 저장, 검색, 삭제와 관련된 동작을 수행하여 성능을 평가한 표이다.

표4의 결과를 본다면, XML 문서 자체의 저장, 검색, 삭제의 경우(1,4,6번의 동작), 100,000번을 수행하는데 2~7초 정도의 시간이 걸렸으며, 한 번의 동작의 경우 0.00002~0.00007초 정도의 시간이 걸렸다. 이 정도 속도면 대용량 XML의 검색도 가능하며, 특히 검색의 경우 100,000번의 동작을 2~3초 정도로 저장, 삭제에 비해 빠른 성능을 보여주고 있다. 그러나 XML Parser를 사용하는 2,3,5,7번의 동작의 경우 XML 문서를 Parsing하여 정보를 처리해야하는 시간 때문에 1,4,6번의 동작에 비해 약 400~1700배 정도 느리다는 것을 알 수 있다. 따라서 제안하는 방식으로 시스템 모델을 구성한다면, XML Parser의 성능이 중요한 요소가 되며, XML Parsing을 하지 않는 처리의 경우에는 높은 처리 속도를 보장한다. 즉, URI나 단어와 관련된 XML 문서 자체를 검색하고, 제공하는 경우에는, 높은 처리 속도를 보장한다.

6. 결론 및 향후과제

차세대 웹에서의 검색은 분산된 정보들을 기계가 이해하고 분석하여 정보를 가져오는 지능형 웹에서의 검색으로, XML이나 RDF과 같이 의미 표현이 가능한 언어로 표현되어지며, 분산된 많은 에이전트간의 협력, 처리를 통한 고도화된 서비스를 제공할 것이다. 본 논문에서는 차세대 웹에서의 에이전트간의 동작 방식의 정리 및 차세대 웹에서 문서를 검색하여 제공하는 검색 에이전트의 모델을 제안하였으며, XML 처리에 있어서의 속도에 관한 성능 평가를 수행하였다.

향후과제로는 XML이나 RDF의 검색 모델 및 중요 Keyword 추출 방법, Ontology의 활용 방안 등의 연구를 진행할 예정이다.

7. 참고 문헌

[1] Berners-Lee, T., Hendler, J., Lassila, O. "The Semantic Web", Scientific American, Vol. 284 (4). (2001) 34-43
[2] A. Palival, N. Adam, C. Bornhovd, J. Sehaper "Semantic Discovery and Composition of Web Services for RFID Applications in Border Control", ISWC04, 2004
[3] EPCglobal (http://www.epcglobalinc.org)
[4] EPCglobal: The EPCglobal Network™: Overview of Design, Benefits, & Security, 2004
[5] The World Wide Web Consortium (http://w3c.org)
[6] Wei-shuo Lo, Tzung-Pei Hong, Shyue-Liang Wang, Yu-Hui Tao, "Semantic web and Multiple-agents in SCM", International journal of Electronic Business Management, Vol. 2. (2004) 122-130
[7] Inceon Paik, Wonhee Park, "Software Component Architecture for an Information Infrastructure to Support Innovative Product Design in a Supply Chain", Journal of Organizational Computing and Electronic Commerce 15(2), (2005) 105-136