제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

XML을 이용한 메타데이터 분산검색 시스템의 설계 및 구현에 관한 연구

송종철*, 홍기채*
*한국전자통신연구원 정보유통연구팀
{ jcsong, gchong}@etri.re.kr

A Study on the Design and Implementation of the System for Distributed Information Retrieval based on the Metadata using of XML

Jong-Cheol Song*, Gi-Chae Hong*
*Information Processing Team, ETRI

요 약

인터넷이 급속히 발전하고 확산되면서 정보를 효율적으로 활용하고 유통시키기 위한 연구가 활발히 진행되고 있다. OAI(Open Archives Initiative)에서는 대용량 정보를 메타데이터를 이용하여 공유하고 검색할 수 있는 프로토콜 및 임프라에 대한 연구와 표준화를 추진하고 있다. 또한 EMAF(An Extensible Multi-Agent Framework)에서는 멀티에이전트를 이용한 정보 유통 및 활용한 대한 연구가 진행 중이다.

이에, 본 논문에서는 메타데이타과 XML, 멀티에이전트를 이용한 분산검색 시스템을 설계하고 구현하였다. 본 시스템은 조정에이전트와 응용에이전트로 구성되고 에이전트간 통신에는 XML 과 OAI 의 메타데이타 하비스팅 프로토콜을 응용하였다. 메타데이터에 대한 검색을 수행하여 검색 성능을 높일 수 있었으며 또한 사용자가 입력한 문장 단위의 질의를 처리할 수 있는 기능과 관련어를 추출할 수 있는 기능도 제공한다.

1. 서론

웹의 급속한 보급과 확산으로 인하여 웹상의 정보들이 기하급수적으로 증가하고 있다. 그러나 지금의 웹환경에서는 단일 시스템상의 검색엔진을 이용한 웹 정보검색 서비스의 한계점이 나타나고 있다. 즉 계속적으로 증가하는 웹 정보를 수집하고 색인함으로써 수집된 웹 정보 및 색인을 관리하는 측면의 문제점과 대용량 웹 정보를 보유한 웹사이트에 대한 정보 수집 및 활용의 문제점, 사용자가 필요로 하는 웹 정보를 제공하는데 있어서 정확성의 문제점 등이 제시되고 있다.

이러한 문제점들을 해결하기 위해 메타검색과 같은 분산검색 기술이 활용되고 있으나 표준화된 정보 교환 환경이 미흡하고 각 검색엔진의 검색 결과에 대한 의존성이 높아 효율적인 분산 검색과 검색 결과의 정확성을 보장할 수 없는 제약점을 가지고 있다.

OAI[1]에서는 전자도서관 등 대용량 정보 및 메타데이터를 보유한 단체에서 사용할 수 있으며 상호 동작성을 가지는 분산환경 기반의 메타데이터 하비스팅 프로토콜 표준안을 제시하였다. 본 표준안에서는 메타데이터를 교환함으로써 분산정보검색 서비스를 지원하는 새로운 기반구조를 제시한다.

본 논문에서는 이러한 문제점들을 해결하기 위해 W3C(World Wide Web Consortium)[2]에서 표준화한 XML(eXtensible Markup Language)과 OAI 에서 제시한 OAI-PMH(Open Archives Initiative Protocol for Metadata Harvesting)를 응용하여 분산환경에서 사용할 수 있는 프로토콜을 설계하였고 또한 멀티에이전트를 이용하여 분산검색시스템을 구성하였으며 사용자가 입력한

질의어와 관련성을 가지는 단어를 추출하는 관련어 추출기를 구현하여 사용자가 필요로 하는 정보를 제공하는 시스템을 설계 및 구현하였다.

2. 요소 기술 및 관련 연구

분산정보검색 시스템을 구성하는 요소 기술 및 관련 연구에 대해 살펴보면 다음과 같다.

2.1 요소 기술

분산검색 시스템과 관련하여 연구 및 표준화가 진행 중인 더블린 코어 메타데이터와 XML, 통신 프로토콜, 멀티에이전트에 대해 알아 본다..

더블린 코어 메타데이터는 DCMI(Dubline Core Metadata Initiative)[3]에서 제정한 메타데이터이며 이기종 시스템 간의 상호 운영성과 기계가독성(Machine-Readable)을 지원한다. 또한 더블린 코어 메타데이터는 단순 더블린 코어(Simple Dublin Core)와 한정 더블린 코어(Qualified Dublin Core)로 구성된다. 단순 더블린 코어는 열 다섯 개의 엘리먼트로 구성되며 더블린 코어 메타데이터 엘리먼트 집합 버전 1.1.이 1999 년 7 월에 발표되었다. 한정 더블린 코어는 정보자원을 더욱 상세히 표현한 부가적인 한정어를 사용하는 메타데이터로 표준안으로 확정되지는 않았으며 계속 연구중이다.

XML 은 SGML(Standard Generalized Markup Language)의 문서구조정의 기능을 단순화하여 인터넷에서 사용할 수 있는 문서구조를 정의할 수 있도록 W3C 에서 제정한 메타언어 이다. 초기 XML 표준은 대용량 전자출판에서 활용할 수 있는 문서 정의에 목표를 두었으나 현재에는 문서 및 정보 교환을 목적으로 하는 프로토콜 정의에도 활발히 사용되고 있다. XML 버전 1.1 이 2002 년 4 월에 발표되었으며 XML 의 링크와 관련된 XML Linking Language 1.0 은 2001 년 10 월에 발표 되었다. 또한 XML 을 스키마 형식으로 구성하는 XML Schema 와 관련된 표준(Primer, Structures, Datatypes, Description)들도 2001 년 5 월에 발표 되었다.

분산검색에서 사용되는 프로토콜로 중 OAI 에서 개발한 OAI-PMH 는 여러 전자도서관들 간에 메타데이터를 공유하고 활용할 수 있는 하비스팅 프로토콜을 표준화하였다. 현재 버전 1.0 은 2001 년 1 월에 발표되었으며 버전 2.0 이 2002 년 6 월에 발표 되었다. 버전 2.0 은 XML 스키마를 사용할 수 있는 기능을 추가하였다.

분산검색에서 이용될 수 있는 멀티에이전트에 관한 연구는 활발히 진행 중이다. 그 중 EMAF 모델과 FIFA (Foundation for Intelligent Physical Agents) 모델에 대해 기술한다.

EMAF 는 에이전트의 관리와 제어를 담당하는 조정에이전트(Coordination Agent)와 다수의 응용에이전트로 구성되며 에이전트간의 협동 작업을 수행하기 위해 KQML(Knowledge Query and Manipulation Language)와 ICL(Inter-agent Communication Language)을 사용한다. EMAF 의 조정 에이전트는 요청한 서비스 내용을 분석해 이를 수행할 에이전트를 선정하고 해당 에이전트에 요구된 서비스 메시지를 전달한다. 그리고, 각 에이전트의 능력과 기능 정보를 메타지식 형태로 저장하며, 에이전트 상태 테이블 등을 유지하여 간단한 형태의 에이전트를 실현하고 있다. [4]

FIPA 모델에서는 AP(Agent Platform)를 제시하고 있다. AP 는 ACC(Agent Communication Channel), ANS(Agent Name Server), DF(Directory Facilitator), AMS(Agent Management System)으로 구성된다. ACC 는 에이전트 사이에 메시지를 전달하며, ANS 는 전역적인 에이전트 이름과 지역적인 전송 주소 사이의 매핑을 저장하며, DF 는 에이전트의 기능과 그들이 제공하는 서비스를 저장하고, AMS 는 에이전트의 생성, 삭제, 정지, 복구, 이동 등을 관리한다.[4]

2.2 관련 연구

본 절에서는 OAI-PMH 를 활용하여 개발한 시스템인 LOVE(Learning Object Virtual Exchange)와 JAFER ToolKit Project 에 대해 살펴 본다.

LOVE 는 통합 프로토타입 DL(Digital Library) 서버상에서 전자도서관의 기능과 OAI 를 통합하여 설계 및 구현한 시스템이다. LOVE 는 플로리다 대학에서 개발한 시스템으로 메타데이터 집합을 분석하기 위한 데이터 마이닝 툴과 하비스트 메타데이터 기반의 검색 서비스, 데이터 프로바이더로 부터 제공된 메타데이터의 포맷을 변화하는 변환기, 사용자가 데이터 프로바이더를 등록할 수 있는 개인화된 메타데이터 레지스트리들로 구성된다.[5]

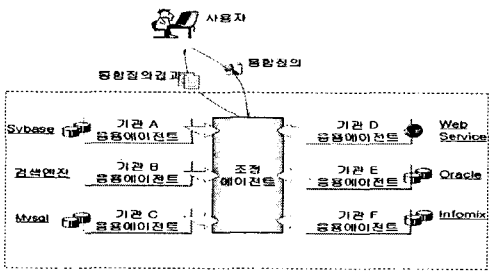
JAFER ToolKit Project 는 Z39.50 프로토콜의 상위 계층에 XML 을 기반으로 하는 API(Application Programming Interface)를 구현하였다. Z39.50 프로토콜과 관련된 서버 및 클라이언트 단의 어플리케이션 개발을 편리하게 하였다는 것이 장점이다.. JAFER ToolKit 은 JAFER 클라이언트와 JAFER 서버로 구성되며 JAFER 클라이언트는 쿼리 빌더와 웹 어플리케이션, XML 시리얼라이저로 구성된다. JAFER 서버는 XML DB 와 RDB, Z39.50 DB 를 포함하며 시리얼라이저와 XSLT 변환기를 포함한다.[6]

3. 시스템의 구조 및 설계

본 연구는 EMAF 구조를 응용한 멀티에이전트를 기반으로 구성하였으며 OAI 에서 제정한 OAI-PMH 를 응용하여 멀티에이전트를 구성하는 조정에이전트와 응용에이전트들 간의 통신 프로토콜을 설계하였다. 또한, 응용에이전트는 메타데이터 DB 에서 검색을 수행한 후 검색결과를 XML 로 변환하여 조정에이전트에게 전송하는 동작을 수행한다 또한 사용자가 입력한 질의 중 문장단위의 질의를 처리하며 관련어를 추출할 수 있는 기능도 구현되었다.

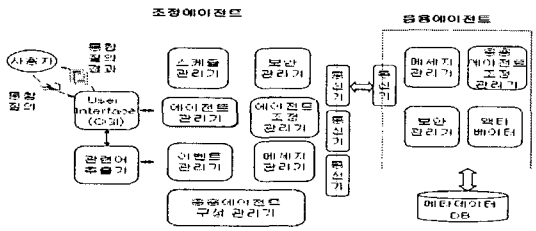
3.1 시스템의 구조

(그림 1)은 조정에이전트와 응용에이전트들 간의 역할과 구성을 보여 준다. 본 시스템은 하나의 조정에이전트와 여러 개의 응용에이전트로 구성된다.



(그림 1) 조정에이전트와 응용에이전트들 간의 구성도

(그림 2)는 조정에이전트와 응용에이전트로 구성된 본 시스템의 상세 구조를 보인다.



(그림 2) 메타데이터 분산검색 시스템의 구성도

조정에이전트와 응용에이전트를 기반으로 구성된 본 시스템은 각 에이전트들이 다음과 같은 모듈들로 구성되었으며 에이전트 사이의 통신은 XML 과 OAI-MHP 를 활용한 프로토콜로 설계되었다. 조정에이전트는 응용에이전트와 통신하여 필요한 작업을 제어하고 등록된 응용에이전트에 대한 상태관리, 통신 제어와 검색 결과 통합 및 관리 등을 수행한다. 조정에이전트의 에이전트 관리기는 조정에이전트에 포함된 스케쥴 관리기와 이벤트관리기, 보안관리기, 에이전트 조정관리기를 제어하는 모듈이다.

에이전트 조정 관리기는 응용에이전트와의 분산작업을 수행하면서 발생하는 이벤트들을 처리하는 모듈로서 통신기의 제어 및 응용에이전트에 필요한 작업의 요청 및 요청에 대한 결과를 처리하고 메세지의 생성 및 검색결과를 통합하는 메시지관리기를 제어한다.

보안 관리기는 에이전트간에 주고받는 메시지(메타정보, 제어명령 등) 의 신뢰성을 부여하기 위해 응용에이전트들에 고유 키(ID)를 생성하여 제공한다. 스케쥴 매니저는 각 응용에이전트의 통신 및 검색 요청, 검색결과 반환 등의 동작 스케쥴을 관리하며, 또한 본 시스템에서 모듈간의 메시지 및 데이터 처리에 쓰이는 메모리와 버퍼 등도 함께 관리한다. 이벤트 관리기는 시스템 운용 상에서 발생하는 모든 이벤트 및 로그를 관리한다.

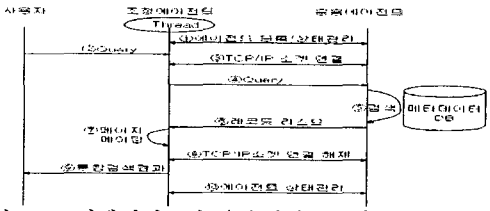
응용에이전트는 조정에이전트에서 전송된 질의어를 수신하며 메타데이터 DB 에 대하여 검색을 수행하고 검색결과를 조정에이전트에 반환하는 기능을 수행한다. 응용에이전트 조정 관리기는 응용에이전트 내의 보안 관리기, 액티베이터 및 메시지 관리기를 제어하는 기능을 수행한다.

보안 관리기는 에이전트간에 주고받는 메시지(메타정보, 제어명령 등)의 신뢰성을 부여하기 위해 응용에이전트들에 고유 키(ID)를 생성하여 제공한다. 액티베이터는 메타데이터 DB 에 접근하여 요청한 정보를 검색하는 인터페이스이다.

메시지관리기는 에이전트간에 주고 받는 메시지를 처리한다.

3.2 통신 프로토콜 설계

본 시스템에서 조정에이전트와 응용에이전트 간의 통신 프로토콜은 HTTP 를 기반으로 구축되었다. 두 에이전트 간의 통신에서 질의어 및 검색 결과는 OAI-MHP 버전 1.0 의 레코드를 응용하여 설계하였다. (그림 3)은 사용자의 질의어 전송에서부터 통합검색 결과를 사용자에게 제공하는데 까지의 통신 프로토콜 흐름도를 보인다. 또한 (그림 4)은 전송되는 검색결과의 레코드 리스트 예를 보인다.



(그림 3) 조정에이전트와 응용에이전트의 통신 과정

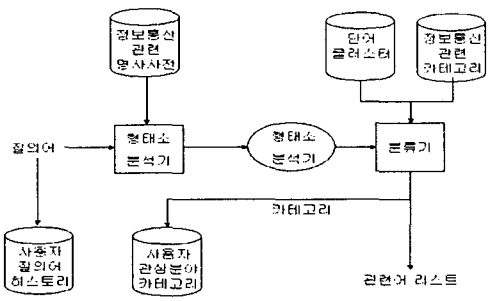
```
<?xml version="1.0" encoding="UTF-8"?>
<GetRecord>
<header>
   <identifier>etri/ejournal</identifier>
   <datestamp>2002-09-08</datestamp>
</header>
<metadata> <dc>
   <title>CDMA</title>
   <creator>jc song</creator>
   <date>2001-12-01</date>
   <identifier>http://www.etri.re.kr/ejournal/cdma
   </identifier>
   </dc>
</metadata>
</GetRecord>
```

(그림 4) 레코드 리스트의 예

3.3 관련어 추출기 설계

본 시스템의 관련어 추출기는 사용자가 문장 단위의 질의를 입력할 경우 명사를 추출하여 분산검색을 수행할 수 있도록 구현되었다. 또한 추출된 명사를 이용하여 기구축된 단어클러스터에서 관련 단어들을 도출하고 도출된 관련어 리스트는 카테고리를 참조하여 분류기에서 관련 분야를 추출하게 된다. 추출된 카테고리는 사용자 관심분야 카테고리에 저장하고 카테고

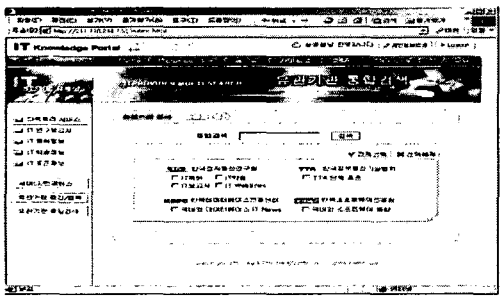
라의 키워드들을 관련어 리스트로 작성하여 조정에이전트에 활용할 수 있도록 한다. 본 관련어 추출기의 명사사전과 단어 클러스터, 카테고리는 정보통신분야에 한정하여 구축되었다. 본 관련어 추출기는 관련 카테고리 정보를 이용하므로 분야별 검색에서 검색 성능을 높일 수 있다.



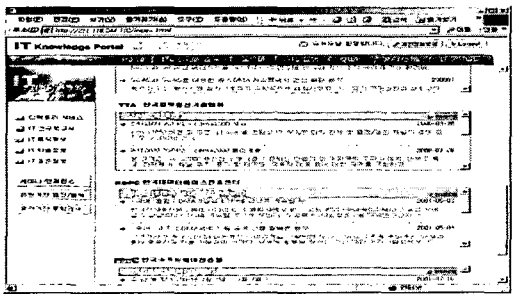
(그림 5) 검색어와 관련된 카테고리와 관련어 추출기

3.4 실험

본 시스템을 설계 및 구현하여 정보통신 유관기관 중 네 개 기관에 대하여 본 시스템을 실험 하였다. 조정에이전트는 사용자가 입력한 질의어를 조정에이전트에 등록된 각 기관의 응용에이전트에 전송하고 검색 결과를 응용에이전트로 부터 반환받는다. 조정에이전트에 전송된 검색결과는 조정에이전트에서 페이지 메이킹을 거쳐 사용자에게 전송된다. (그림 6)과 (그림 7)은 실험 결과 화면을 보인다.



(그림 6) 검색어 입력 화면



(그림 7) 분산검색 결과 화면

4. 결론

본 논문에서는 XML 및 OAI-MHP 를 응용하여 통신하는 멀티에이전트를 기반으로 하고 메타데이터를 대상으로 분산검색을 수행하는 시스템을 설계 및 구축하였다.

본 시스템에서 에이전트간 통신은 HTTP 기반의 소켓을 사용하였으며 질의어 및 검색 결과 전송은 XML 및 OAI-MHP 를 응용하여 설계하였다. 또한 메타데이터 DB 에 대한 검색을 수행하여 분산검색을 수행하는데 있어 검색 속도를 향상시킬 수 있었으며 검색결과에 대한 정확성도 높일 수 있었다.

본 시스템을 발전시키기 위해 향후에는 에이전트간 통신 및 운영에 대한 안전성을 보다 더 확보하기 위해 에이전트간의 자동 등록 및 상태 관리 기능에 대해 연구가 보다 더 필요하며 OAI-MHP 버전 2.0 을 지원하도록 확장하여야 한다.

참고문헌

[1] Open Archives Initiative, http://www.openarchives.org/
[2] World Wide Web Consortium, http://www.w3.org/
[3] Dubline Core Metadata Initiative, http://www.dublinecore.org/
[4] MultiAgent Systems, http://www.multiagent.com/
[5] A DL Server with OAI Capabilities:LOVE, Su-Shing Chen, JCDL '02,July 13-17, 2002, pp388
[6] JAFER ToolKit Project-Interfacing Z39.50 and XML, Antony Corfield, JCDL'02, July 13-17, 2002, pp289-290