학술지목차DB(QTOC)를 활용한 해외학술정보 수집에이전트 시스템

신성수°, 노경란, 권오진, 홍성화
한국과학기술정보연구원 정보콘텐트개발실
e-mail:{kolatree, infor, dbajin, shong}@kisti.re.kr

Agent-Based Gathering System For Foreign Scientific Information Using QTOC

Sung-Su Shin°, Kyung-Ran Noh, Oh-Jin Kwon, Sung-Hwa Hong
S&T Contents Development Dept, KISTI

요 약

인터넷과 정보통신기술의 급속적인 발전으로 수천 또는 수억에 달하는 방대하고 다양한 정보들이 웹상에서 존재하게 되었다. 웹상에서 획득가능한 학술정보가 증가함에 따라 다량의 정보를 효율적으로 수집하기 위하여 다양한 웹 로봇기반 수집에이전트를 활용하고 있다. 본 논문에서는 웹에 산재되어 있는 해외학술정보를 체계적이고 주기적으로 수집, 분류하기 위하여 학술지목차DB(QTOC)를 활용한 해외학술정보 수집에이전트 시스템을 설계하였다.

1. 서론

인터넷과 정보기술의 급속한 발전으로 웹은 방대한 양의 학술정보를 포함하게 되었다. 이러한 웹상에 존재하는 다양한 정보를 효율적으로 검색, 추출, 분류하기 위한 다양한 방법 등이 제시되었다. 웹에 존재하는 학술문서를 효율적으로 수집하기 위하여 웹 로봇 에이전트를 사용한다. 그리고 웹상에서 수집한 학술문서를 효율적으로 분류하여 사용자에게 적합한 정보를 신속하게 제공하기 위하여 인터넷 디렉토리 시스템을 적용하는 방안이 연구되었다[2,4].

이러한 웹 로봇 에이전트는 초기화된 URL정보를 바탕으로 HTTP상에 존재하는 웹 서버의 문서위치를 파악하고 수집, 분석한다. 그리고 문서에 연결된 다른 문서를 추출하는 방식으로 동작한다. 인터넷상에서 존재하는 다양한 학술자원을 추출하기 위한 많은 웹 로봇 에이전트들의 동작은 웹 서버 및 네트워크의 과부하를 가중시킬 수 있다. 또한 수집된 정보에 대한 정확한 분류가 이루어지지 않을 경우, 사용자에게 불필요한 데이터를 가중시키는 결과를 가져올 수 있다. 이러한 문제점들을 해결하기 위하여 특정 전문분야의 웹 문서들을 수집하기 위한 방법들이 제시되었다[2].

본 논문에서는 KISTI의 학술지목차DB를 활용하여 과학·기술분야에 관련되는 학술정보를 효율적으로 수집·분류하는 학술정보 수집에이전트 시스템을 설계한다. 웹로봇 에이전트는 서버의 과부하를 최소화하기 위해, 동작하게 될 웹 서버의 위치와 관련 웹 문서에 대한 정보 및 설정된 수집범위에 따라 동작한다. 수집된 학술정보에 대하여 KISTI 과학기술문헌 분류표를 적용하여 체계적으로 문서를 분류한다.

본 논문의 구성은 다음과 같다. 2장에서는 웹 로봇 에이전트의 개념 및 동작, 그리고 제안된 에이전트 시스템의 기반으로 사용되는 학술지목차DB에 대하여 기술한다. 3장에서는 시스템의 구성과 동작에 대해 기술한다. 4장에서는 구현된 시스템의 동작결과에 대해 기술하고, 마지막 5장에서는 결론 및 향후 연구과제에 대하여 기술한다.

2. 관련연구

본 장에서는 웹 로봇 에이전트의 개념 및 동작에 대해, 그리고 에이전트의 초기 프로파일을 제공할 학술지목차 DB에 대하여 기술한다.

2.1 웹 로봇 에이전트 (Web Robot Agent)

웹 로봇 에이전트는 웹 크롤러, 웹 스파이더 등으로 불려지며, 웹 서버를 순회하며 각 홈페이지에 있는 수많은 정보를 수집하는 프로그램이라 할 수 있다. 분산된 웹 환경에서 이질적인 시스템 상에서 존재하지 않고 하나의 웹 서버에서 존재하는 웹 문서를 수집하기 위하여 초기화된 URL을 바탕으로 한다. 웹 문서를 획득하여 문서내에 포함된 데이터를 분석한 후 가용 데이터를 추출하는 기능을 수행한다. 또한, 웹 문서내에 포함된 다른 문서에 대한 하이퍼링크를 바탕으로 수집영역을 확대한다. 이렇게 수집된 문서가 체계적으로 분류되고 색인되어 이용자가 검색을 수행할 때 보다 정확한 학술정보를 얻을 수 있도록 한다. Web상에서 존재하는 다양한 학술자원, 미디어, 이미지, 문서 등의 추출을 목적으로 하는 많은 웹 로봇 에이전트들의 동작은 웹 서버 및 네트워크의 과부하를 가중시킬 수 있다. 그리고 수집된 정보에 대한 정확한 분류가 이루어지지 않을 경우, 불필요한 데이터를 검색결과로 가져올 수 있다[1,5].

2.2 학술지목차DB (QTOC : Quick Table Of Content)

한국과학기술정보연구원에서 가공, 서비스하고 있는 학술지목차DB(QTOC)는 EBSCO로부터 도입하는 해외학술지목차와 직접 제작하는 일문학술지목차로 구성된다.

QTOC는 ISSN을 키값으로 학술지명, 발행국, 자료유형, 각 학술지의 주제분야에 대한 정보를 학술지목록DB로부터 가져와 저장하고 있다[3]. QTOC의 제작과정은 그림 1과 같다.

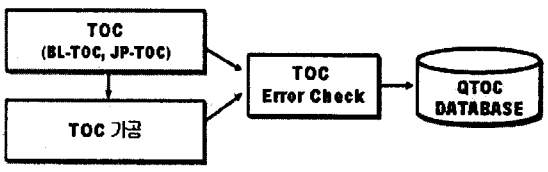


그림 1 학술지목차DB 제작과정

3. 시스템설계

본장에서는 학술정보수집 에이전트시스템의 각 구성모듈에 대해 설명하고, 에이전트 시스템의 동작에 대하여 기술한다.

3.1 시스템구성

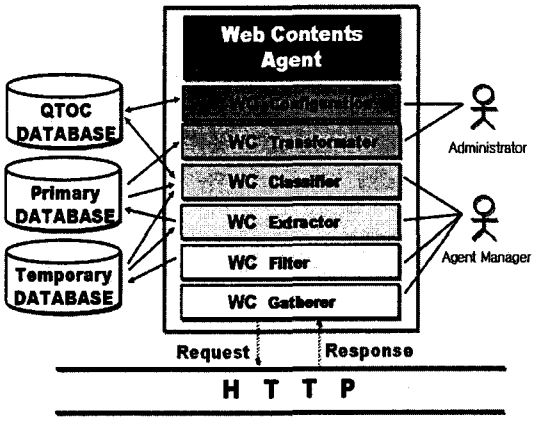


그림 2 시스템 구성

● Configurations Module

각 에이전트에 대한 프로파일을 설정한다. 본 논문에서 제안하는 에이전트의 프로파일은 접속할 서버의 URL과 인증정보, 수행주기, 수집된 문서에 포함된 링크정보의 탐색순위, 연관된 정보를 수집하기 위하여 에이전트에 의하여 생성될 하위 에이전트의 생성 및 동작, 생명주기에 관한 제약사항들을 수록하여 설정하고 있다.

● Transformation Module

수집된 후 분류가 완료된 데이터에 데이터변환을 수행한다. 수집된 데이터를 내부 데이터베이스 Format으로 변환한 후 저장하고 인덱싱한다.

● Classifier Module

수집된 데이터에 대하여 분류작업을 수행한다. 각 에이전트가 주제분야별로 데이터를 수집하므로, 대상 학술지는 KISTI의 과학기술분류표에 의하여 1차적으로 분류된다. 이렇게 분류된 데이터는 각 학술지의 과학기술분류코드에 해당하는 용어사전을 적용하여 키워드를 추출하고 키워드에 대한 가중치 부여방식을 통하여 수집된 데이터에 대하여 2차 분류작업을 수행한다.

● Extractor Module

수집된 데이터로부터 DB구축에 필요한 가용자원

(제목, 저자, 초록, 페이지)들을 추출한다. 이러한 원시데이터를 추출하기 위하여 정의된 각 학술지 패턴정보를 이용한다. 학술지패턴은 에이전트가 추출하게 될 대상 저널의 웹 페이지를 파스(Parse)하여 가용데이터의 위치와 데이터간의 관계를 기술함으로써 정의되고 학습된다.

● Filtering Module

데이터에서 가용데이터를 분류하는 기능을 수행한다. 수집된 데이터가 해당 주제분야에 속하지 않거나 패턴에 일치하지 않거나, 수집대상 데이터가 아닌 경우 불용문서로 판단하여 이에 대한 정보를 해당 에이전트의 불용문서리스트에 저장한다.

● Gatherer Module

에이전트에 의하여 요청된 서버에 대하여 실질적인 요청을 수행하고 서버로부터 응답된 결과를 반환한다. 이 과정에서 어떠한 처리과정도 이루어지지 않으며 수신된 문서에 대하여 저장된 문서에 대한 크기와 수정일자에 대한 HEAD정보를 비교하여 Filtering 모듈로 전송된다.

3.2 시스템 동작

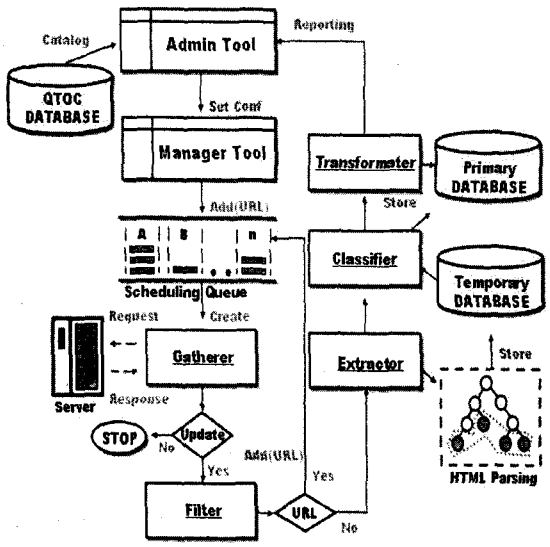


그림 3 수집에이전트 시스템 동작

그림 3에서는 본 논문에서 제안하는 해외학술정보 수집에이전트의 수행과정을 보여준다. 우선 각 주제분야에 해당하는 수집에이전트는 초기화된 URL을 기반으로 서버에 접속하여 웹 문서를 요청한다. Filter모듈에서는 서버로부터 응답된 웹 문서에서 가용데이터를 추출하기 위하여 불필요한 데이터를 제거한다. 이러한 데이터의 제거는 에이전트 관리자에 등록된 각 학술지에 대한 문서패턴에 의하여 가용데이터를 필터링하는 것이다. 그리고 문서내에 포함된 URL정보를 바탕으로 새로운 수집에이전트를 생성하게 된다. 데이터 필터링 이후, Extractor 모듈에서 실질적인 데이터추출작업이 이루어진다. 추출된 데이터는 임시데이터베이스 저장되며, 저장된 데이터에 대하여 분류작업이 수행된다. 데이터에 대한 분류작업은 각 주제분야에 해당되는 주제분야코드를 적용하여 분류하고 색인하게 된다. 수집된 데이터에 정보와 에이전트의 수집활동에 대한 시각적인 정보를 관리자가 확인할 수 있도록 각 에이전트의 행위정보를 제공하고 활동을 종료한다.

4. 시스템 구현

본 논문에서 제시하는 학술정보수집 에이전트 시스템의 환경설정과 수행결과는 다음과 같다.

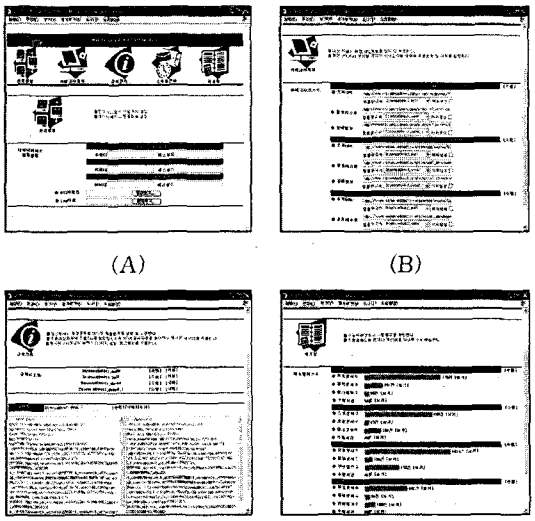
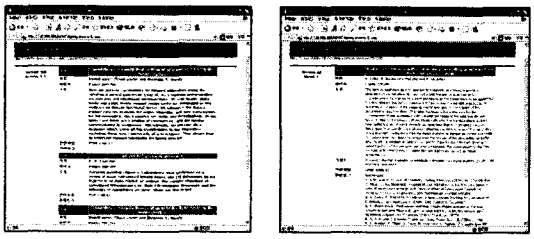


그림 4 시스템관리자 화면

그림 4의 (A)는 본 논문에서 구현한 시스템에서 연결된 각 데이터베이스시스템의 로그인 정보를 설정한다. 또한, 주기적으로 갱신되는 QTOC파일의 입력과 수행 과정에서 발생되는 예외적인 상황들을 저장할 로그파일을 설정하는 화면이다. (B)에서는 각 해당저널의 URL정보와 해당저널의 웹 문서의 링크정보를 파악하여 실제적인 가용데이터를 추출하게 될 URL을 설정하는 화면이다. (C)는 HTML을 분석

하여 구조적인 패턴을 추출하며 데이터에 대한 추출 규칙을 설정하는 화면이다. (D)는 웹 문서에 대한 수집활동이 완료된 수집에이전트의 수집결과를 보여주는 화면이다.



(A) (B)

그림 5 수집된 데이터의 결과

그림 5에서는 해외학술정보 수집에이전트에 의하여 수집된 데이터에 대한 사용자화면을 보여준다. (A)는 수집대상 저널에 수록된 각각의 제목과 초록 데이터, 링크된 원본파일에 대한 정보를 보여주며, (B)는 (A)에서보다 상세한 데이터가 수집되어 저자가 작성한 키워드와 링크정보가 유지된 참고문헌에 대한 정보를 보여준다.

5.결론 및 향후과제

웹의 방대한 정보를 효율적으로 수집하기 위하여 웹 로봇 에이전트를 활용한 다양한 시스템들이 제시되고있다. 본 논문에서는 웹에서 획득가능한 학술정보를 수집하기 위하여 웹 로봇 에이전트기반의 해외학술정보 수집에이전트를 설계하고 구현하였다.

수집에이전트는 정의된 수집영역 내에서 연관된 정보를 추출하고 링크정보를 선별하여 제거함으로써 수집에이전트로 인한 서버의 부하를 최소화할 수 있도록 하였다. 각 주제분야별 수집에이전트가 학술지 목차DB를 기준으로 데이터를 수집함으로써 학술정보의 연속성을 유지할 수 있도록 하였다. 수집된 데이터는 과학기술분류표에 의하여 분류되고 색인됨으로서 정보검색효율을 높여준다. 그리고 주기적인 수집을 통하여 동일한 주제분야의 다양한 정보들을 제공하여 사용자에게 양질의 학술정보를 제공할 수 있게된다.

향후 사용자의 주제분야별 프로파일을 적용하여 개개의 사용자가 요구하는 학술정보를 제공할 수 있도록 하며 현재 각 주제분야별 전문가에 의하여 세분화된 분류작업을 통하여 서비스되고 있는 해외학술정보의 분류패턴을 분석하여 수집에이전트에 의한 자동분류가 가능토록 하여야 한다.

[참고문헌]

[1] 김동범, 2002. "웹 로봇 에이전트의 효율적인 인터넷 정보검색", 한국정보과학회 학술발표논문집 29 (2), pp.574-576.

[2] 김상경, 2001. "인터넷 웹에서의 특정 분야의 전문 지식 획득", 한국정보과학회 학술발표논문집 28(1). pp.346-348.

[3] 노경란, 2003. "가용자원을 활용한 해외학술정보 데이터베이스제작방법에 관한 연구", 한국콘텐츠학회 학술발표논문집 1(1), pp.323-326.

[4] 노영희, 2001. "기계학습을 기반으로 한 인터넷 학술문서의 효과적 자동분류에 관한 연구", 한국도서관 · 정보학회지 32(3), pp.307-330.

[5] 서희경, 2002. "준구조화된 정보소스에 대한 지식기반의 Wrapper 학습 에이전트", 정보과학회지 29(1), pp.42-52.

[6] 최중민, 2000. "인터넷 정보 추출 에이전트", 정보과학회지 18(5), pp.48-53.

[7] 윤구호, 2002. "웹 색인작성에 관한 연구", 한국도서관 · 정보학회지 33(2), pp.235-258.

[8] 최재황, 1998. "인터넷 학술정보자원의 디렉토리 서비스 설계에 있어서 DDC 분류체계의 활용에 관한 연구", 정보관리학회지 15(2)

[9] Caglayan, A. and Harrison C., 1997. "Agent," Wiley computer Publishing

[10] Keiichiro Hoashi, 2001. "Document Filtering Method using Non-Relevant Information Profile," Proc, Of the 23th annual international ACM-SIGI,

[11] Mark Craven, 1998. "Learning to Extract Know- ledge from the World Wide Web", Proc. of the international Workshop on AAAI

[12] Susan Dumais, 2000. "Hierarchical Classification of Web Content", Proc of the 23th annual international ACM-SIGIR.

[13] Won-Kyun Joo, 1998. "Improving Retrieval Effectiveness with Link Information", Proc. of the International Workshop on IRAL