디지털 도서관을 위한 로봇 에이전트의 설계 및 구현

주원균*, 김용**, 조은일*, 맹성현*

충남대학교 컴퓨터과학과*, 한국통신 연구개발본부 기술조사팀**

A Design and Implementation of Robot Agent in a Digital Library

Won-Kyun Joo, Yong Kim, Eun-Il Cho, Sung-Hyoun Myaeng

Dept. of Computer Science, Chungnam National Univ.*, Korea Telecom R&D Group**

요 약

인터넷이 그 규모 및 정보의 내용에서의 대형화를 지향하면서, 분산환경에서 사용자들의 정보에 대한 요구가 날로 증가하고 있고, 이를 만족시킬 수 있는 전 단계로 정보 수집이라는 것이 새로운 문제로 떠오르고 있다. 이에 에이전트를 기반으로 구성된 분산 디지털 도서관상에서 웹 검보 서비스를 제공하기 위한 로봇 에이전트를 설계·구현하였다. 구현된 로봇은 JAVA 와 RMI 를 기반으로 에이전트화를 통해 분산 환경 구조를 갖추고 있으며, 멀티 쓰레드와 로봇 제어 알고리즘을 통해 정보의 수집 및 관리에 있어서 로봇의 신뢰도를 향상시켰고, DBMS 와 에이전트의 연계를 통한 대용량 처리 기능과 관리자의 관리의 편리성을 도모하기 위해 자동화 기능을 갖추었으며, 로봇과 관련되어 발생할 수 있는 부하를 최소화 할 수 있도록 하였다

1 서론

인터넷 환경에서의 사용자의 검보에 대한 요구가 증가하면서, 사용자가 원하는 정보의 보다 빠르고 정확한 서비스를 위한 정보수집이 중요한 문제가 되고 있다. 정보수집을 위해서는 로봇이라는 것을 사용하는데, 이것은 1993 년에 매튜그레이에 의해서 인터넷에 탄생이 되었고, 단순히 주어진 호스트를 시작점으로 링크를 따라 문서를 수집하는 역할을 하였다. 단순한 로봇을 시작으로 하여 인터넷에는 많은 로봇들이 탄생하였고, 국내에도 공식적인 로봇들이 이십여개 정도 있는 것으로 알려져 있다.

그러나 시대의 흐름과 함께, 지금의 대용량 웹 환경에서는 문서와 네트웍의 규모에서 큰 차이가 있고, 이와 관련하여 다양한 문제들이 파생될 수 있기에, 예전과 같은 단순 구조의 로봇으로 이런 일을 처리하기에는 부적합하다. 최근에 국내에서는 변화된 웹 환경에서의 로봇의 구현에 대한 연구[3,4]가 있었고, 국외에서도 독립 모듈을 기반으로 한 HARVEST[1] 에이전트를 기반으로 한 HARNESS[3] 와 같은 시스템에서 좀더 발전된 형태의 정보 수집의 가능성을 보여주고 있다. 변화하는 웹 추세와 발맞추어가기 위해서 로봇은 제각기 나름대로의 일을 처리 할 수 있는 에이전트화가 되어야 하며 이와 더불어 네트웍과 문서에 크기에 무관한 로봇의 신뢰도 향상 대용량 처리를 충족시켜줄 수 있어야 한다.

본 논문에서는 분산 디지털도서관[6]상에서 웹 정보 서비스를 가능하게 하기 위해서 에이전트화 된 로봇을 설계·구현하였고, 로봇의 신뢰도의 대용량 처리 능력을 극대화 시켰다. 2 장에서는 로봇 에이전트가 갖추어야 할 요소들을 현재의 웹 환경과 관련 지어 언급하고, 3 장에서는 우리가 구현한 로봇 에이전트의 디지털 도서관상에서의 작동 모습에 대해서 설명하고, 4 장에서는 로봇에이전트의 구현 및 작동시나리오를 보이며, 5 장에서 결론을 맺는다.

2 로봇이 갖추어야 할 요건

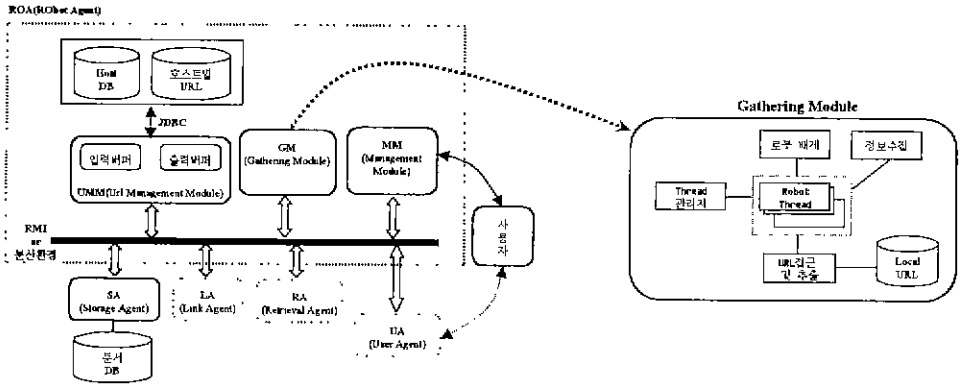
현재의 웹 추세와 분산 디지털 도서관과의 관계를 고려해 볼 때, 로봇은 다음과 같은 요건을 갖추어야 한다.

■ 에이전트화

현재와 같은 대용량/대기능을 처리하기 위해서는 각각의 특립된 기능들이 에이전트로 나뉘어지 서로 자기의 전문 분야를 처리할 수 있도록 하여야 한다. 이렇게 함으로서 시스템과 네트웍 자원을 많이 필요로 하는 로봇의 기능을 분산 서키고 지원을 효과적으로 사용함으로써 로봇의 성능을 높일 수 있다. 또한 새로운 기능의 추가나 문제 발생시 다른 환경에 영향을 주지 않고 에이전트만을 교체할 수 있다. 에이전트화 됨으로써 얻는 가장 큰 장점은 정보의 공유이다. 고전적인 로봇들은 자기 혼자 동작하고, 자원에 대한 효율적인 접근 방법이 없었으나 HARVEST 와 같은 수집저들에서는 구조적인 접근 및 SOIF 라는 검보 공유 포맷을 통해 보다 진보된 형태를 지향한다. 에이전트화를 위해서는 전형적인 미들웨어인 CORBA 나 JAVA 에서 지원하는 RMI, 또는 전용 에이전트 환경인 KQML[2]이 사용될 수 있다.

■ 신뢰도 향상

본 논문에서 제기하는 신뢰도는 수행속도의 극대화, 유용한 문서의 수집, 로봇의 오동작의 배제이며, 수행속도를 최대화하면서 시스템에 적은 부하를 주기 위해서는 로봇의 병렬화, 쓰레드 화, 에이전



[그림 1] 분산 디지털 도서관상에서의 로봇

트화로 설계가 되어야 한다. 또한 웹 문서는 단순 HTML, CGI등 다양한 형식으로 구성되며, 수집결과 중복되거나, 이미지 맵과 같이 그 자체로는 불필요한 문서가 많이 존재하고, 호스트와 문서의 깊이에 따라 문서의 유용성이 많이 달리지며, 이질적인 네트웍과 웹 메모을 통한 문서 수집 과정 상에서 병렬화나 로봇의 방문정책에 따른 오류가 많이 발생한다는 점도 감안해야 한다.

■ 대용량 처리

적은 규모의 웹 환경에서는, 몇 개의 시작 URL(우리나라를 대상으로 할 경우에는 한국 내 호스트 리스트)만을 이용한 단순 깊이 방문만을 통해서도 성공적으로 웹 문서를 얻어올 수 있었다. 이런 환경하에서는 특별히 URL 자원 관리, 로봇 제어와 같은 개념이 필요치 않았다. 그러나 문서는 나날이 증가하고 있는 환경에서 URL 자원의 관리 없이 문서를 기져 오는 것은 로봇의 처리능력 감소와 더불어 데이터의 신뢰도를 떨어뜨리고, 네트웍의 부하만을 가중시키게 된다. 실제로 InterGate의 조사[7]에 의하면 국내에서 제작된 로봇별 문서 수집능력의 확연한 차이를 볼 수 있는데, 로봇에 따라서는 4배정도의 차이를 보이기도 한다.

대용량 처리를 위해서는 도메인이나 호스트 별에 따른 URL의 관리 방법과 이를 저장하기 위한 건문 DB와의 연계가 필요하며. 갱신의 측면에서 새로 추기된 문서에 대한 확인, 기존 문서의 갱신의 가능성 및 정도의 파악에 대한 방법이 필요하다.

■ 자동화

단순 방문을 이용한 정보 수집 방식들에서는 로봇 시작의 명시, 얻어온 문서와 URL 자원에 대한 처리, 로봇 작동 중 발생하는 문제로 인해 관리자의 많은 개입이 필요했다. 편리함과 유용성으로 인해 널리 확산되어 가는 웹을 이용해 인터페이스를 설계한다면, 관리의 편리성을 도모함과 동시에 시간적인 측면에서도 많은 이득을 볼 수 있다. 또한 웹 인터페이스를 이용하면 로봇에게 빠르고 쉽게 많은 정보를 알려 줄 수 있어 로봇의 자율화 측면이 중대되고, 수행이 완료되었을 경우에는 메일 등을 통해서 수행의 결과를 관리자에게 보고할 수 있으며, 관리자는 로봇의 작동 중에 수집 상태를 한눈에 파악할 수도 있다.

■ 로봇 부하의 최소화

정보 서비스를 위해서 로봇의 역할은 매우 중요하지만, 로봇 작동 중에 로컬 호스트, 네트웍, 접근호스트에 대한 부하는 심각한 문제가 되고 있다. 이에 대한 해결책으로 여러 방법이 제시되었는데, 로컬 호스트에 대한 문제는 로봇 서버의 성능을 향상시킴으로써 해결하고 있고, 네트웍에 대한 부하를 줄이기 위해서는 로봇의 문서 요청 주기를 조절(하비스트, 분당 1건)함으로서 해결하거나 정보 공유 형식(SOIF 등)의 제안[1] 혹은 사용자가 없는 시간대(저녁부터 새벽, 주말, 공휴일)에 로봇을 작동[4]시키는 방법(한국내 도메인과 같이 특정 도메인에 한정시켰을 경우에만 가능하다는 단점이 있다)으로 가능하다. 접근 호스트에 대한 부하(rapid-fire)를 해결하기 위해서는 BFS의 변형된 형태인 도메인별 처리(까치네), 메타 검색엔진을 이용[5]하는 방법이 있다.

3 로봇 에이전트의 설계

본 논문에서 설계한 로봇은 디지털 도서관에서 웹 서비스를 지원하기 위한 것으로, 2장에서 언급한 로봇의 요건을 갖추고 있으며, 에이전트 기반 분산 디지털 도서관에서 정의한 여러 에이전트들과 상호 협조하도록 설계하였다. <그림 1>의 왼쪽은 로봇에이전트의 구조 및 디지털 도서관상의 각 에이전트들과의 관계를 보여준다.

3.1 분산 디지털 도서관과의 관계

제안된 에이전트 기반 분산 디지털 도서관[6]상에서는 UAP(User Application Program), UA(User Agent), RA(Retrieval Agent), LA(Link Agent), SA(Storage Agent)의 다섯 요소가 정의 되어 있는데, UAP는 사용자 응용 프로그램으로 일종의 인터페이스 역할을 하며, 사용자 에이전트인 UA는 사용자를 대신한 검색 및 재랭킹을 담당하고, 검색 에이전트인 RA는 검색을 위한 색인정보를 바탕으로 검색을 담당하며, 링크 에이전트인 LA는 재랭킹[8]및 브라우징을 위한 링크 정보의 추출과 링크·문서의 분리 저장을 담당하며, 저장 에이전트인 SA는 원문의 저장을 담당한다. 이들간의 관계를 보면, SA는 RA와 LA에게 원문을 제공하여 각각 색인 정보의 생성과 링크 정보의 분리 저장을 가능하게 하고, UA는 UAP의 요구에 따라 RA 및 LA를 사용하여 검색/재랭킹/브라우징을 하게 된다.

<그림 1>의 왼쪽의 가로로 된 굵은 선은 에이전트들의 작동 및 통신을 위한 분산 환경을 나타내는데, 본 논문에서는 CORBA보다는 규모가 작으면서 효율적으로 지바/웹과 연계 시킬 수 있는 RMI 환경을 개발하였다. RMI환경을 중심으로 사용자, 로봇에이전트(ROA), 분산 디지털 도서관을 위한 에이전트(SA, LA, RA, UA)가 산재해 있다. 이 환경에서 사용자의 요구에 따라 로봇에이전트가 작동하여 내부적으로 결과를 처리하면서, SA에게로 문서를 전송한다. 전송된 문서는 디지털 도서관의 에이전트의 역할에 따라 처리되고, 사용자의 검색 및 브라우징이 가능하게 된다.

3.2. 로봇의 주요 모듈

■ UMM(Url Management Module)

UMM은 로봇의 방문 대상이 되는 URL의 획득, 저장, 관리를 담당하는 모듈로서 외부의 데이타베이스와 JDBC를 통해 연결되며 로봇 에이전트중 수집을 담당하는 GM에서 방문할 URL이나 특정 URL의 저장을 요청할 경우에 외부 데이터베이스와의 적절한 통신

을 통해 요구에 응답한다.

UMM은 관리를 담당하는 관리 모듈인 MM으로부터 로봇의 방문 시작 포인터가 될 수 있는 URL을 얻어, 초기화시에 HOST DB에 입력하게 된다 모든 GM들은 UMM을 통해서만 방문 URL을 얻을 수 있으며, 방문 후 찾아 낸 호스트와 상태정보(URL,방문 깊이, 문서의 최종 수정일, 문서의 방문 날짜)는 각각 Host DB와 Host별 URL DB에 입력 또는 갱신된다.

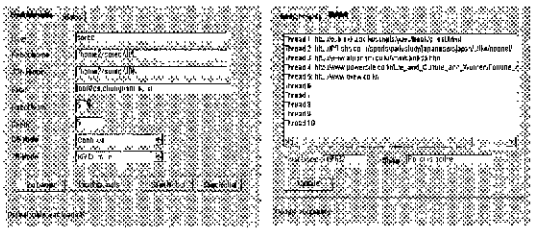
로봇은 주어진 호스트를 방문하면서 빈번한 URL요청, 수정을 요구하기에 시스템의 성능을 고려하여 일종의 캐시 역할을 할 수 있는 입력버퍼/출력버퍼를 두었다. DB측면에서는 방문할 호스트 정보를 얻기 위한 Host DB와 업데이트를 위한 호스트별 방문 URL 정보를 저장할 수 있는 호스트별 URL DB를 두고 있다

■ GM(Gathering Module)

GM은 직접적으로 문서 수집을 하는 모듈로서 UMM으로부터 얻은 호스트 정보를 바탕으로 멀티 쓰레드를 이용하여 정보를 수집하고 저장하는데, 자세한 내용은 <그림 1>의 오른쪽 부분에 나와 있다 GM을 보면 각 기능별로 Thread 관리자, Robot Thread, 로봇배제, 정보 수집, URL 접근 및 추출, Local URL DB로 구성이 된다 Thread 관리자는 Robot Thread의 관리 및 제어를 담당하며, 생성/제거/상태확인을 통해서 Robot Thread의 문제 발생 여부를 확인/복구한다 Robot Thread는 실제의 수행단위로 정보수집/로봇배제/URL 접근 및 추출의 기능을 가지고 있는데, 문서를 수집하여 필요한 정보를 추출한 후 URL 정보는 DB에 문서 내용은 SA로 전송한다. 로봇 배제는 특정 로봇의 특정 URL로의 접근을 막기위한 웹서버 관리자의 의도를 반영하고자 하는 것이며 정보 수집 부분은 실제 웹 문서를 가져오고 필요한 내용을 추출하는 기능을 가지고 있으며, URL 접근 및 추출은 Local DB를 통해 Robot Thread가 방문할 URL의 제공 및 저장의 역할을 맡고 있다.

■ MM(Management Module)

MM은 관리 모듈로서 사용자가 로봇이 제공하는 기능을 효과적으로 이용하여 문서 수집을 가능하게 하고, 로봇 사용 시 로봇의 여러 상태 및 제어를 가능하게 하는 부분으로 사용자 인터페이스라 할 수 있다 현재 인터페이스상에서 제공되는 것들은 로봇의 환경설정/수행/상태확인/종료이며, 로봇의 환경 설정에서는 로봇 수행 전에 앞서 로봇 쓰레드의 개수, 방문 깊이, 방문 정책(주어진 URL/현 호스트/KR 도메인/모든 도메인)을 조정함으로써 사용자의 요구에 맞는 문서 수집을 가능하게 하며, 로봇의 상태 확인 기능은 로봇의 수행 상태(각 쓰레드별 문서 수집상태, 전체 문서 수집상태)를 확인하여 사용자가 능동적으로 대처할 수 있도록 해준다



[그림 2] 로봇 관리를 위한 인터페이스

4. 구현 및 작동 시나리오

본 논문에서 제안한 로봇은 분산환경을 기반으로 구축된 디지털 도서관의 각 에이전트들과 원활하게 작동될 수 있어야 한다 또한 로봇 지체의 효율성의 증대. 로봇은 웹과 밀접한 관계가 있기에 웹과의 관계를 고려해야 한다 종합적으로 볼 때 RMI라는 분산 환경을 지원하고 멀티 쓰레드와 웹을 지원하는 JAVA가 가장 적합하다고 판단되어, 사용언어로는 JAVA를 분산환경으로는 JAVA에서 지원하는 RMI를 채택하였다. 이렇게 함으로서 JAVA와 웹에서 지원하는 기능을 모두 로봇에 적용시킬 수 있고, 웹을 통한 인터페이스(그림 2)로 관리의 편리성 측면을 증진시킬 수 있다

다음은 로봇이 작동하는 과정에 대한 시나리오이다.

① 사용자 인터페이스를 통해 로봇의 각종 설정을 한다(처음 작동 시에는 미리 구축된 초기 URL 리스트를 이용하여 UMM의 DB를 초기화 시킨다)
② GM이 UMM과 SA를 이용하기 위한 초기 작업을 수행한다
③ Thread 관리자가 수행되어 Robot thread의 상태를 확인하면서 특정 개수의 유효 쓰레드를 유지시키고, 각 쓰레드는 다음을 수행한다
   A. UMM으로부터 호스트 정보를 가져온다
   B. Robot thread가 하나 작동되어 한 호스트에 대한 내용을 방문하면서 문서 정보는 SA로 새로 발견한 호스트 정보는 UMM으로, url 정보는 Local URL DB에 저장한다.
   C. 모든 방문이 끝나면 Local URL 정보 중 유효한 것만을 UMM의 호스트별 URL에 반영시킨다
④ 모든 수행이 끝나면 관리자에게 메일을 통해 통보한다

5 결론 및 향후 연구 계획

제안된 에이전트 기반의 디지털 도서관에서는 현재의 웹의 확산에 따른 웹 검색 서비스의 도입이 필요하며, 이에 맞추어 에이전트화, 신뢰도 향상, 대용량 처리 기능, 자동회복 갖추고 온라인 상에서 작동하는 로봇을 설계 구현하였다 이것은 기존의 로봇의 성능을 향상 시켜 대용량 웹에서 작동하도록 한 것으로, 에이전트화 되면서 수집한 정보의 공유를 가능하게 하였고, 자동화를 갖추어 사용자의 편리성을 증대 시켰다

그러나 대용량 처리의 효율성을 위해서는 URL 정보의 저장을 위한 저장 구조 및 로봇의 방문 정책에 대한 연구가 더 필요하다. 또한 구현된 에이전트 기반 디지털 도서관에서의 검증이 필요하다

6. 참고 문헌

[1] C.Mic Bowman, Peter B Danzig, and Darren R Hardy "Harvest: A Scalable, Customizable Discovery and Access System", Technacal Report CU-CS-732-94, Department of Computer Science, University of Colorado-Boulder, March 12, 1995

[2] Tim Finin, Richard Fritzson, Don McKay, and Robin McEntire, "KQML as an Agent Communication Language", Proc Of the Third International Conference on Information and Knowledge Management(CIKM'94), 1994

[3] Venkat N. Gudivada, and Siva Perraju Tolety "A Multiagent Architecture for Information Retrieval on the World-wide Web", Proc. of RIAO '97

[4] 심해청, 김병만, 김태남, 이준호, 최윤수, "효율적인 웹 로봇의 설계 및 구현에 관한 연구", 한국정보과학회 가을 학술발표논문집, 1997

[5] 양영걸, 안미정, 옥철영, "META 검색 엔진을 이용한 웹 로봇에 관한 연구", 한국정보과학회 가을 학술발표논문집, 1997

[6] 주원균 외. "디지털 도서관을 위한 문서와 링크 정보의 분리", 한국정보과학회 봄 학술 발표논문집(B), 137-139, 1998

[7] InterGate, "검색엔진 비교 분석 및 한글 검색엔진 검색 정밀성 실험", http://www.igate.co.kr/review/hody-re.html, 1997

[8] 김동욱, 류준형, 주원균, 맹성현, "링크정보를 이용한 검색 신뢰도의 향상", 정보과학회 봄 학술발표논문집(B), 446-448, 1998