정확도 높은 검색 엔진을 위한 문서 수집 방법

하은용, 최선원
안양대학교 컴퓨터공학과, 정보통신공학과

A Document Collection Method for More Accurate Search Engine

Eun-Yong Ha, Sun-Wan Choi
{eyha,sunchoi}@aycc.anyang.ac.kr
Dept. of Computer Engineering, Anyang University

요 약

인터넷상의 정보 검색 엔진들은 웹 로봇을 실행해서 인터넷에 연결되어있는 수많은 웹 서버들을 방문해서 웹 문서를 획득하고, 인덱싱 기법을 써서 자료를 추출하고 분류해서 검색 엔진의 기초가 되는 데이터 베이스를 구축한다. 정보 추출을 위해 웹 로봇을 운영할 때 웹 서버에 대한 사전 지식 없이 진행된다면 수많은 불필요한 요구가 전송돼서 인터넷 트래픽을 증가시키는 요인이 된다. 하지만 웹 서버가 사전에 자신이 공개할 문서에 대한 요약 정보를 웹 로봇에게 통보하고, 웹 로봇은 이 정보를 이용해서 웹 서버의 해당 문서에 대한 정보 수집 작업을 처리한다면 불필요한 인터넷 트래픽을 줄일 수 있을뿐만 아니라 검색 엔진의 정보의 정확도를 높이고, 웹 서버와 검색 엔진의 부하도 줄일 수 있는 효과를 가질 수 있을 것이다. 따라서, 본 논문에서는 웹 서버상의 웹 문서 파일의 변동 사항을 자동으로 점사하고 변동된 사항들을 종합 정리해서 등록된 각 웹 로봇에게 전송하는 문서 감시 통보 시스템과 통보된 요약 정보를 토대로 웹 서버로부터 해당 문서를 전송받이 필요한 인덱스 정보를 추출하는 효율적인 웹 로봇을 제안한다.

1. 서 론

인터넷에서 웹 사용자에게 손쉬운 정보 검색 서비스를 제공하는 대표적인 검색 엔진인 Northernlight, Google, Yahoo, Lycos, AltaVista, Excite, HotBot, MetaCrawler 들은 자신이 제공하는 정보의 최신성과 신뢰도를 높이기 위해 수많은 웹 서버 사이트들을 항해하면서 정보수집을 위해 독자적인 웹 로봇을 운영하고 있다. 하지만 이렇게 얻이진 자료가 방대하기 때문에 정확하게 유지 관리하는비용과 이로 인한 인터넷 트래픽도 대단할 것이다. 지금까지 알려진 인터넷상에서 자료를 수집하는 웹 로봇의 수가 수백 개에 이르고, 알려지지 않은 웹 로봇의 수도 상당할 것이다[1]. 이런 웹 로봇들이 항해하는 횟수가 커질수록 네트워크 트래픽은 엄청나게 증가되고 웹 로봇의 요구를 처리하는 웹 서버의 작업 부담은 자신의 로컬 작업을 처리하는데 있어 효율 저하를 가져오는 원인이 되고 있다. 이렇듯 인터넷 상에서 정보 검색이 대량화되면서 정보 검색에 대한 연구로는, 1) 정보 검색 속도를 빠르게 하기위해 정보의 근접성을 유지하는 Internet Caching 방법에 대한 연구가 있고[2], 2) 여러 검색 엔진에 분산되어 있는 정보를 종합해서 검색 서비스를 제공하는 메타 정보 검색 방법에 대한 연구도 있고, 3) 정보 검색 결과의 신뢰도를 높이기 위한 방법에 대한 연구가 있다 다시 말하면, 검색된 결과에 포함되어 있는 항목이 더 이상 존재하지 않는 죽은 링크(Dead Link)이거나 웹 서버상의 문서 변경 사항이 반영되지 않고 부정확한 인덱스 정보를 갖고 있는 등의 ,문제를 해결하기 위한 연구가 진행되고 있다

정보의 생명은 정확성에 있다 최신의 정보를 유지하기 위해 검색 엔진 중심의 정보 수집 방법을 탈피해서 정보의 원천지인 웹 서버의 적극적인 협조를 얻어야한다. 그래서 신뢰도 높은 최신의 정보를 유지하고, 잘못된 인터넷 트래픽의 발생도 감소시키고, 웹 서버의 부하도 줄일 수 있는 연구가 필요하다 따라서, 본 논문에서는 웹 서버상의 웹 문서의 변동 사항을 자동으로 검사하고, 변경된 사항들을 종합해서 정보를 원하는 등록된 각 웹 로봇에게 전송하는 문서 감시 통보 시스템과 통보된 종합 정보를 토대로 웹 서버로부터 해당 문서를 획득해 인덱스 정보를 추출하는 효율적인 웹 로봇을 제안한다.

본 논문의 구성은 다음과 같다 2장에서는 인터넷상의 정보를 수집하는 일반적인 방법의 문제점을 분석하고, 3장에서는 웹 서버의 협조를 통한 정보 수집 모델에 대해 설명하고, 4장은 웹 로봇 등록절차에 대해 설명하고, 5장은 문서 상태 정보 파일 생성에 대해 설명하고, 6장에서는 변경 정보 파일을 종합하는 알고리즘에 대해 설명하고, 7장에서는 변경된 문서를 획득해서 인덱스를 추출하는 웹 로봇에 대해 설명하고, 결론에서는 연구의 결과를 요약한다.

2. 웹 로봇의 정보 수집 방법의 문제점

웹 로봇은 일종의 웹 클라이언트로 초기의 URL로 지정된 웹 서버를 접근해 원하는 정보를 수집해서 자신의 데이타베이스에 저장한다. 이때 중복되는 URL 방문을 방지하기 위해 별도의 상태정보를 유지한다. 그러나 URL에 명시된 웹 서버가 존재하지 않은 경우이거나 다운된 경우에도 웹 로봇은 웹 서버에게 TCP/IP 연결 요구을 보낸 후 서버의 무응답에 대한 처리를 한다 또한 URL에 명시된 문서가 존재하지 않는 경우에도 웹 서버에게 연결 요구를 해서 설정된 TCP 연결을 통해 해당 문서의 전송 요구를 보낸다. 이때 웹 로봇은 전송 요구에 대한 에러 응답 메시지를 받아 처리한다. 이런 작업들은 웹 서버에 대한 정보가 부정확하기 때문에 발생하는 문제다. 이는 결국 인터넷상에 불필요한 트래픽을 발생시킬 뿐만 아니라 웹 서버 또는 웹 로봇이 실행되는 시스템에 부하를 가중시킨다

3. 웹 서버의 협조를 통한 정보수집 모델

앞에서 설명한 방법의 문제를 해결하기 위해 (그림 1)과 같은 정보 수집 모델을 제시한다 이 모델은 기본적으로 정보의 근원지인 웹 서버가 공개할 문서에 대한 정보를 전파하는 임무를 수행하고, 웹 로봇은 전달 받은 정보를 이용해서

실제 필요한 문서를 획득해서 데이터 베이스를 구축하게 된다. 웹 로봇은 웹 서버로부터 변경사항이 있음을 통보받은 문서에 대해서 인덱스를 추출하므로 인덱스의 신뢰성을 향상시킬 수 있을 뿐만 아니라, 웹 로봇이 불필요하게 인터넷 상을 항해하는 것을 막을 수 있으므로 인터넷 트래픽을 줄일 수 있는 장점을 갖는다.

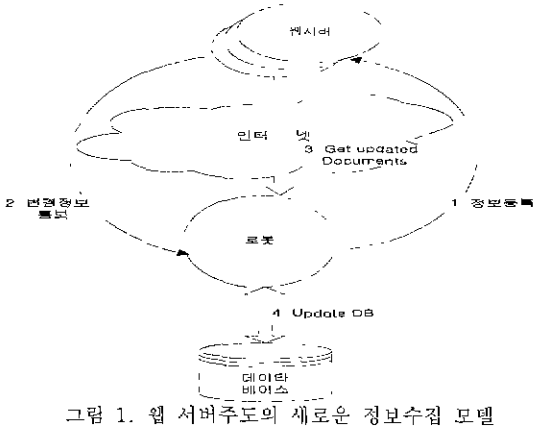

그림 1. 웹 서버주도의 새로운 정보수집 모델

제시한 문서 수집 절차는 (그림 1)에서 보듯이, 1) 먼저 웹 로봇이 문서 변경 통보 주기(latency), E-Mail 주소 등의 사항을 웹 서버에게 등록한다. 2) 웹 서버는 자신의 공개 문서들의 변경을 감시하고, 그 결과를 종합해서 원하는 웹 로봇에게 통보한다. 3) 웹 로봇은 해당 웹 서버로 부터 변경된 문서만을 읽어 인덱스 데이터 베이스를 변경하는 과정을 반복한다. 제안한 정보 수집 모델의 시스템 구성은 (그림 2)와 같다.

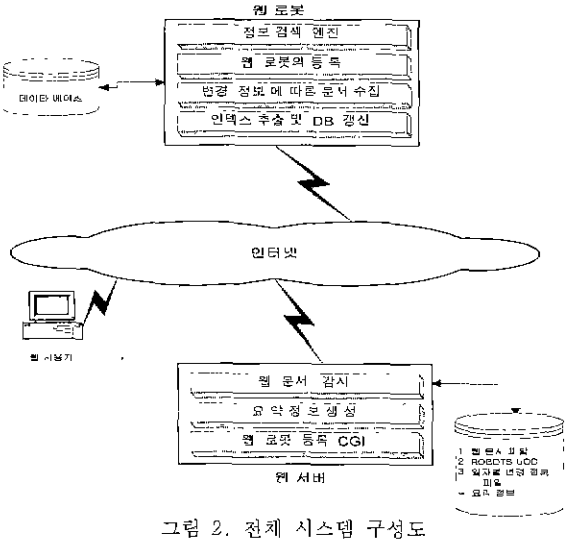

그림 2. 전체 시스템 구성도

4. 웹 로봇 등록 절차

웹 로봇은 웹 서버에게 자신에 대한 정보를 등록해야만 웹 서버가 이를 토대로 각 웹 로봇이 원하는 정보를 제공하게 된다. 이것은 웹 로봇의 과도한 요구에 의한 시스템 부하로부터 시스템을 보호하려는 로봇 배제 방법[3]과 비교할 때, 웹 로봇이 관심 있는 웹 서버를 선택하고 웹 서버가 변경사항을 통보한다는 점에서 다르다. 웹 로봇의 등록은 웹 서버상의 CGI 프로그램을 실행해서 처리한다. 웹 서버는 자신의 디렉토리에 "robots.uod" 파일에 각 웹 로봇에 대한 정보를 기록한다. 이 파일에는 문서 변경 통보 주기, 웹 로봇의 E-Mail 주소, 변경 통보 일시 중단 여부 등과 같은 항목이 수록된다. (그림 3)은 등록절차를 보여주고 있다.

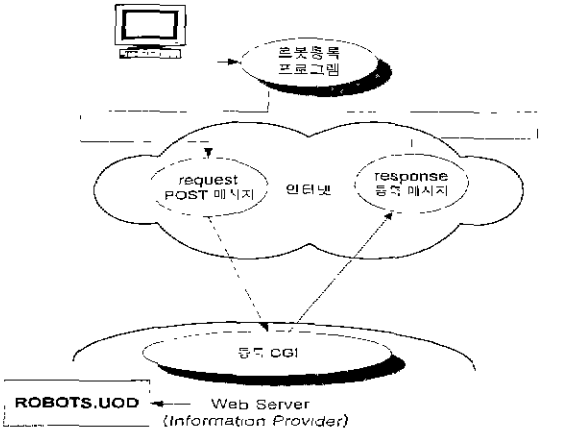

그림 3. 웹 로봇이 웹 서버에 등록하는 절차

먼저 웹 로봇은 웹 서버와 TCP 연결을 설정하고, HTTP [4]에 따라 웹 서버상의 등록 CGI 프로그램을 호출하기 위해 POST 메시지를 구성해서 웹 서버에게 전송한다. 이 POST 메시지에는 웹 서버의 등록 CGI 프로그램의 URL과 웹 로봇에 대한 상세 정보가 포함된다. 웹 서버에서 등록 CGI가 실행되면서 최초 등록일, 최종 전송일 등의 관리상 필요한 부가적 필드를 "robots.uod" 파일에 기록한다.

5. 문서 상태 정보 파일 생성

웹 서버는 공개할 문서 전체를 감시하고 문서의 snapshot을 주기적을 생성해서 파일에 기록해둔다. 이 파일은 일자별로 생성된다. 문서 감시 프로그램의 실행은 주기적으로 프로세스를 실행시키는 기능을 이용한다. 문서 감시 프로그램의 처리과정은 (그림 4)와 같다.

일자별로 문서에 대한 변경사항을 유지 관리하는데 일자별 상태 정보 파일의 각 엔트리는 다음 항목으로 구성된다.

- *Document's Status* : 문서 변경 상태로, 'A' 생성, 'U' 변경, 'D' 삭제
- Document's Name : 문서명
- *Document's Size* : 문서의 크기
- Last Modified Date : 문서의 최종 변경 시간

하나의 웹 문서에 대해 이전 상태와 현재 상태를 비교하여, 다른 경우 현재 상태를 변경 정보 파일에 기록하고, 이 과정을 모든 파일에 대해 반복 처리하여 문서 변경 정보를 생성한다

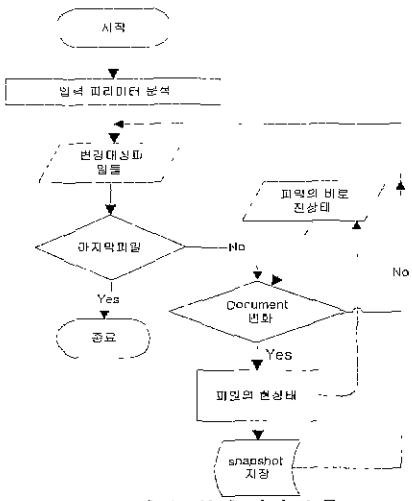

그림 4. 문서 감시 흐름도

다음은 웹 서버의 문서 감시 프로그램의 처리과정이다.

① 전체 감시 대상 파일에 대한 이전 상태를 읽는다.
② 한 파일에 대한 현재 상태를 얻는다.
③ 이전 상태와 비교해서 다르면 일자별 변경 정보 파일에 상태를 'U'라고 기록한다.
④ 파일의 이전상태는 있는데 현재 상태가 없는 경우는 상태를'D'라고 기록한다.
⑤ 현재 상태만 있는 경우는 일자별 변경 정보 파일에 문서의 상태를 'A'라고 기록한다
⑥ 모든 감시 대상 파일에 대해 ②③④⑤과정을 반복 처리한다.

6. 변경 요약 정보 파일 생성

웹 로봇에게 변경 사항을 통보해야 할 시간이 되면 최종 통보일로부터 현재일까지 상태 정보를 처리하여 요약정보를 생성하게 된다. 이 기간동안 한 웹 문서가 여러 번 변경되어도 처음 상태와 마지막 상태에 따라 요약 정보를 만들어 웹 로봇에게 E-mail로 전송한다. 이에 대한 구체적 처리는 다음과 같다. 'A'는 생성, 'U'는 변경, 'D'는 삭제를 각각 의미한다.

● 문서가 처음 생성된 상태에서 변경되는 경우:
① 문서가 생성되면 상태는 'A'다.
② 'A'상태에서 수정되면 'U'가 되고, 그 이후에 계속 수정이 발생해도 'U'상태를 유지한다.
③ ②의 'U'상태에서 삭제되면 'D'상태로 된다.
④ ③의 'D'상태에서 생성 외에 다른 사건은 발생할 수 없으며, 문서가 다시 생성되면 상태는 'A'다.
⑤ ①의 'A'상태에서 삭제되면 'D'상태가 된다.

이밖에 변경 상태와 삭제 상태에서 시작되는 경우도 같은 방식으로 생각할 수 있다.

7. 변경 통보된 문서 수집

웹 로봇은 웹 서버가 통보한 문서 변경 메일을 확인하고, 각 레코드를 처리하면서 변경된 문서를 웹 서버로부터 읽어온다. 이 웹 문서를 분석해서 색인을 추출하고 DB에 저장한다. 그 과정은 (그림 5)에 나타나 있다.

웹 로봇은 웹 서버와 TCP 연결을 설정하고 해당 문서들을 모두 수집한 후 TCP 연결을 닫은 후, 문서들에 대해 일괄 처리한다. 불필요한 처리를 없애기 위해서 수집된 문서의 앞에 Hostname, URL, Action, Size LastModified Date를 포함시켜 인덱스를 추출하는 과정에서 데이터 베이스의 기존 인덱스 내용과 비교할 수 있도록 한다

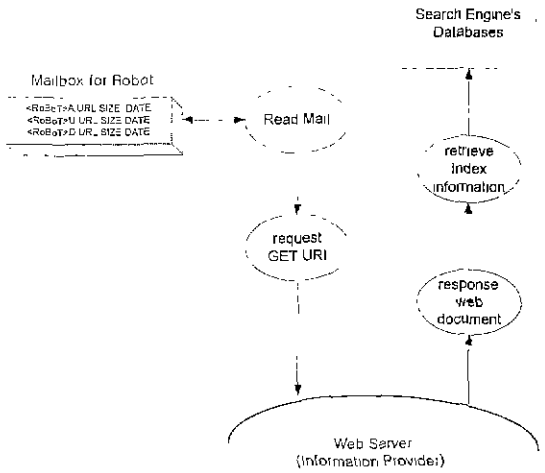

그림 5. 변경 통보된 문서 처리과정

8. 결론 및 향후 연구

본 논문은 최신의 정보를 위해 검색 엔진 중심의 정보 수집 방법을 탈피해서 정보의 원천지인 웹 서버의 적극적인 협조를 얻어 정확하고 필요한 정보를 획득하는 방법에 대해 연구했다. 웹 서버상의 웹 문서의 변동 사항을 자동으로 검사하고 변경된 사항을 종합해서 등록된 각 웹 로봇에게 전송하는 문서 검사 통보 시스템과 통보된 정보를 이용해 해당 문서를 전송 받아 인덱스 정보를 추출하는 웹 로봇을 설계 구현하였다. 제시한 모델에 따라 웹 검색 엔진과 웹 서버들이 상호 협조하면 인터넷 상의 불필요한 트래픽을 줄일 수 있을뿐만 아니라 검색 엔진과 웹 서버 시스템의 부하를 줄일 수 있을 것이다.

향후 제시한 모델에 대한 성능 평가를 위해, 먼저 웹 서버에서 문서 감시를 위해 부가되는 시스템 부하를 측정하고, 검색 엔진의 데이타 베이스에 저장되어 있는 전체 자료량에 대한 부정확한 자료량, 자료 갱신 주기의 변화에 따라 발생하는 네트웍 트래픽을 분석할 것이다.

참 고 문 헌

[1] M. Koster. " Robots in the Web: threat or treat ?" , April 1995
[2] D.Wessels and K. Claffy. "Application of Internet Cache Protocol (ICP), Ver 2", National Laboratory for Applied Network Resarch/UCSD, September 1997
[3] M. Koster,"A Standard for Robot Exclusion ", http://info.webcrawler.com/mak/projects/robots/exclusion.html
[4] R. Fielding, J. Gettys, J. C Mogul, H Frystyk, L. Masinter, P. Leach,T Berners-Lee, "Hypertext Transfer Protocol--HTTP/1.1 ", September 1998