웹 로봇 에이전트의 효율적인 인터넷 정보검색

김동범[0] 곽병정 김연옥 오용철 이재영
한국산업기술대학교 컴퓨터공학과
{imbumi[0], ocs_navy98, okthink, oh, jylee }@kpu.ac.kr

Efficient Information Retrieval of A Web Robot Agent on the Internet

Dongbum Kim[0], Byoungjung Kwak, Yeon-ohk Kim, Yongchul Oh, Jaeyoung Lee
Dept. of Computer Engineering, Korea Polytechnic University

Key Words: Agent, Web Robot, Robot agent

Abstract

인터넷상에서의 정보검색은 검색엔진을 이용하여 이루어지는데, 방대한 사이트들을 검색하여야 하므로 검색효율이나 검색된 정보의 유용성에 문제가 있게 된다. 만약 이러한 정보들을 미리 자동적으로 검색, 분류해서 저장한다면 위의 두 가지 문제들을 해결할 수 있을 것이다. 자동적으로 이런 일을 처리하도록 고안된 것이 웹 로봇 에이전트라고 하며 현재 국내에도 여러 개의 웹 로봇 에이전트를 이용한 검색엔진이 사용되고 있다.

본 논문에서는 검색엔진을 구현하기 위해 하이퍼텍스트 전송규약에 대한 연구와 웹 로봇 에이전트에 대한 연구를 하여 올바른 로봇 에이전트를 구현하여, 구현된 검색엔진을 통한 효율적인 정보검색을 실현하는데 목적이 있다.

1. 서 론

World Wide Web의 발전으로 인하여 웹 상에서 제공되는 정보가 기하 급수적으로 증대되고 변경됨에 따라 웹 사용자가 자신이 원하는 정보를 정확하고 신속하게 검색하기에 많은 어려움이 있다. 이러한 원인에 의하여 웹 로봇 에이전트(Agent)라는 새로운 분야가 등장하게 되었다.

에이전트의 사전적인 의미는 사람의 일을 대신 도와주는 것이며, 웹 에이전트와 일반 프로그램과의 차이점은 에이전트는 맡은 목표를 달성하기 위하여 외부환경변화에 적응을 하면서 자율적인 행동을 한다는 것이라고 할 수 있다. 웹 에이전트 또는 웹 로봇, 웹 로봇 에이전트란 자동으로 네트워크를 순회하며 웹서버의 위치를 파악, 웹서버의 문서정보를 수집하고, 이렇게 수집한 정보를 바탕으로 효율적인 검색서비스를 제공하는 것이다. 즉 자동적으로 웹의 하이퍼텍스트 구조를 따라 다니며 문서를 추출하고, 그 문서에서 참조되는 다른 문서들을 추출하는 식으로 동작하는 프로그램이다

현재 국내에는 10여 개 이상의 웹 로봇 에이전트를 이용한 검색엔진들이 동작 중에 있으며 많은 로봇 에이전트들이 활동하고 있다. 이렇게 활동하고 있는 국내 로봇에 의해 풍부한 정보검색서비스를 제공하는 긍정적인 면도 있지만 단순한 문자열 비교에 의한 정보검색에 지나지 않아 그 정보검색의 정확성에는 사용자의 요구를 충족시키지 못하고 있고, 또한 네트워크나 상대방 웹서버에 부하가중 등 역기능이 발생하기도 한다. 따라서 사용자에게 인터넷망에서의 효율적인 정보검색 서비스를 제공하기 위해서 로봇 에이전트에 대한 필요성 인식은 물론, 올바른 제작이나 사용에도 관심을 가져야만 한다.

본 논문의 목적은 올바른 로봇 에이전트를 구현하여 웹 문서 정보에 대한 데이터베이스를 구축하여 효율적인 정보검색을 실현할 수 있는 타당성을 보이는 것이다. 본 논문의 구성은 다음과 같다. 2장은 관련연구로 로봇 에이전트의 연구와 구현 및 효율적인 정보검색을 위한 설계와 구현에 대하여 기술한다. 3장에서는 본 논문에서 구현한 검색엔진에 대하여 기술하고 4장에서는 결론을 맺는다.

2. 관련연구 및 설계와 구현

본 논문에서는 웹 로봇으로 인해 인터넷망에서 발생되는 트래픽(Traffic)과 웹서버의 부하를 최소화하며 동작되는 에이전트를 구현하여 정보를 수집하고, 수집된 정보를 문자열의 가중치에 의한 분류저장, 역파일 및 색인파일을 이용하여 사용자에게 정확도와 검색속도를 향상하고자 한다.

2.1 로봇에이전트를 이용한 효율적인 정보검색의 설계와 구현

2.1.1 로봇 에이전트(Robot-Agent)

에이전트란 사용자를 도와주기 위해 자동으로 자율적으로 환경을 감지하고 목적에 맞게 행동함으로써 환경을 변화시켜 나갈 존재로 볼 수 있다. [1]

로봇 에이전트는 웹서버를 순회하며 각 홈페이지에 있는 수많은 정보를 수집하는 프로그램이다. 결국 웹서버에 접속해 데이터를 가져오는 기능적인 측면만 보면 웹브라우저와 같은 기능을 하는 셈이다. 단지 웹브라우저는 가져온 데이터를 화면에 보여 주고 하이퍼링크 등의 기능을 사용할 수 있지만, 로봇 에이전트는 데이터를 분석하고 그 안의 URL을 추출해 인터넷상의 웹 호스트들을 방문하며 정보들을 수집한다.

2.1.2 로봇 배제의 표준

현재 많은 로봇 에이전트가 활발한 활동을 하고 있으며, 어쩌면 사용자가 홈페이지에 접근하는 횟수보다 로봇 에이전트가 자동으로 접근하는 횟수가 많을 수도 있다. 일반적으로 로봇 에이전트가 웹서버를 방문 시 짧은 시간에 많은 페이지를 읽어간다. 또한 효용성을 검사하지 않기 때문에 쓸데없는 페이지도 많이 가져간다. 이렇게 로봇이 짧은 시간동안 많은 양의 통신을 요구하기 때문에 순간적으로 네트워크에 많은 트래픽을 발생시키며 서버에 부하를 준다. 때문에 웹서버가 원하지 않는 로봇 에이전트 또는 웹서버 로컬영역의 특정URL 접근하지 못하도록 설정하는 방법이 제안됐는데, 이것이 바로 '로봇 배제에 대한 표준(A Standard for Robot Exclusion)'[2]이다.

2.1.3 로봇 에이전트의 구현

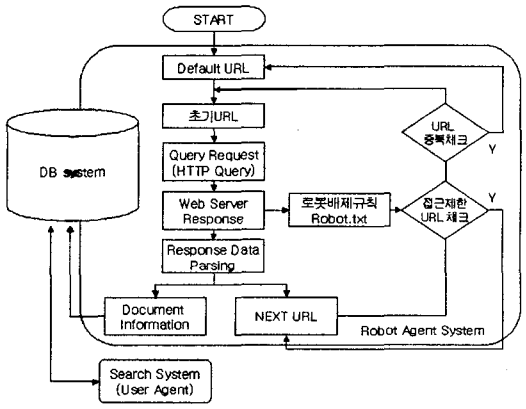

그림 1 로봇 에이전트 구성도

그림 1은 본 논문에서 구현한 로봇 에이전트의 구성도를 나타낸다. 로봇이 인터넷망을 순회하기 위해서 방문을 시작할 URL은 사용자가 입력해야만 한다. 일반적으로 많은 웹서버를 방문하기 위해 외부로 나가는 링크가 많은 페이지를 설정해 주어야만 한다. 로봇은 사용자가 입력한 URL을 바탕으로 문서를 수집한다. 방문할 웹서버에 HTTP(Hyper Text Transfer Protocol)[3] Query를 요청하면 웹서버는 자신의 시스템 환경정보와 요청한 웹 문서의 내용을 응답한다. 웹서버에서 응답 받은 정보는 html문서로서 이 응답 받은 html문서를 로봇이 분석하여 다음에 방문할 URL을 결정하게 된다. 일반적으로 모든 웹 페이지는 하이퍼링크를 필수적으로 포함하고 있다.[4] 응답받은 문서정보의 하이퍼링크는 로봇에 의하여 분석되며 로봇이 인터넷상에서의 웹서버와 웹서버의 문서를 수집하는데 이용되는 기준이 된다. 이렇게 웹 로봇은 HTTP를 통하여 서버에 요청하고 응답 받고 분석하여 다시 요청, 응답의 작업이 짧은 시간동안 이루어지기 때문에 인터넷망의 트래픽과 웹서버에 부하를 발생시킨다. 인터넷망의 트래픽과 웹서버의 부하를 최소화하기 위해 구현된 웹 로봇은 웹서버의 Robot.txt에 있는 로봇 배제규칙을 따랐고 HTTP Qurey 요청시 GET에 의한 요청대신 HEAD에 의한 요청으로 인터넷망에서 발생되는 트래픽을 최소화하였다.[5]

2.2 효율적인 정보검색의 설계와 구현

현재 국내에서 10여 개 이상의 웹 로봇을 이용한 검색엔진이 동작 중에 있으나 단순한 문자열 비교에 의한 정보검색에 지나지 않아 그 정보검색의 정확성에는 사용자의 요구를 충족시키지 못하고 있다. 많은 양의 자료를 검색하지만 사용자가 요구한 정보는 불과 몇 건에 지나지 않다.

이에 본 논문에서는 사용자가 요구한 보다 정확한 문서를 검색하기 위해 응답 받은 웹 문서의 문자열중에서 가중치가 있는 키워드를 추출하여, 키워드사이의 거리를 이용하여 수집되는 문서의 제한을 두었고, 또한 역파일과 B^+-Tree를 이용하여 보다 정확한 문서의 정보를 추출할 수 있다. 이렇게 추출된 문서의 정보와 가중치를 이용하여 사용자에게 보다 정확한 정보를 검색할 수 있게되었다.

2.2.1 적합성 판정 알고리즘의 설계

적합성 판정 알고리즘은 전송 받은 웹 문서가 수집하려는 해당 분류의 웹 문서인지 아닌지 판정하여 가중치를 부여하는 알고리즘이다.

본 논문에서는 분야별로 해당하는 키워드(Key Word)를 두어 문자열 사이의 유사성에 따라 가중치를 판별 일정 기준치 이상의 자료들만 수집하여 정보 검색의 효율을 높였다.

- 정의 1. 유사모델은 문자열 s1, s2, s3이 다음과 같은 성질을 만족할 경우, 거리함수 d에 의하여 정의됨

$d(s1, s1) = 0$, $d(s1, s2) \geq 0$, $d(s1, s3) \leq d(s1, s2)+d(s2, s3)$

- 정의 2. 해밍거리 : 같은 길이의 문자열에 대하여 정의, 함수 d는 동일한 위치에 있는 서로 다른 기호의 수

예) d(text, that) = 2

- 정의3. 편집거리 : 문자열 s1을 s2로 변환하기 위해 입력, 삭제, 치환되는 기호의 최소 수로서 정의

$d(s1, s2) \geq | length(s1) - length(s2) |$

위의 정의는 다음의 정규식 으로 표현될 수 있다.

Σ : 어떤 기호의 집합

[a1,..,am] : Σ내의 symbol 범위, symbol은 순서 화 되어야함

$r<=k$: $\sum_{i=0}^{k} r^i$ (유한반복)

즉 수집되는 문서에 제한을 두고 서버에서 응답 받은 데이터의 문자열중에서 키워드가 등장하는 회수와 키워드사이의 거리를 계산하여 일정 개수와 일정 거리 이상의 문서들만 수집함으로서 정보 검색의 효율을 높일 수 있었다.

2.2.2 역파일 설계 기법

본 논문에서는 탐색효율을 높이기 위해 역파일 기법을 이용하였다. 역파일 기법은 색인 파일과 포스팅 파일로 구성되어 있다.

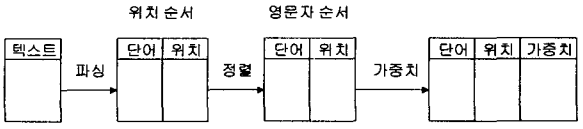

그림 2 역파일 생성의 구조

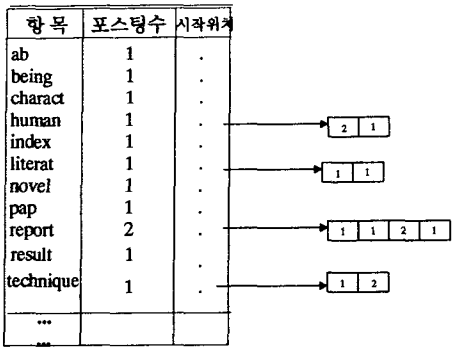

그림 3 구성된 포스팅 파일

그림2와 그림3은 역파일생성의 구조와 구성된 포스팅 파일을 나타낸다. 색인 파일은 각 분류별로 키워드에 해당하는 단어들로 구성하였고 효율적으로 탐색하기 위하여 B+-Tree, Hash등의 인덱스 기법을 이용하여 구성한다. 포스팅 파일은 색인어와 색인어가 출현한 문서간의 관계를 저장하는 파일로써 색인어와 문서간의 밀접도, 문서 내에서 색인어가 출현한 위치 정보 등으로 이루어진다.

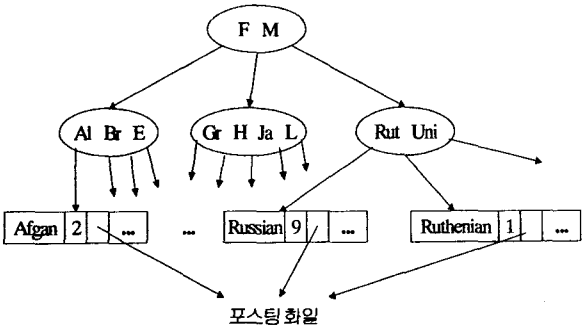

그림 4 B+-Tree 색인과 포스팅 파일의 연결

그림 4는 B+-Tree 색인과 포스팅 파일의 연결관계를 나타낸다. 본 논문에서는 이상과 같이 역파일을 구성함으로서 원하는 정보를 찾는데 대상URL이나 문서 수와 상관없이 일정한 속도를 유지하며 검색할 수 있다. 또한 각 분야별 키워드사전을 통해 색인어를 추출함으로서 검색효율을 개선하였다.

3. 웹 검색엔진의 구현

3.1 웹 검색엔진의 시스템 구성

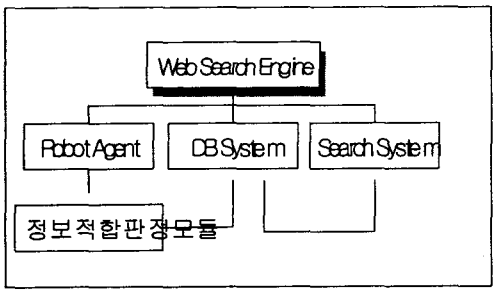

그림 5 웹 검색엔진 시스템 개요

그림 5는 본 논문에서 구현한 웹 검색엔진의 시스템 개요를 나타낸다. 웹 검색엔진은 크게 3가지로 볼 수 있다. 네트워크를 순회하면서 자료를 얻어오는 웹 에이전트와 얻어온 데이터를 저장하는 DB저장소, 사용자에게 검색서비스를 제공하는 검색엔진의 3부분으로 나뉠 수 있다. Red Hat 계열의 Wow Linux 7.1(Kernel 2.4.2-3 on an I686)의 운영체제 기반으로 GCC (Ver 2.96) 컴파일러를 사용하여 네트워크를 순회하는 웹 에이전트를 구현하였고 DB저장소는 Oracle 8i (8.1.7)를 사용, DBMS에 접근하여 저장된 데이터를 검색하여 사용자에게 효율적인 정보검색 서비스를 제공할 검색엔진은 PHP로 구현하였다.

4. 결론

인터넷상의 정보는 현재 엄청난 양이며, 매년 두 배 이상의 빠른 속도로 증가하고 있다[6]. 또한 뉴스그룹, 광고, 개인자료, 학술자료 등 인터넷 상에서의 정보의 종류는 매우 다양하며, 정보의 질 역시 매우 편차가 크다. 이들을 통제하는 수단이 없으므로 웹 브라우저의 서핑 방식은 본질적으로 인터넷상에서 사용자들에게 올바른 자료를 찾아 주는 것은 어렵다. 이러한 이유로 만들어진 로봇에이전트는 효율적인 정보검색을 가능하게 해줄 수 있다.

로봇 에이전트에 대한 기존의 연구는 대부분 통계적인 목적이나 검색엔진을 위한 데이터의 수집을 목적으로 사용되었다. 그러나 웹에서 제공되는 정보의 기하급수적인 증가와 더불어 산재된 많은 정보를 수집하기 위해 더 높은 성능의 로봇 에이전트들이 제작되었고 이러한 프로그램들이 팽창하면서 실제로 전체 네트워크를 과부하 시키는 현상을 초래하게 되었다. 재귀적인 방법으로 수행되는 로봇 에이전트의 사용을 억제하기 위한 연구들이 많이 발표되었으나 수동적인 방법에 의존하는 연구가 대부분이며 대표적인 것이 로봇 배제를 위한 표준안이다.

본 논문에서는 서버와 클라이언트사이에서 수행되는 로봇 에이전트에 의한 자동적인 정보획득 방법을 시도하여 네트워크의 과부하를 억제하면서도 정보의 신뢰성과 정확성을 보장하였다.

참 고 문 헌

[1] 최중민, "에이전트의 개요와 연구방향", 정보과학회지, 제15권, 제3호, 1997

[2] 로봇 배제에 대한 표준
http://www.robotstxt.org/wc/exclusion.html, 2002

[3][5] RFC 2616, Hypertext Transfer Protocol -- HTTP/1.1
ftp://ftp.isi.edu/in-notes/rfc2616.txt, 2002

[4] 박사준, 김상경, 황수철, 김기태. "전문가 검색 엔진에서 개념 그래프를 이용한 웹 정보 획득" 2000 봄 학술발표논문집(B) pp295-297. 한국정보과학회.

[6] 인터넷통계보고서, http://stat.nic.or.kr/stat_report.html