2002년 한국정보처리학회 춘계학술발표논문집 제9권 제1호

특정 사이트내의 검색 프로그램 구현에 관한 연구

장덕성*, 구세완**
*동원대학, e-비즈니스과
** LG 전자기술원
e-mail : dsjang@tongwon.ac.kr

A Study on Implementation for Web Search Program in Specific Web Site

Doc-Sung Jang*, Se-Wan Gu **
*Dept. of e-Business, Tongwon College
**LG Electronics

요 약

본 논문은 검색엔진을 이용하여 대상 웹사이트의 링크 사이트 전체를 수집하고, 각 링크 사이트의 페이지를 인덱싱하여 데이터베이스화하는데, 특히 가장 최신의 페이지를 분류하여 시간에 의해 검색단어의 정확도가 가려지는 경우, 이를 이용할 수 있도록 하였다. 본 논문은 검색엔진에 의해 검색 서비스를 제공하는 기본적인 웹 로봇의 구현에 대한 연구이며, 웹 로봇의 역할은 크게 링크 사이트를 수집하는 것 이외에 제목, 메타태그, 멀티미디어 다운로드등의 역할들을 수행하며 이를 인덱싱하여 데이터베이스화한다.

1. 서론

인터넷과 웹 사이트의 폭발적 증가로 인해 인터넷에 연결된 컴퓨터는 이미 거대한 데이터베이스의 집합군이 되었고, 원하는 정보를 언제, 어디서나 얻을 수 있게 되었다. 그러나, 그 많은 정보를 체계적으로 정리하고, 정말 원하는 정보를 검색하여 주는 지능형 검색 엔진의 연구는 아직도 지속되어야 할 숙제로 남아 있다. 하나의 사이트가 포함하는 정보는 텍스트, 파일, 이미지나 동영상등 멀티미디어 데이터, 다른 사이트를 링크하고 있는 링크 사이트등도 포함한다고 하겠다.

이는 쇼핑몰과 같은 사이트에도 매우 유용하게 적용될 수 있다. 쇼핑몰은 매우 방대한 자료의 물품에 대한 정보를 담고 있으며, 사용자가 원하는 물품이나 정보를 검색하기는 쉬운 일이 아닐 것이다. 따라서 그 안에서라도 원하는 정보를 쉽게 찾을 수 있는 방법을 강구하는 것은 정보전달 및 정보탐색을 주요 목적으로 하는 정보 제공자 및 사용자에게는 중요한 이슈가 될 것이다. 본 논문에서 구현한 검색엔진의 검색방법은 다음과 같다. 먼저 대상 사이트내의 모든 웹 페이지, 멀티미디어 데이터등을 인덱싱하여 데이터베이스에 저장한다. 이 때 레벨을 미리 설정하여 메인 홈페이지에서 몇 번째 링크 페이지까지의 웹 페이지를 데이터베이스화 할 것인지를 결정한다. 그리고, 검색엔진에 의해 검색할 단어의 입력에 따라 데이터베이스에서 검색하여 그에 해당하는 페이지를 출력하여 준다. 이 때 웹페이지를 데이터 베이스에 저장하는 역할은 웹로봇의 몫이며, 웹로봇의 성능과 기능이 전체 검색엔진의 성능을 좌우할 것이다.

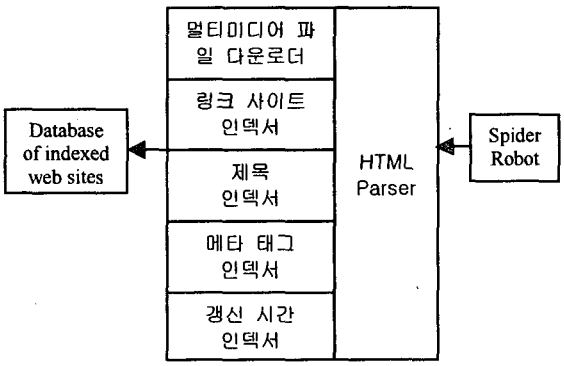
본 논문의 검색 엔진은 다음과 같은 기능으로 구성된다. 첫째, 홈페이지를 중심으로 그 사이트내의 모든 링크 사이트를 모두 수집하는 웹 로봇이 있으며, 둘째, 링크된 문서, 멀티미디어 파일들을 다운로드 하는 다운로드 로봇이 있으며, 셋째, 수집

된 각 링크 사이트로부터 제목, 메타 태그에 의한 정보와 그 사이트 자체의 페이지등을 카테고리별로 분류하고, 파싱을 수행하여 데이터 베이스화하는 인덱서, 넷째, 웹 페이지등의 시간별 업데이트 정보를 추출하는 타임 인덱서, 다섯째는 전체 검색엔진을 윈도우 컴포넌트화 하여 윈도우 플랫폼에서 검색엔진을 구현하려는 곳에서는 범용적으로 사용이 가능하도록 하였다.

웹로봇의 동작에 대한 알고리즘과 검색엔진에의 적용, 그리고 구현에 대하여 설명하겠다.

2. 목적 및 특징

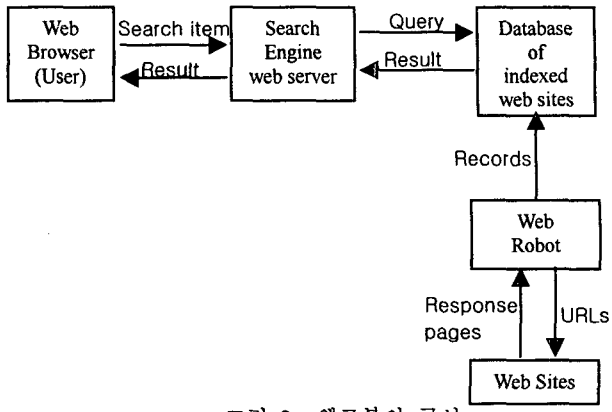
본 논문은 인덱스 된 전체 사이트를 대상으로 하는 검색 엔진의 구현 보다는 하나의 사이트가 포함하고 있는 정보를 정확히 찾아주는 것을 목표로 하였다. 구현한 검색엔진은 일반적인 검색 엔진의 구성을 따르면서, 웹로봇으로 부터 수집된 사이트로부터 다양한 컨텐츠정보를 인덱싱하고 이를 갱신된 날짜에 의한 검색이 가능하도록 시간 인덱싱을 추가하였다. 검색 엔진의 일반적인 구성은 그림 1 과 같다. 웹로봇으로부터 수집된 사이트는 인덱스하여 데이터베이스에 저장한다. 사용자측의 검색엔진은 데이터베이스에서 검색 단어를 찾아 사용자에게 보여주는데, 웹로봇이 수집하는 데이터와 인덱싱하는 방법이 전체 검색엔진의 성능을 좌우한다고 할 수 있다.



<그림 1> 검색 엔진의 구성

그림 2 는 본 논문의 웹로봇이 수집하는 웹사이트 내의 컨텐츠를 인덱싱하여 데이터베이스화 하는 과정에서 다양한 인덱싱이 가능하도록 필터와 파서를 갖는 추출기와 사이트별 갱신 시간을 인덱싱할 수 있는 기능에 대한 블록도를 보인다.

본 논문은 웹 로봇에서 수집된 다양한 웹사이트로부터 사용자의 다양한 요구에 부응할 수 있는 인덱서와 필터들을 구현하고, 이를 갱신 시간 인덱서에 의해 갱신 시간별로 검색할 수 있도록 서비스하는 것과 이를 컴포넌트화하는 것이 목적이다.

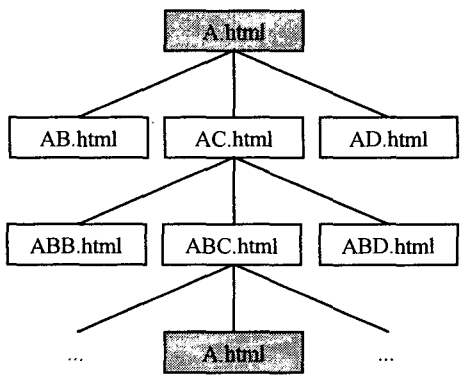


<그림 2> 웹로봇의 구성

3. 구현

3.1 웹사이트로부터의 전체 링크 페이지 수집기

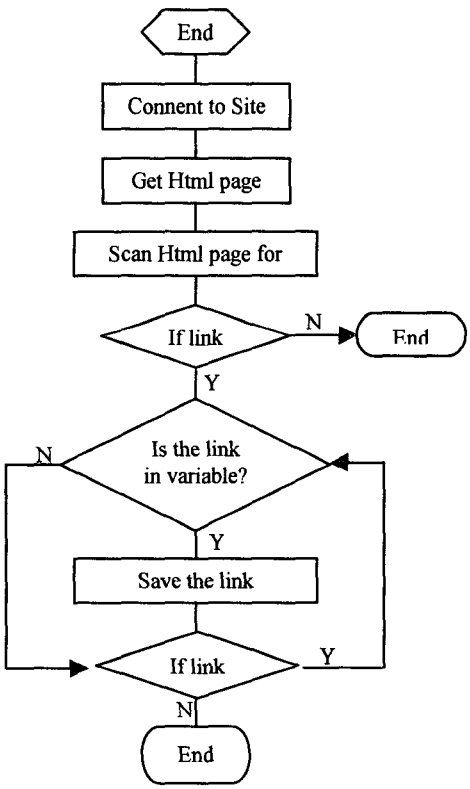
정해진 웹사이트에 접속하여 그 페이지내의 다른 링크 사이트(html, asp, php 등)를 수집하는 역할을 하는 것으로 가장 많은 시간이 걸린다. 먼저 각 링크 사이트에 대한 정보를 수집한 후에 각각의 목적과 용도에 알맞은 인덱서, 즉 필터의 역할을 하는 곳으로 전달하면 이들 인덱서에 의해 각 컨텐츠가 분류가 되어 데이터베이스에 저장이 되게 된다. 링크 페이지 수집기를 구현하기 위해 주의해야 할 몇가지가 있는데 그것은 첫째, 반복되어 링크가 되어 있는 사이트는 중복을 피해야 하며, 둘째, 파일 타입을 잘 구분해야 하며, 셋째, 링크된 사이트가 절대 경로인지 상대경로 인지에 대해 잘 판단해야 한다. 넷째, 첫번째 웹사이트에서 얼마만큼 깊이 있게 페이지를 수집할 것인지를 결정하는 것이다. 즉, 그림 3 과 같은 경우가 자주 발생할 것과 필요한 컨텐츠가 어느 깊이까지 존재할 것인지를 예측하여 깊이를 조절하여 링크 페이지를 수집하는 것이다.



<그림 3> 웹사이트의 링크 구조의 간략

본 논문의 검색엔진에 대한 시뮬레이션을 할 때, 특정 사이트들은 깊이가 2,3 정도만 되더라도 수백개에서 1000 이상의 링크가 존재하는 사이트도 상당히 많았었다.

그림 4 는 링크 페이지를 수집하기 위한 일반적인 알고리즘으로 간략하게 설명하면, 대상이 되는 사이트의 전체 페이지 소스를 가져와서 그로부터 <a href="...">내의 html, asp, php 등 링크 사이트를 수집하고 기존에 수집된 링크 사이트에 현재 링크 사이트가 있는지를 체크하는 루틴으로 구성이 되어있다.



<그림 4> 웹사이트로부터 링크사이트 수집알고리즘

3.2 멀티미디어 파일 다운로더

위의 HTML Parser 로부터 수집된 각 사이트의 페이지로부터 멀티미디어 파일, 즉 이미지(jpg, gif 등)와 음성 및 동영상(mpg, mp3 등) 파일에 대한 링크를 찾아서 다운로드 하는 것 역할을 하는 것으로 사용자가 원하는 어떠한 확장자도 가능하도록 하는 것이다. 이에 대한 응용으로 다양한 문서파일도 가능하다. 즉 hwp, doc, ppt, pdf 등 웹사이트에서 다운로드 받을 수 있는 모든 문서의 포맷이 가능할 것이다.

3.3 제목 인덱서

각 페이지에 담고 있는 제목(title)에 대한 정보를 추출하여 데이터베이스에 저장하는 역할을 하는 것이다. 예를 들면, "<title>Daum - 우리 인터넷, Daum</title> " 와 같은 html 소스에서 title 이라는 태크에 의한 정보를 저장하는 것이다. 이를 통해 각 페이지에 대한 것을 제목에 의해 정렬하고 검색할 수도 있을 것이다.

3.4 메타 태그 인덱서

html 소스에서 사용자에게는 보이지 않는 다음과 같은 meta 태그에 의한 정보를 추출하는 것으로서 이는 일부 검색엔진에서 키워드 검색에 사용이 되기도 한다.

```
<meta name="Description" content="...">
<meta name="keywords" content="...">
```

3.5 갱신 시간 인덱서

각 웹 페이지로부터 그 페이지에 대한 다양한 정보를 얻기 위해 HTTP 헤더 필트로부터 얻어 올 수 있다. 즉, Date, Content-type, Last-modified, Content-length, Content-Encoding 등의 정보들이며, 이 중에서 갱신시간을 나타내는 Last-modified 에 대한 것을 특별히 인덱싱하여 저장하는 것이다. 방대한 인터넷의 자료로부터 가장 최근 순으로 정리하였을 경우 가장 잘 찾을 수 있는 정보들의 수가 점차 늘어나기 때문이다. 즉 정보라는 것이 시간에 의해 그 중요성이 강조되는 경우가 점차 증가하기 때문이며, 이는 대개의 뉴스등에 많이 해당이 될 것이다. 이로 인해 카테고리를 시간에 의해 지정하여 검색을 할 수 있는 검색엔진이 나올 수 있으며, 언제 이전의 페이지는 폐기하는 기능들도 가능할 것이다.

3.6 조정자(Cordinator)

그림 2 에는 나타내지 않았지만, 위에 설명한 다양한 인덱서들을 적절하게 조합하여 데이터베이스에 저장하는 역할을 하는 조정자가 있다. 조정자에 역할은 크게 다음과 같이 나눌 수 있다.

1. 각 웹 페이지에 독특한(Unique)한 숫자 및 기호를 부여 (Numbering)
2. 데이터 베이스에 인덱서에 의해 필터링 된 데이터 저장
3. 각 웹 로봇을 일정 간격으로 구동시켜서 가장 최신의 데이터 베이스 유지
4. 각 웹 페이지가 혹시 죽은 사이트가 아닌지 일정 간격으로 체크

4. 실험

4.1 개요

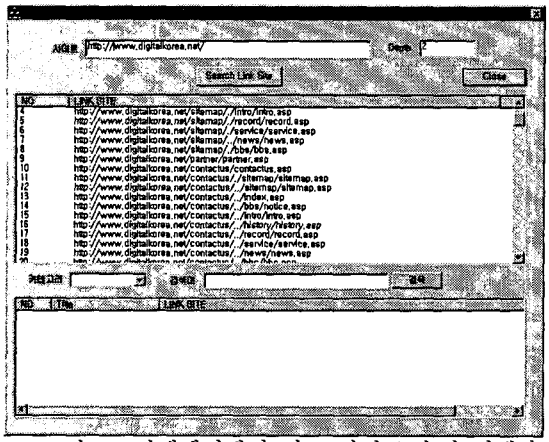
수집기에 의해 웹페이지를 수집하고 파싱하여 각

인덱서에 의해 필터링을 거쳐 그로부터 나온 각 결과를 출력하는 시뮬레이터를 구현하였으며, 윈도우 환경에서 비쥬얼 C++ 로 구현하였다. 따라서 WinInet API 클래스들을 많이 활용하였다. 또한 후에 사용자의 의도에 맞추어 다양하게 구현되도록 하기 위해 컴포넌트로 구현을 하여 Visual Basic, ASP 등에서 손쉽게 구현할 수 있도록 하였다.

4.2 웹 사이트로 부터의 전체 링크 페이지 수집기

그림 5 의 테스트 프로그램에서 사이트와 어느 정도의 깊이로 링크 페이지를 수집할 것인지를 설정하면 다음과 같이 링크 사이트등이 출력된다.

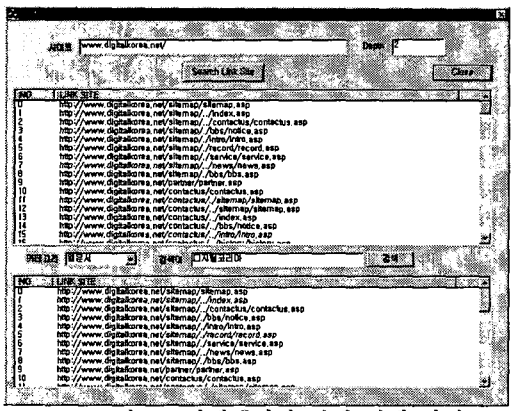
테스트는 디지털코리아㈜의 홈페이지인 http://www.digitalkorea.net 에 대하여 수행하였다.



<그림 5> 검색엔진에서 링크 사이트 수집 실행화면

4.3 검색엔진 수행

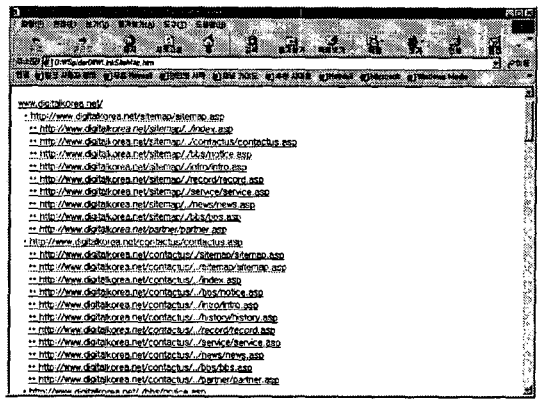
링크 사이트 수집을 통하여 수집된 사이트내에서 입력된 검색어에 대해 사이트를 찾는 것으로 그림 6 과 같은 결과를 볼 수 있다. 카테고리는 웹문서에서 찾았다.



<그림 6> 검색엔진의 검색 실행 화면

4.4 사이트맵 구축에 응용

링크 사이트 수집기를 구축된 사이트의 사이트 맵과 링크들이 올바르게 연결되어 있는지를 검사하는 도구로서도 응용할 수 있으며, 그림 7 은 이에 대한 응용으로 http://www.kips.or.kr 의 사이트 맵을 출력한 것이다. 링크 사이트를 수집하는 깊이는 3 으로 설정하였다. 따라서 그 이상의 링크 사이트는 수집되지 않았기 때문에 출력되지 않는다.



<그림 7> 링크사이트 수집기에 의한 사이트 맵

5. 결론 및 향후 연구 방향

본 연구는 특정 사이트에 대한 검색을 철저히 하기 위한 제목, 메타태그, 멀티미디어 파일등을 통한 여러 가지 방법을 모색하였다. 이를 실제 적용할 시에는 병렬처리등에 의해 성능을 고려해야 하고, 인덱싱 이후에 지능형 검색엔진을 접목하여 좀 더 정확한 결과를 보여주려는 노력이 있어야 할 것으로 본다.

참고문헌

[1] 양명석, "메타 검색에서 검색 결과 통합 랭킹 방법 및 결과 분석", 충북대학교 컴퓨터학과 석사논문, 2001

[2] 이상영, "상품 비교·검색을 위한 검색엔진 구현", 순천향대학교 전자상거래학과 석사논문, 2000

[3] 이상영,주경수, "최적화된 상품정보 비교검색을 위한 검색기 및 수집기 구축", 2001 인터넷 정보학회 제 3 회 춘계학술발표대회, pp.29-34, May 19, 2001.

[4] Roger S. Pressman. "Software Engineering, A Practitioner's Approach", 3rd Ed. McGraw Hill, 1997