빈 연결을 제거하는 메타 검색 엔진의 구현

김연곤, 엄채임, 변정용
동국대학교 전자계산학과

An Implementation of Meta Search Engine of Removing Empty-Link

Youn Gon Kim, Chae Im Um, Jeong Yong Byun
Dept. of Computer Science, Dongguk University.

요 약

지금까지 정보 검색에 대한 많은 연구가 있어 왔지만, 여전히 여러가지 문제들로 인해 사용자는 많은 시간을 소비하게 된다. 본 논문에서는 이러한 문제를 해결하기 위해 검색 결과를 분석하여 중복된 URL을 제거하고, 접근이 불가능한 URL 정보를 사용자에게 보여준다. 해결 방안으로는 멀티 쓰레드를 이용한 로봇 에이전트가 자동으로 각 URL을 방문함으로써 가능하게 했으며, 사용자는 직접 방문하지 않고도 접근 불가능한 이유를 미리 알게된다. 구현된 메타 검색 엔진을 기존의 검색 엔진들과 비교 했을 때 약 13%의 효율성 향상을 가져왔으며, 앞으로 시소러스 등을 이용한 더 많은 연구가 진행될 것이다.

1. 서론

전 세계의 수 많은 정보들이 인터넷 상에 있으며, 그 정보들을 효율적으로 관리하고 제공하기 위해 정보 검색이란 용어가 나타났다. 이미 수 많은 정보 검색 엔진들이 개발[3]되어 있으며, 많은 사람들이 한번 씩은 정보 검색을 해 보았을 것이다.

지금까지 정보 검색의 정확도를 높이기 위한 많은 연구가 있어왔다. 검색 키워드에 대한 높은 재현율을 보이기 위해 로봇이 자동으로 웹 사이트들을 돌아다니면서 큰 데이터베이스를 구축[2]하도록 했고, 키워드에 대한 정확도를 높이기 위해 질의어를 확장하거나 전문(full-text) 검색이 가능하게 하는 노력들이 한창 진행중이다.

현재 두드러지게 나타나고 있는 정보 검색의 문제점을 보면 데이터베이스 대량화에 따른 관리 부족으로 결과에 대한 정확도가 떨어진다는 것이다. 즉, 100개 이상의 결과를 가져올 때 실제 사용자 요구에 대응되는 결과는 많지 않다는 것이다. 또한 많은 검색 결과 중에서 중복된 결과를 가져오는 경우 사용자는 그 결과들 중에서 이러한 사실을 인지하지 못하고 두 번 이상 같은 결과에 접근하는 것이다. 그리고 실제 접근이 불가능한 사이트가 있다는 것이다. 이밖에도 정보 검색은 많은 문제들을 가지고 있으며, 이러한 문제의 가장 중요한 점은 사용자의 검색 시간을 낭비하게 한다는 것이다.

본 논문에서는 사용자의 검색 시간을 최소화하기 위해 검색 결과에 대한 중복과 실제 접근이 불가능한 연결을 제거하고자 한다. 문제의 해결 방안은 사용자가 검색을 요구할 때 로봇 에이전트가 자동으로 웹 사이트들을 방문하여 사용자에게는 접근 신뢰도가 높은 결과만을 보여주는 것이다.

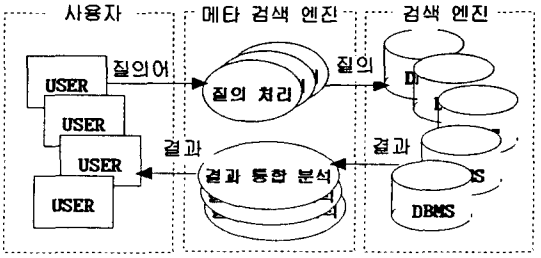
논문의 2장에서는 메타 검색 엔진 및 구체적인 문

제점을 기술하고, 현재까지 URL을 관리하고 있는 로봇 에이전트에 대해 알아본다. 3장에서는 로봇 에이전트를 이용한 메타 검색 엔진이 어떻게 빈 연결을 제거하는지에 대한 설계에 대해 기술하고, 4장에서는 멀티 쓰레드를 이용한 자바 환경에서의 구현과 검색 예를 보인다. 5장에서는 다른 검색 엔진과의 비교 실험을 통하여 구현된 메타 검색 엔진의 효율에 대해 기술하고, 6장에서 결론을 맺도록 한다.

2. 기존 연구

2.1 메타 검색 엔진의 수행방식

메타 검색 엔진은 그림 1과 같이 이미 구축된 여러 개의 검색 엔진에 단 한번의 질의문 입력으로 동시에 질의를 던지고 결과를 받아 사용자에게 보여준다. 이는 사용자가 각 검색 사이트에서 질의를 매번 해야하는 번거러움을 덜어줄 뿐만 아니라, 각 검색 사이트마다 서로 다른 사용자 인터페이스를 일관성 있게 함으로써 사용자에게 편리성을 제공한다.



[그림 1] 메타 검색 엔진의 수행방식

일반적으로 메타 검색 엔진들은 데이터베이스를 따로 구축하지 않고, 기존의 검색 엔진에서 리턴되는 질의 결과를 통합 분석하여 사용자에게 제공한다. 그러나 대부분의 메타 검색 엔진들은 질의문 확장이나 네트워크 부하[2], 사용자 인터페이스[3]만을 고려하여 설계하였으며, 질의 결과에 대한 접근 가능 신뢰도에 대해서는 고려하지 않고 있다.

현재 서비스 중인 메타 검색 엔진에는 12개의 검색 사이트에 질의를 던지는 미스 다찾니[9]와 자바로 구현된 깨비[7], 까치네[6]등이 있다.

2.2 검색 결과 분석

많은 메타 검색 엔진들과 직접 데이터베이스를 구축하고 있는 검색 엔진들의 검색 결과를 보면 결과들에 대한 접근 가능 신뢰도가 떨어지는 경우가 많다. 첫째로 결과 중에서 중복된 URL이 나타나 사용자가 똑같은 사이트를 여러번 방문하게 되거나, 둘째로 결과 사이트에 접근을 시도했을 때 아래와 같은 이유들로 인해 접근이 불가능할 때가 있다. 이러한 경우 사용자가 직접 접근을 시도한 후에 그 사이트의 접근이 불가능하다는 것을 알게 되므로 사용자 측면에서 많은 불편을 초래하게 된다.

두 번째 문제점인 결과 사이트의 접근이 불가능한 경우는 다음과 같이 구분해 볼 수 있다.

① 요청 시간내에 서버에서 응답이 없음
② 서버에 연결이 불가능
③ 서버의 이름을 DNS Table에서 찾을 수 없음
④ File을 찾을 수 없음
⑤ File을 읽을 수 없음
⑥ HTTP 개체를 찾을 수 없음
⑦ 결과 화면이 데이터를 가지고 있지 않음

이와 같은 사항들로 인해 사용자는 에러 메시지를 받을 때까지 에러를 인지하지 못한 채 시간을 소비해야 한다.

2.3 URL 관리 로봇

① 넷스케이프 네비게이터

네비게이터에서 북마크를 열면 "Update Bookmark" 메뉴가 있다. 이 메뉴를 선택하면 현재 북마킹되어 있는 사이트에 접근해 그 사이트가 유효한지 검사를 한다. 만약 북마킹한 내용이 변경되었다면 몇 개의 사이트가 변경되었는지를 알려준다.

② URL-minder

로봇 에이전트를 이용해 등록된 홈페이지를 검사하고 변경 사항이 있으면 홈페이지를 등록한 사용자에게 알려준다[11].

③ 체크봇(checkbot)

네덜란드의 Hans de Graaff가 개발한 것으로 HTML 파일을 따라 가면서 각 페이지의 문제점을 지적한다.

본 논문에서는 2.2에서 밝힌 기존 메타 검색 엔진들의 두 가지 문제점을 해결하고자 한다. 첫째로 중복된 URL을 제거하기 위해서 임시 파일을 이용하여 검색해야할 사이트를 정렬하고 사용자에게도 하나의 URL만 보여준다. 두 번째로 불확실한 접근을 제거하기 위해서 각 검색엔진들의 검색 결과에 대해

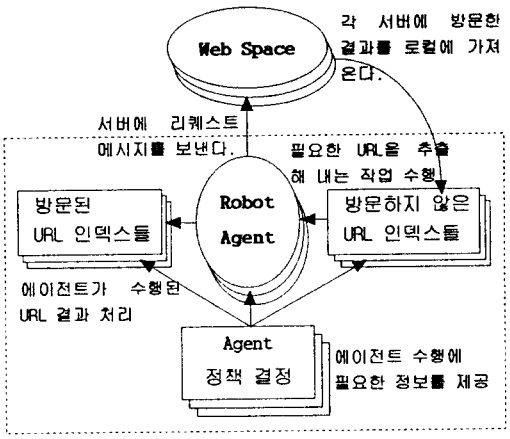
미리 접근을 해보고 실제 접근이 가능한 URL과 접근 불가능한 URL을 구별하여 사용자에게 보여준다.

3. 메타 검색 엔진의 설계

메타 검색 엔진에서 사용자가 질의를 던질 때 자동으로 여러 검색 엔진들에 알맞은 형태의 질의를 던지고, 각 검색 엔진들에서 리턴되는 결과들을 처리하는 일은 로봇 에이전트가 담당하게 된다. 즉, 웹 브라우저는 가져온 데이터를 그대로 화면에 보여주는 역할만을 하지만, 로봇 에이전트는 데이터를 분석하고 URL을 추출하여 자동으로 각 URL을 방문하는 역할을 한다[4].

3.1 로봇 에이전트의 동작

메타 검색 엔진은 로봇 에이전트에 의해 수행되며, 이 로봇은 전체를 관리하는 조정 에이전트와 실제 각 검색 엔진들을 검색하는 에이전트로 구성된다.



[그림 2] 로봇 에이전트의 동작 방식

로봇 조정 에이전트는 사용자가 질의를 던질 때 검색할 각 검색 엔진들에 대해 그림 2와 같은 에이전트를 각각 생성하며, 생성된 에이전트들의 수행이 모두 끝났을 때 사용자에게 검색 결과를 보여주는 역할을 한다.

생성된 각 로봇들은 검색 엔진의 데이터베이스에서 검색된 결과들에 대해 자동으로 방문을 하고 결과를 처리하는 역할을 한다. 즉, 로봇은 방문하지 않은 URL 인덱스에 있는 URL에 대해 방문을 하여 각 결과 메시지를 가져온다. 결과 메시지와 에이전트의 정책에 따라 로봇은 방문된 URL 인덱스를 생성한다. 이때 에이전트의 정책은 결과 메시지에 방문한 URL에 대한 에러 메시지가 있을 경우 사용자에게 에러를 알려줌으로써 검색 결과에 대한 접근 신뢰도를 향상시킨다.

3.2 로봇 에이전트 수행 알고리즘

① 전체 로봇 조정 에이전트가 검색할 검색 엔진과 URL을 정하고, 메타 검색 엔진에서 받은 질의를 각 검색 엔진에 맞는 포맷으로 생성한다.

② 각 검색 엔진에 대한 멀티 쓰레드를 이용한 각각의 로봇을 생성하고 소켓으로 서버와 연결한다.

③ 서버에 성공적으로 연결이 되면 방문하지 않은 URL 인덱스에 있는 URL로 Request Message를 보내고 그러면 서버는 Response Message를 보내온다. 서버에 성공적으로 연결 되지 않으면 방문된 URL 인덱스에 이 URL은 Server Error임을 알린다.

④ 서버로부터 Response Message를 받았을 때는 HTTP Header 또한 볼 수 있는데 header의 정보 중에서 상태 코드를 분석한다.

⑤ header의 상태 코드에 따라 현재 URL이 성공적인 상태 코드를 가졌으면 URL의 title을 방문된 URL 인덱스에 알려주고 실패한 상태 코드를 가졌으면 코드의 종류에 따라 에러 메시지를 방문된 URL 인덱스에 알려준다.

⑥ 방문하지 않은 URL 인덱스에 있는 모든 URL을 방문 했을 때 하나의 로봇은 수행을 끝나게 된다.

⑦ 각 로봇의 수행이 모두 끝났을 때 로봇 조정 에이전트는 방문된 URL 인덱스를 사용자에게 보여준다.

4. 메타 검색 엔진의 구현

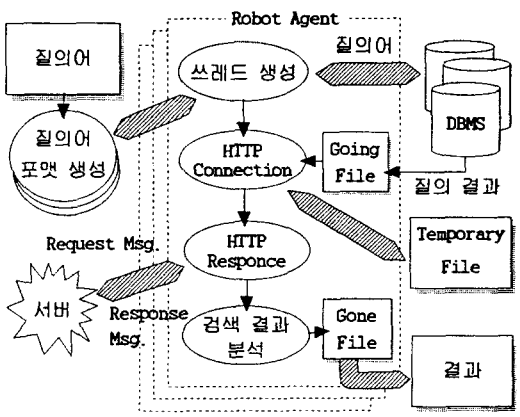
본 논문에서는 기존의 데이터베이스를 구성하고 있는 네이버[8], 코리아 야후[10], 미스 다찾니[9]를 검색 대상으로 메타 검색 엔진을 구현한다. 사용자의 질의 입력시 세 개의 쓰레드를 생성하여 각 데이터베이스에 대한 검색을 하고 결과에 대한 URL에 모두 방문한 후 각 쓰레드에 대한 결과를 통합하여 하나의 결과를 사용자에게 보여준다.

4.1 시스템 환경

① 플랫폼 :: UNIX Sun Enterprise5000
② 오퍼레이팅 시스템 :: sun-solaris2.4
③ 언어 :: java
④ 컴파일러 :: jdk 1.1.3
⑤ 주요 패키지 :: HTTPClient

4.2 각 모듈별 구현

빈 연결을 제거하는 메타 검색 엔진의 전체 구조는 그림 3과 같다. 사용자 인터페이스에서 질의어를 받아 세 가지 검색 엔진에 알맞은 질의 포맷으로 스트링 처리를 하고 각 검색 엔진을 검색할 쓰레드를 생성한다. 이 때 각 쓰레드는 파일을 공유하게 되므로 동기화가 필요하다.



[그림 3] 빈 연결을 제거하는 메타 검색 엔진

쓰레드가 생성되면 각 검색엔진의 데이터베이스에 질의를 하고 데이터베이스로부터 결과 리스트를 "Going File"에 저장한다.

"Going File"이 생성되면 HTTPClient 프로토콜[5]을 사용하여 "Going File"에 있는 URL을 하나씩 방문하게 되는데, 이때 세 개의 쓰레드가 같은 URL을 방문하는 낭비를 막기 위해서 쓰레드들이 공유하는 "Temporary File"에 방문하려는 URL을 저장하고 각 쓰레드는 URL을 방문하기전 "Temporary File"에서 방문되었던 URL인지 조사하여 중복성을 제거한다. 쓰레드들이 하나의 URL을 방문 과정을 보면 다음과 같다.

먼저 "Going File"에 있는 URL에서 SERVER_NAME과 PATH_NAME을 분리하여 서버에 HTTP Connection을 보낸다.

```
HTTPConnection con =
    new HTTPConnection("SERVER_NAME");
```

서버에 연결이 되면 실제 찾아갈 디렉토리나 페이지를 찾아가 HTTPResponse를 받는다.

```
HTTPResponse rsp =
    con.Get("PATH_NAME");
```

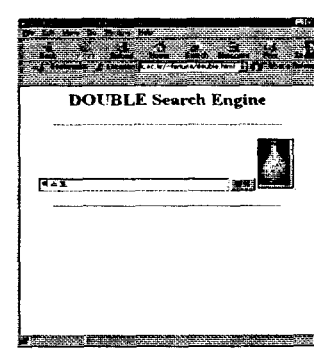
검색 결과의 상태 코드를 분석하여 에러 코드이면 "Gone File"에 URL에 대한 에러 메시지를 적어 주고, 성공한 코드이면 URL에 해당하는 "TITLE"을 적어준다.

```
if(rsp.getStatusCode() >= 300)
    System.err.println("Received Error:"
                       + rsp.getReasonLine());
```

각각의 쓰레드들에 대한 "Gone File"이 생성이 되면 세 개의 "Gone File"을 하나의 결과 파일로 생성하여 사용자에게 보여준다.

4.3 검색 예

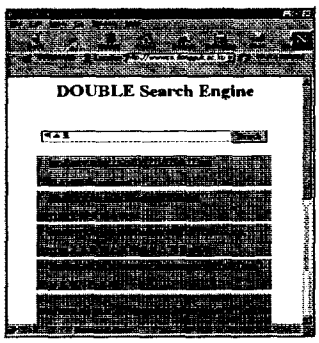
빈 연결을 제거하는 메타 검색 엔진의 그림 4와 같은 초기 화면에서 엑스포라는 용어로 검색을 했을 때 결과는 그림 5와 같다.



[그림 4] 검색 초기 화면

검색 결과는 각 검색 엔진에 대한 결과들이 현재 접근이 가능한지에 대한 여부와 가능한 경우는 사이

트의 제목을 보여주고, 불가능할 경우는 불가능한 이유를 사용자에게 보여준다.



[그림 5] 검색 결과 화면

5. 실험 및 평가

구현된 메타 검색 엔진이 기존의 검색 엔진의 문제점을 개선하였는가에 대한 평가를 하기 위해 실험한 검색 엔진은 네이버, 코리아 야후, 미스 다찾니 세 개의 검색 엔진이다. 여기서 네이버와 코리아 야후는 자체적으로 로봇을 돌려 데이터베이스를 구성하고, 미스 다찾니는 데이터베이스를 구성하기는 하지만 로봇에 의한 데이터베이스 구성이 아니라 기존의 구성되어 있는 데이터베이스를 검색해 로컬 데이터베이스를 구성한다.

5.1 가정

네이버와 코리아 야후에는 디렉토리 검색과 키워드 검색 두 가지를 지원하고, 미스 다찾니는 키워드 검색을 지원하고 있다. 디렉토리 검색은 사용자가 일일이 디렉토리를 찾아가야 하는 반면 URL에 접근 신뢰도가 높은 편이며, 키워드 검색은 사용자가 간단하게 검색 할 수 있는 반면 접근 신뢰도가 낮은 편이다. 실험에서는 세 가지 검색 엔진이 모두 지원하고있는 키워드 검색 방법을 한다.

검색 용어는 가변성 및 시사성을 고려하여 결정해야 하는데, 그 이유는 가변성의 유무는 URL의 수정 여부를 결정하고, 시사성의 여부는 URL의 삭제 여부를 결정할 수 있기 때문이다. 실험에서의 검색 용어는 시사성 및 가변성의 유무에 따라 한글, 시티폰, 삼풍백화점, 엑스포, 전자상거래 다섯 가지로 한다. 여기서 한글은 비가변성 용어이며, 나머지 네 용어는 가변성 용어로 간주한다. 가변성 용어중 시티폰과 삼풍백화점은 이미 과거의 용어이므로 비시사성 용어이고, 엑스포와 전자상거래는 시사성 용어로 간주한다.

5.2 실험

정보 검색 엔진 네이버, 코리아 야후, 미스 다찾니에 대한 결과 그래프는 그림 6, 7, 8과 같으며 그래프에서 검색 용어는 다음 표와 같이 사용된다.

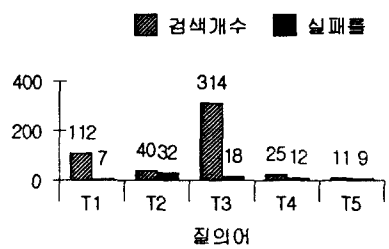
[표 1] 검색 용어

	T1	T2	T3	T4	T5
	한국	시티폰	삼풍백화점	전자상거래	엑스포

그래프는 각 용어에 대해 총 검색 된 개수와 검색 결과에 대해 접근 했을 때 실패율을 나타내고 있다.

① 네이버

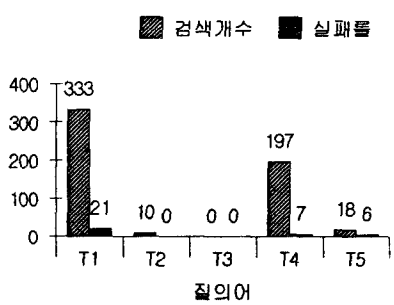
검색 결과 네이버는 각 용어들에 대해 평균 15.6%의 실패율을 보이고 있다.



[그림 6] 네이버 검색 결과

② 코리아 야후

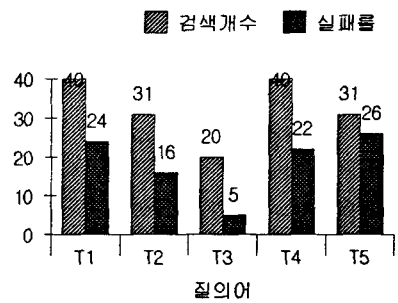
검색 결과 코리아 야후는 각 용어들에 대해 평균 4.7%의 실패율을 보이고 있다.



[그림 7] 야후 검색 결과

③ 미스 다찾니

검색 결과 미스 다찾니는 각 용어들에 대해 평균 18.6%의 실패율을 보이고 있다.



[그림 8] 다찾니 검색 결과

5.3 평가

실험 한 검색 엔진들 중에서 실패율이 가장 낮은 것은 코리아 야후이며, 실패율이 가장 높은 것은 미스 다찾니이다. 야후는 검색 용어 시티폰에 대해 실패율이 0%였으며, 검색 용어 삼풍백화점에 대해서는 검색 결과가 0였다.

각 검색 용어에 대한 실패율은 한글 17.3%, 시티폰 24%, 삼풍백화점 11.5%, 전자상거래 13.7%, 엑스포가 18.2%이다.

실패율의 요인들 중에는 "DNS entry"에 대한 에러율이 33.8%였고, 서버 에러가 30.4%, 파일을 찾을 수 없는 경우가 33.1%, 파일에 데이터를 가지고 있지 않은 경우가 2.7%였다.

비록 3개의 검색 엔진과 5개의 용어에 대해 제한적인 실험이 되었지만, 구현한 메타 검색 엔진은 평균 12.86%의 검색 결과 접근 신뢰도를 향상 시켰다.

6. 결론 및 향후 과제

평가에서도 볼 수 있듯이 정보 검색 결과의 실패 요인들을 보면 서버에 문제가 있거나 파일을 찾을 수 없는 경우가 많다. 이러한 사항들을 미리 사용자에게 알려주고 검색 결과의 접근 신뢰도를 향상 시키기 위해 로봇 에이전트를 이용한 빈 연결을 제거하는 메타 검색 엔진을 멀티 쓰레드를 이용하여 구현하였다.

또한 실험을 통하여 검색 용어에 성격에 따른 기존 검색 엔진들에 대한 검색 결과의 접근 실패율을 알아보고 구현 된 메타 검색 엔진이 문제점들을 개선하였음을 보았다.

본 논문에서 해결한 문제점은 앞으로 정보 검색 분야에서 많은 검색 엔진들이 지향해야할 바이며, 시소러스등을 이용한 문서의 내용 분석등을 통해 사용자에게 더 많은 정보를 제공해야 한다.

참고문헌

[1] 도신희, 곽재창, "개인용 인터넷 메타정보검색기의 설계 및 구현", 한국정보처리학회 추계 학술발표논문집, 제4권, 제2호, pp.915~919, 1997.

[2] 심해청, 김병만, 김태남, "효율적인 웹 로봇의 설계 및 구현에 관한 연구", 한국정보과학회 가을 학술발표논문집, 제24권, 제2호, pp465~468, 1997.

[3] 정동열, 전서현, "메타 검색 엔진을 이용한 사용자 인터페이스 에이전트의 구현", 한국정보과학회 봄 학술발표논문집, 제25권, 제1호, pp.503~505, 1998.

[4] 장명욱, 이강찬, "웹로봇과 정보추적자", 마이크로소프트웨어, p268~296, 정보시대, 1996.10.

[5] HTTPClient 패키지 매뉴얼 URL "http://www.innovation.ch/java/HTTPClient/api/Package-HTTPClient.html"

[6] 까치네 URL "http://www.kachi.net"

[7] 깨비 홈페이지 URL "http://www.kebi.com"

[8] 네이버 URL "http://www.naver.com"

[9] 미스 다찾니 URL "http://www.mochanni.com"

[10] 코리아 야후 URL "http://www.yahoo.co.kr"

[11] URL-minder URL "http://www.netmind.com/URLminder/URL-minder.html"