2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

MySQL과 PHP를 이용한 Internet 가격 비교 검색 엔진의 설계 및 구현

하은용, 정명교
안양대학교 컴퓨터학과
e-mail:eyha@aycc.anyang.ac.kr, mkjung@cse.anyang.ac.kr

Design and Implementation of a Comparative Price Search Engine Using MySQL and PHP

Eun-Yong Ha, Myung-Gyo Jung
Dept of Computer Engineering, An-Yang University

요약

인터넷 사용의 급격한 증가와 방대한 자료로 인하여 검색엔진에 대한 요구가 높아지고 있으며, 인터넷을 통한 쇼핑이 확대됨에 따라 가격에 대한 정확한 검색과 필터링이 불가피하게 되었다. 현재 정보를 찾기 위한 많은 검색엔진이 존재하지만 실제로 사용자가 필요로 하는 정확한 정보를 찾아주지는 못하고 있다. 따라서 특화된 검색엔진이 필요하게되고, 이로 인해 가격비교 검색엔진이라는 특화된 비교 검색엔진을 제안한다. 구현에 사용된 데이터베이스는 MySQL이며 스크립트 언어는 PHP이다.

1. 서론

최근 인터넷 발전과 더불어 웹사이트들이 포탈서비스 중심으로 경쟁이 심화되면서 웹 검색 시스템 시장에 참여하는 업체도 많아졌고, 기존의 웹 검색 시스템간에도 치열한 경쟁 상황이 전개되고 있다. 하지만 고객이 인터넷상에서 쇼핑을 하기 위하여 각각의 사이버 쇼핑몰들을 찾아다녀야 하고, 해당 사이버 쇼핑몰들을 일일이 기억하여야 하는 부담이 있다. 이것은 고객에게 전통적 방법으로 물건을 고르고 사는 것과 유사하며, 고객들이 사이버 공간에서 쇼핑을 하는데 상당한 부담을 주고 있다.

이 같은 부담을 덜어주기 위하여 인터넷상에서 복수의 쇼핑몰을 중개해주는 역할을 하는 쇼핑몰 형태인 'Meta 쇼핑몰' 즉, 'Mall of Malls' 라는 개념이 등장했으나, 이 역시 쇼핑몰들의 해당 웹사이트를 단순히 나열해 주기만 할 뿐 고객이 원하는 단일 상품에 대한 직접적인 정보를 제공해 주지는 않는다. 따라서 가격 비교 검색엔진이라는 개념을 도입하여 고객들이 검색엔진을 통하여 쉽게 찾고자 하는 물건을 가장 저렴한 가격으로 빠르게 살수 있도록 하기 위해 이 논문을 제안한다.

2. 관련연구

2.1 HTTP 프로토콜 [1]

HTTP는 HyperText Transfer Protocol의 약자이며, 분산환경 및 공동작업 환경에 이용할 하이퍼미디어 정보시스템의 개발을 목적으로 설계된 응용계층의 프로토콜로서 WWW에서의 하이퍼텍스트 문서의 전송을 위해 쓰이는 프로토콜이란 뜻이다.

HTTP 프로토콜은 요구/응답 (Request/Response) 방식을 이용하여 동작하고 있다. 즉, 원하는 프로토콜 기능(예: GET, HEAD, POST)에 대해 서비스 요구를 하면 데이타 송수신을 위한 TCP 연결이 만들어지고, 서버가 응답을 보내어 데이타 전송을 끝내면 자동적으로 연결이 끊어지게 되는 것이다.

2.2 웹 로봇

웹 로봇이란 자동적으로 웹의 하이퍼 텍스트 구조를 따라 다니며 문서를 가져와 검색하고, 그 문서에서 참조한 다른 문서들을 계속해서 가져와 검색하는 동작을 반복하는 프로그램이다.

2.3 검색엔진

검색 엔진이란 어떤 자료집합에서 검색을 수행해

주는 프로그램이다. 웹 분야의 경우에서는 흔히 로봇에 의해 모아진 HTML 문서 데이타베이스로부터 검색을 해주는 것을 말한다.

2.4 PHP [2]

PHP(HyperText PreProcessor)는 서버에서 해석되는 스크립트 언어이다. 이런 서버용 언어는 C나 Perl 등으로 만들었던 CGI 프로그램과 마찬가지로 서버에서 해석되고 그 결과만을 HTML 형태로 만들어서 클라이언트로 보내주기 때문에 내부 소스코드를 볼 수 없어 보안상으로 상당한 장점을 가지고 있다. PHP의 언어 해석과정은 그림1과 같다.

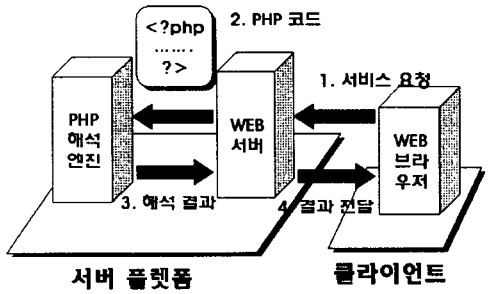

그림1. PHP의 해석과정

2.5 MySQL [3]

MySQL은 간단히 데이터베이스를 만들고 사용할 수 있게 해주는 프로그램으로 '터미널 모니터' 혹은 간단히 '모니터'라고도 한다. MySQL은 대화식 프로그램으로서 서버에 연결하고, 질문을 수행하고, 결과를 화면에 보여주는 일을 한다.

MySQL은 커널 thread를 이용한 Multi thread를 지원하므로 CPU가 여러 개일 경우 이들 CPU를 잘 활용할 수 있고 다양한 플랫폼을 지원하며 아주 큰 데이터베이스도 다룰 수 있다. 또한 다양한 언어로 에러 메시지를 표시하고 속도도 상당히 빠른 편이다.

3. 가격 비교 검색엔진의 구현

3.1 구현환경

구현에 사용된 서버는 일반 PC로 운영체제는 안정성이 검증된 Linux 6.0 이고, 웹 서버는 Apache를 사용하였다. DBMS는 비교적 규모가 작은 MySQL을 사용했고, 서버에서 해석되는 스크립트 언어는 PHP4인 Zend가 나왔지만 PHP3를 사용했다. 웹 로봇은 C로 만들었으며 사용된 컴파일러는 gcc이다.

3.2 검색 시스템 개요

본 논문은 웹 로봇으로부터 얻은 데이터를 데이터베이스에 저장하여 고객이 원하는 정확한 정보를 주는데 목적이 있다. 따라서 웹 로봇은 반드시 필요한 가격과 상품이름, 상품 URL, 간단한 설명 등을 파일로 저장시킨다. PHP 프로그램을 통해서 파일에 있는 내용을 중복 없이 데이터 베이스에 저장시키고 like 검색과 %(와일드카드)문자를 이용해서 필요한 정보를 검색한다. 검색된 정보는 가격이 싼 순서대로 정렬을 하며 카테고리를 이용하여 체계적으로 보여준다.

3.3 웹 로봇의 정보 추출 절차

1. index.html문서를 가져온다.
   ex) webrobot icoda.co.kr
2. index.html문서에서 최상위 분류를 추출한다.
   ex) shophtml index.txt > topclass.txt
3. 세분류 제품정보를 갖고 있는 문서들을 가져온다.
   mkdir tmp
   mkdir tmp/goods
   gchunk icoda.co.kr topclass.txt
4. 제품정보 html화일에서 제품정보(url, cost, item name, note)를 추출한다.
   ls -1 tmp > subclass.txt
   goodsdetail subclass.txt

결과는 tmp/goods 디렉토리에 각 세분류별로 정보가 파일로 저장된다.

▶ index 파일 가져오는 메인 루틴

```
main (int argc, char *argv[])
{
  .....중략
  connect_webserver(); //지정된 웹서비와 연결 설정
  send_GET_request(html_url);  //문서요청
  read_GET_response(html_url); //문서받기
  disconnect_server(); // 연결 종료
  .....중략
}
```

▶ html 문서를 요청하는 서브루틴

```
send_GET_request(char *url)
{
  .....중략
  sprintf(get_mesg,"GET %s HTTP/1.0\r\n\r\n",url);
  .....중략
}
```

▶ 소켓을 통해 원하는 문서 가져오기

```
read_GET_response(char *html)
{
  char buf[LEN];
  FILE   *fp;
  char *a, c='<';
  fp = fopen(storefile,"w");
  while (fgets(buf,LEN,infp) != NULL) {
      if(a=strchr(buf, c))
      fprintf(fp,"%s",a);
  }
    fclose(fp);
}
```

웹로봇의 동작원리는 그림2와 같이 먼저 index.html을 요청하고, 이것을 통해서 세부 문서들을 가져오게 된다. 하나의 쇼핑몰이 끝나면 다음의 쇼핑몰에서 했던 작업을 반복한다.

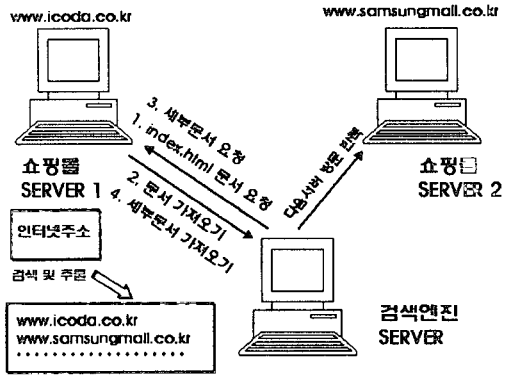

그림2. 웹 로봇 작동원리

3.4 검색엔진의 동작원리

검색엔진은 웹 로봇이 가지고 온 데이터를 데이터베이스에 저장하고, 이것을 중복이 없이 유지하며 검색어를 통해서 필요한 정보를 얻을 수 있게 한다. 본 논문에서 사용된 검색법은 어절 검색법이 아닌 와일드카드 문자 검색법이다. 어절 검색법으로 검색하면 검색어와 똑같은 것만 찾지만 와일드카드 문자 검색법은 비슷한 단어도 찾기 때문에 고객이 필요로 하는 정보를 더 정확히 찾아 준다. 그리고 카테고리를 사용하여 체계적인 데이터를 분리하여 검색어를 입력하지 않아도 쉽게 찾고자 하는 정보를 찾을 수 있다. 다음 그림3과 그림4가 카테고리를 사용한 디렉토리 구조의 형식을 잘 보여주고 있다.

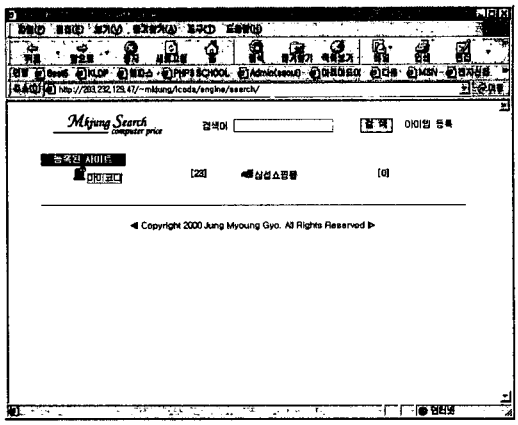

그림3. 카테고리 상위 디렉토리 구조

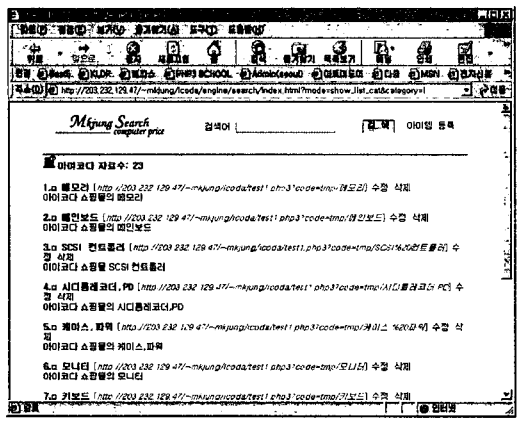

그림4. 카테고리 하위 디렉토리 구조

3.5 데이터베이스 구조

정보를 저장하기 위한 data와 data1 테이블의 구조는 다음과 같다.

```
create table data (
    url varchar(60),    // URL 저장
    goodstitle varchar(50), // 상품이름 저장
    content text,       // 상품설명 저장
    price varchar(10), // 가격 저장
    fname varchar(20) // 파일 이름
);
```

본 검색엔진은 2개의 데이터 테이블을 이용해서 항상 중복이 없는 데이터 베이스를 유지한다. 웹 로봇이 파일시스템으로 데이터를 저장한 것을 PHP를 이용해서 data 테이블에 저장하고, data 테이블에 있

는 레코드가 data1 테이블에 존재하는지를 검색한 다음 레코드가 존재하면 중복을 없애기 위해서 중복된 레코드를 삭제하고 INSERT를 수행한다. 또한 data 테이블의 레코드가 data1 테이블에 존재하지 않으면 INSERT만 수행한다.

data1 : 실제 데이터 유지 테이블, data : 임시 테이블

```
<?php
// 데이터베이스 연결 부분
$conn=mysql_connect("HOST","DB_USER","DB_PASS");
mysql_select_db("DATABASE_NAME",$conn);
// 실제 SQL문 처리 부분
$result1=mysql_query("select * from data");
while($result2=mysql_fetch_row($result1))
{
$flag=mysql_query
      ("select * from data1 where url='$result2[0]'");
if($flag) {
mysql_query
("delete from data1 where url='$result2[0]'");
}
mysql_query("insert into data1 values
            ('$result2[0]','$result2[1]','$result2[2]',
             '$result2[3]','$result2[4]')");
}
?>
```

위에서 입력된 데이터들은 가격을 기준으로 정렬되어 화면에 보여 지게 된다. 그림5는 검색엔진의 작동원리를 나타낸 것이다.

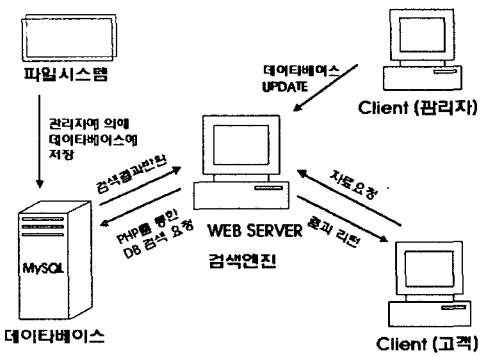

그림5. 검색엔진의 작동원리

4. 결론

본 논문은 일반 사용자들이 컴퓨터 가격을 알아보는데 있어 쉬운 인터페이스를 제공하여 빠르게 정보를 검색할 수 있는 환경을 제공한다. 카테고리를 이용하여 품목별로 데이터를 분류한 것과 여러 쇼핑몰에 대한 정보를 가지고 있는 것이 장점이다. 하지만 웹 로봇이 가지고 온 데이터를 데이터 베이스에 저장하는 부분이 자동화되지 않았고, 쇼핑몰에서 로봇의 접근을 막아 놓으면 데이터를 가지고 올 수 없다는 것이 단점이다. MySQL과 PHP의 사용으로 빠른 검색과 신뢰성 있는 검색을 할 수 있다. 하지만 지식기반 검색과 자연어 검색을 도입하는 것 및 모든 작업을 자동화하는 것이 향후 과제이다.

참고문헌

[1] HTTP RFCs, "HTTP Specifications and Drafts", http://www.w3.org/Protocols/Specs.html

[2] PHP 온라인 메뉴얼, "PHP Manual Online", http://www.php.net/manual/

[3] MySQL 온라인 메뉴얼, "MySQL Manual", http://www.mysql.com/documentation/mysql/commented/