韓國 OA學會 論文誌
第7卷 第3號, 2002. 9.

2002-7-3-3-3

사용빈도 높은 웹정보의 자동 검색

이 범 근* 송 호 정**

Automatic Retrieval Of Frequently-Used Web Information

Beom-geun Lee* Ho-jeong Song**

요 약

인터넷을 사용하는 가장 중요한 목적중의 하나는 새로운 정보를 신속하게 얻는 것이다. 예를 들면, 수동으로 접속해야 하는 어떤 웹사이트에서 주기적으로 정보를 얻기를 원하는 것은 많은 시간과 노력이 필요하므로 비효율적이다.

그러므로, 본 논문에서는 인터넷상의 지정된 웹사이트에서 정보를 자동으로 검색하고 서버에 저장하여 이를 필요에 따라 서비스 해주는 시스템을 제안하였다. 본 시스템은 웹 정보를 모으고 그것을 효과적으로 제공하기 위하여 JSP에 의하여 구현되었다.

Abstract

One of the most important purposes of using Internet is to get new informations rapidly. For instance, we want to get information on some web sites regularly, the web sites should be manually accessed several times a day. It is not efficient because a lot of time and effort are necessary.

Thus, this thesis proposes a system which automatically gets information on the designated sites on the Internet and keeps the information in the server and provides it to anybody who needs it. The system was implemented by means of JSP to collect web informations and provide it efficiently.

* 경희대학교 전자공학과 박사과정
** 충북대학교 컴퓨터공학과 박사과정

논문접수 : 2002. 6. 9
심사완료 : 2002. 9. 14

Ⅰ. 서론

20세기말의 컴퓨터 기술의 발달과 통신 기술의 발달은 이 두 가지를 통합한 인터넷의 등장을 초래하였으며 이를 이용한 모든 부분의 엄청난 변화를 일으켰다. 그중에서도 신속하게 새로운 정보를 얻고자 할 때 인터넷을 많이 이용하게되었다. 학교나 회사, 또는 관공서에서 여러 웹사이트에서 정기적으로 정보를 얻을 경우 지금까지는 일반적으로 직원이 지정된 웹사이트들에 매일 접속해서 새로운 정보를 확인해 왔다. 이것은 인적인 면이나 시간적인 면에서 낭비를 초래한다.

본 연구에서는 위와 같이 자주 사용되는 인터넷 웹사이트의 정보를 자동으로 검색하여 서버에 저장하고 이를 필요에 따라 서비스해주는 시스템을 목표로 하고 있다. 즉 특정 웹사이트의 URL을 저장하여 관리하면서 이들을 매일 접속하여 새로운 정보나 변화가 있으면 시스템 서버의 데이터베이스 에 저장하여 관리하며 저장된 정보를 다수의 사용자에게 서비스하는 시스템을 구현하고자 한다.

Ⅱ. 인터넷 환경에서의 정보수집

1. 일반적인 정보수집 방법

HTTP(Hyper Text Transfer Protocol)는 클라이언트와 서버의 분산 모델을 기초로 하는 요청/응답 페러다임중 하나이다. HTTP는 클라이언트와 서버 사이에서 메시지를 주고받는다. 주로 클라이언트는 요청하고 서버는 응답한다.

즉, 새로운 정보가 웹사이트에 올라와도 갱신된 시점 뒤에 사용자가 해당 사이트를 방문하지 않으면 갱신 여부를 알 수 없다. 따라서 사용자는 어떤 사이트가 갱신되었을지 모르기 때문에 자신이 자주 방문하는 모든 사이트를 주기적으로 방문할 수밖에 없다.

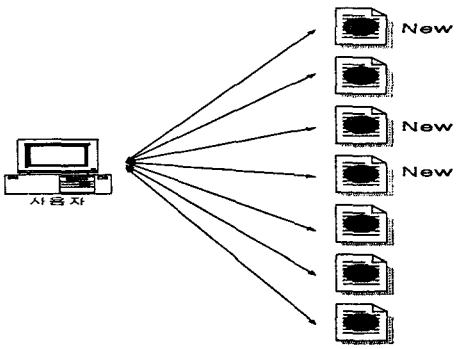


그림 2. 일반적인 정보검색과정
Fig. 1 Progress of general information retrieval

그림 1은 일반적인 정보검색 과정을 나타내고 있다. 사용자는 어떤 사이트에서 언제 정보가 갱신될 지 모르기 때문에 꾸준히 다른 모든 사이트를 방문해야 한다. 특히 시간을 촉박하게 다투는 중요한 정보라면 하루에도 수십차례 계속 방문을 해야만 한다. 어느 시간대에 정보가 갱신될지 모르기 때문이다. 따라서 이러한 시간 의존적인 정보들은 별도의 어플리케이션을 이용하여 실시간으로 제공되기도 한다.

2. 자동 시스템을 통한 정보검색 방법

HTTP는 비지속형 프로토콜이기 때문에 정보를 수집하기 위해서는 어쩔 수 없이 웹페이지를 지속적으로 로딩해야만 한다. 기존에는 각 사용자들이 이러한 작업들을 해왔지만 본 논문에서는 서버에서 해당 사이트의 정보를 처리하고 저장하며 서비스하는 방식을 제시한다.

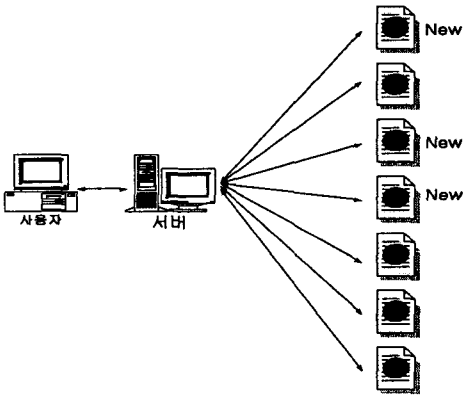


그림 3. 자동화된 정보검색 방법
Fig. 2 progress of automatic information retrieval

그림 2의 경우 사용자는 서버에 한번 접속하는 것만으로 3곳의 새로 갱신된 사이트의 정보를 받아볼 수 있다. 물론 이때 사용자가 방문해야할 여러 곳의 사이트를 서버에서 자동으로 방문해주는 것이기 때문에 네트워크 전체적인 트래픽에 있어서는 동일한 결과를 가져온다. 하지만 사용자가 한 명이 아니고 여러 명일 경우, 정보를 얻고자 하는 사이트가 사용자들간에 중복될 확률이 높다. 왜냐하면 본 시스템을 사용하게 될 회사나 기업, 관공서의 성격이 비슷하고, 얻고자 하는 정보나 해당 사이트가 비슷하기 때문이다. 따라서 부처별로 매번 별도 방문을 할 경우보다 서버가 모든 사이트를 방문한 후 모든 해당 부처에 정보를 보내주기 때문에 네트워크 전체적인 트래픽은 상당히 감소하게 된다.

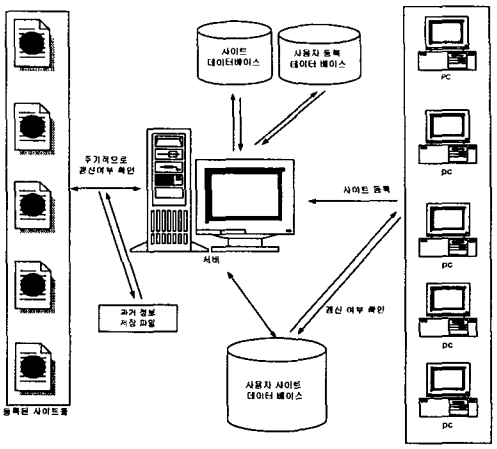


그림 3. 전체적인 시스템 구성도
Fig. 3 System organization

Ⅲ. 웹 자동 검색 시스템의 설계 및 구현

1. 전체 시스템 구조

본 논문에서 제시된 시스템의 작동 프로세서는 다음과 같다.

1) 사용자 등록을 마친 후 정보 갱신여부를 알기 원하는 사이트를 등록한다.
2) JSP에서 사용자가 등록을 원하는 사이트가 이미 타인에 의해 등록되어 있는지 여부를 판단하여, 적절한 데이터베이스에 사이트를 등록한다.
3) 웹 정보 수집 엔진은 새로 등록된 사이트 목록을 데이터베이스로부터 수신하여 현재 사이트 정보를 수집하고 새로운 과거 정보 파일을 생성한다.
4) 웹 정보 수집 엔진은 정해진 시간 간격으로 주기적으로 등록된 사이트를 조사하여 갱신 여부를 판단하고 사용자가 로그인 할 때 그 사실을 알려준다.
5) 사용자가 로그 아웃하면 JSP는 이미 갱신된 것을 사용자가 본 것으로 판단하고, 데이터베이스에 갱신판단 여부 필드를 'old'로 변경한다.

2. ITC(Internet Transfer Control)

2.1 개요

웹 정보수집 시스템 개발에 있어서의 문제점은 각 정보를 제공하는 사이트들의 데이터베이스 포맷이 일정치 않을 뿐 아니라 사용자 임의의 접근을 막기 위한 보안설정이 존재한다는 것이다.

초기 정보수집 시스템에서는 웹의 정보수집을 위해 웹 로봇 에이전트를 설계하여 사용할 것을 고려한 바 있었다. 하지만 웹 로봇 에이전트는 이미 웹 상에서 포화상태에 이르러 네트워크 트래픽을 방해하는 요소로 인지되고 있으며 이러한 문제점 때문에 각 사이트 설계자들은 웹 로봇 에이전트의 방문을 거절하는 코드를 삽입하여 사이트를 설계하는 추세이다 [1,2,6].

따라서 본 논문에서는 ITC(Internet Transfer Control)를 이용해서 정보를 담고있는 사이트의 HTML 코드 전체를 당겨오는 방식을 제시하고자 한다.

일반적인 경우 각 사이트의 데이터베이스가 임의의 접근을 막고 있다 하더라도 정상적인 루트를 통해 사이트를 보는 경우 데이터베이스의 정보가 자동으로 HTML문서로 생성되어 보여지는 것이므로 굳이 데이터베이스에 액서스 하지 않아도 생성된 HTML문서 내부에서 필요한 정보만 파싱(parsing)하여 추출하면 된다.

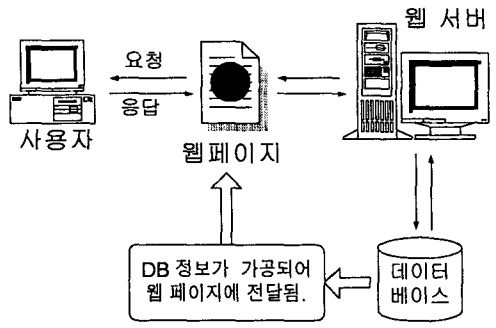


그림 4. 웹서비스에서의 자료흐름
Fig. 4 Dataflow of web service

그림 4는 일반적인 웹서비스에서의 자료흐름을 보여주고 있다. 사용자가 해당 사이트를 통해 얻고자 하는 정보는 서버의 데이터베이스에 저장되어 있다.

3. 웹 정보 수집 엔진

웹 정보를 주기적으로 수집하고 갱신 여부를 판단하는 역할을 웹 정보 수집 엔진에서 처리한다. 본 논문에서 제안하는 웹 정보 수집 엔진은 비주얼 베이직으로 제작되었고, 윈도우즈 OS를 사용하는 곳에 배포, 설치할 수 있도록 패키지로 제작되었다. 웹 정보 수집 엔진은 크게 수집된 정보를 파일로 저장하는 부분, 그리고 조사 사이트 목록 획득과 갱신 여부를 저장하기 위한 데이터 베이스 접속 부분으로 나뉜다.

3.1 사이트 정보 저장

본 논문에서는 사이트의 정보를 데이터 베이스에 저장하는 것이 아닌, 파일 I/O를 사용하여 각 사이트의 고유한 코드 이름을 사용하여 파일로 생성, 관리하는 방법을 사용하였다. 파일 생성에 대한 프로세서는 [그림 7] 과 같다.

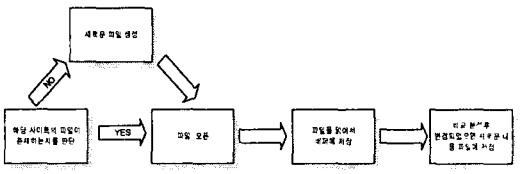


그림 5. 정보수집 엔진에서의 파일 생성 프로세스
Fig. 5 File generation process

3.2 RDO(Remote Data Object)

웹 정보 수집 엔진은 본 논문에서 사용되는 여러 데이터 베이스 테이블 중에서 사이트 데이터 베이스와 사용자 사이트 데이터 베이스에 액서스 한다. 이들 데이터베이스에 접근하기 위해 RDO를 사용하였다.

RDO는 Remote Data Object의 약자로써, 우리말로는 원격 데이터 개체로 해석할 수 있다. 즉, RDO는 말 그대로 원거리에 있는 데이터베이스에 접근하여 원하는 데이터를 액세스하고 이를 제어할 수 있는 개체를 말한다. 다시 말해서 RDO는 원격 ODBC(Open DataBase Connectivity) 데이터베이스 시스템의 컴포넌트를 생성하고 조작하기 위한 코드를 사용하는 프레임웍을 제공한다.

RDO는 마이크로소프트사에서 탄생시킨 개념인데, 이와 유사한 것으로 DAO(Data Access Object)라는 것도 있다. 이것 역시 ODBC에 접근하여 데이터베이스와의 연결성을 확보하기 위한 목적으로 사용되는 것이다. 본 논문에서 DAO를 사용하지 않고 RDO를 사용한 이유는 시스템의 구축시 DAO는 데이터베이스가 동일 시스템에 구축되어 있어야 하지만, RDO의 경우 원격 제어가 가능해, 데이터베이스와 서버와의 이원화 시스템을 구축할 수 있기 때문이다. 본 논문에서는 Microsoft Remote Data Object 2.0을 사용하였다.

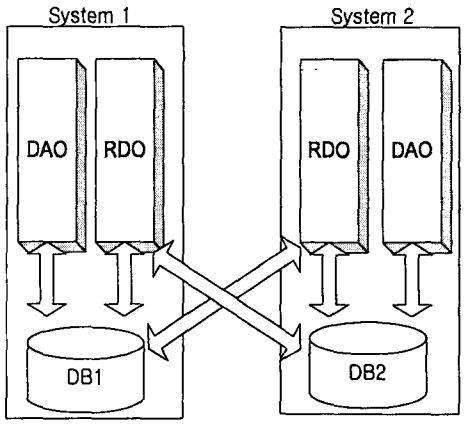


그림 6. RDO와 DAO의 개념도
Fig. 6 Conception of RDO and DAO

3.3 웹 정보 수집엔진의 실제

웹 정보 수집 엔진 실제 작동 모습은 [그림 7]과 같다.

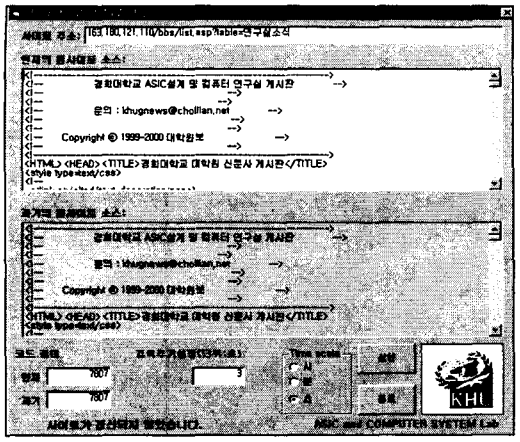


그림 7. 웹 정보 자동수집 엔진의 실제모습
Fig. 7 Actuality appearance

사이트 주소창에는 현재 정보 수집중인 사이트 주소가 나타난다. 이 주소는 정보수집을 마치면 다음 주소로 변한다. 현재의 웹사이트 소스 창에는 검색하는 순간의 실제 웹서비스 되고 있는 홈페이지의 소스가 나타난다. 과거 웹사이트 소스 창에는 기존에 저장되어있던 동일 홈페이지의 과거 소스가 나타난다. 검색주기 설정 창에는 원하는 검색 주기를 숫자로 입력하고 time scale을 설정하면 해당 시간이 지날 때마다 한번씩 웹 정보를 수집하게 된다.

과거 사이트 내용과 현재 사이트 내용을 비교하는 방법에는 크게 2가지가 있다.

1) 전체 코드 길이 비교

저장되어 있는 사이트 소스코드의 전체 길이를 조사하여 현재 사이트와 비교하는 방식이다. 사이트의 내용이 변경되었다면 소스코드의 길이가 변하는 것이 일반적인 경우이므로 소스코드의 길이를 비교하면 쉽게 변경 유무를 알아낼 수 있다. 이 방식은 비교의 루틴이 비교적 간단해서 비교속도가 빠르다는 장점은 있지만 소스코드의 길이를 그대로 유지하면서 변경되는 정보에 대해서는 감지할 수 없다는 단점이 있다.

2) 문자열 비교

이 방식은 저장되어 있던 소스코드와 현재 서비스되고 있는 웹사이트의 소스코드의 모든 문자를 비교, 분석하는 방법이다. 이 방식의 특징은 문자열의 길이를 유지하면서 변경되는 경우도 감지할 수 있다는 것이다. 하지만 글자를 한 글자씩 모두 비교해서 분석하는 문자열 비교 방법에서는 그러한 경우의 변경도 감지해 낼 수 있다. 하지만 소스코드 길이 비교의 방법보다는 비교시간이 오래 걸린다는 단점이 있다.

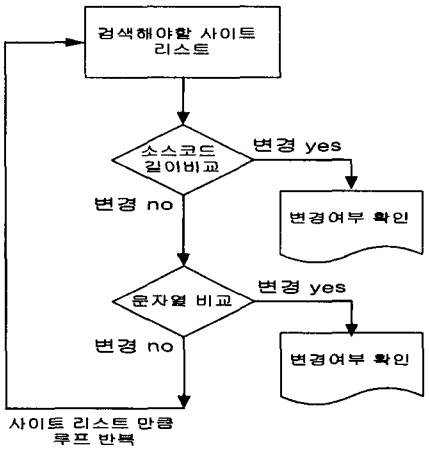


그림 8. 과거 사이트 정보와의 비교 과정
Fig. 8 Comparison process

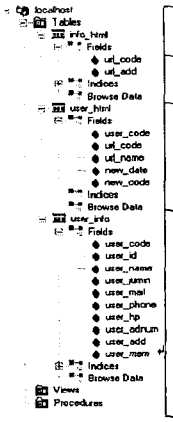
따라서 본 논문에서는 일차적인 검사로써 소스코드의 길이비교를 하고, 이차적 검사로 문자열비교 검사를 시행한다. 이렇게 함으로써 대부분의 변경 사이트는 소스코드 길이 비교로 빠른 시간내에 걸려내고, 그 검사를 통과한 일부 사이트에 대해 다시 한번 문자열 비교를 시행하여 완벽하게 변경 유무를 체크하게 된다. 그림 8은 검색 과정을 보여준다.

4. DBMS(Data Base Management System)

본 논문에서 사용된 DBMS는 MySql이다. MySql은 현존하는 DBMS중에서 완전 무료로 사용할 수 있는 몇 안되는 DBMS중 하나이다. MySql의 특징은 관리하기가 편하고, 고사양의 시스템을 요구하지 않는다는 것이다.

본 논문에서는 그다지 복잡한 DB구조를 다루지 않기 때문에 무료이고, 사용하기도 편리한 MySql을 채택하였다.

그림 9는 본 시스템의 데이터베이스 테이블 구조이다. 데이터베이스 테이블은 크게 사이트 리스트를 저장하는 info_html과 사용자 사이트 리스트를 저장하는 user_html, 그리고 각 사용자의 신상정보를 저장하는 user_info 테이블이 존재한다.



Tables	Fields	Data type	Primary Key
info_html	url_code	char(10)	url_code
	url_add	varchar(255)	
user_html	user_code	char(10)	
	url_code	varchar(255)	
	url_name	varchar(25)	
	new_date	char(25)	
	new_code	char(5)	
user_info	user_code	char(10)	user_code
	user_id	char(8)	
	user_name	varchar(15)	
	user_jumin	char(13)	
	user_mail	varchar(20)	
	user_phone	varchar(20)	
	user_hp	varchar(15)	
	user_adnum	char(6)	
	user_add	varchar(50)	
	user_men	varchar(30)	

그림 9. 데이터베이스 테이블 구조
Fig. 9 Database table structure

그림 10은 갱신여부를 판단하여 출력되는 결과 화면이다.

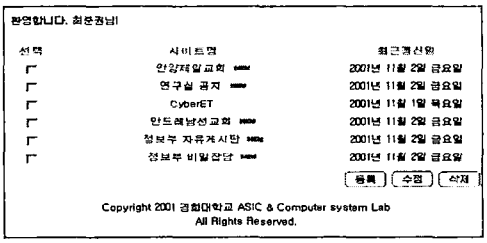


그림 10. 실제 시스템 작동 결과화면
Fig. 10 System operation result picture

Ⅳ. 결론

대부분의 공공 기관이나 연구기관 내지는 유,무형의 단체들도 조직내의 의사소통이나, 대내외적으로 알리고자 하는 소식들을 여러 가지 방법을 이용하여 전달하고 있으며 그 방법 중에는 반드시 인터넷이라는 매체도 포함되어 있다. 뿐만 아니라 개인적인 생활에서도 인터넷은 중요한 정보수집원으로 자리잡아가고 있다. 각종 상거래, 다양한 이벤트, 심지어는 회사의 구인 정보 등 실생활에 유익한 수많은 정보들이 매초마다 쉴새없이 쏟아져 나오고 있다. 이 모든 정보를 습득하고 위해서 많은 사람들은 중요 사이트를 북마크(book mark)해두고 자주 사이트를 방문하여 새로운 정보를 찾아다닌다.

하지만 본 논문에서 제시하는 시스템을 사용하면 더 이상 사용자가 일일이 사이트를 방문할 필요가 없어진다. 정보를 수집하기 위해 방문해야만 하는 사이트를 입력만 시켜두면 단 한번의 웹서핑으로 원하는 정보를 모두 얻을 수 있기 때문이다. 이것은 매우 빠르고 효과적이며 능률적인 방법일 뿐만 아니라, 어느 장소에서도 자신이 자주 방문하는 사이트 주소를 외우지 않고 찾아갈 수 있는 서버 기반의 북마크의 기능도 하게 된다. 공공기관에서도 기관 전체의 시스템에 본 시스템을 활용한다면 각 부서마다 정보를 얻기 위해 동일 사이트를 반복해서 접속함에 따라 발생하는 트래픽을 대폭 감소시킬 수 있을 것이다.

참고문헌

[1] 박정훈, "인터네트 정보 자원 데이타베이스 구축 및 정보 발견 시스템 개발", 최종연구보고서, 시스템공학연구소, pp. 13-17, 1995. 7

[2] 박정훈, 조현성, 이강찬, 이규철, "인터네트 정보 발견 시스템의 개발 및 구현", 정보과학회 논문, pp. 12-13

[3] Joon Ho Lee, "Combining Multiple Evidence from Different Properties of Weighting Schemes," pp. 3-6,8, Cornell University

[4] 한국전자통신연구소, 기술정보센터, "멀티쓰래딩 기법과 동향(I, II)", 주간 기술동향 pp. 93-30,31, 1993

[5] Michael Wooldridge and Micholas R. Jennings, "Agent Theories, Architectures, and Languages: A Survey," In Michael J, Intelligent Agents, pp. 1-39, Springer-Verlag, Germany, 1995.

[6] 백은경, 김영환, "에이전트와 시스템 개발", 정보통신 연구, 제 9권, 제 2호 pp 98-110, 7월, 1995

저 자 소 개

이 범 근

1973년 2월 14일생

1995년 2월 청주대학교 전자공학과 졸업(공학사),

1997년 2월 청주대학교 전자공학과 졸업(공학석사),

2001년 8월 경희대학교 전자공학과 박사수료

2001년 3월~2002년2월 극동정보대학 전자통신과 겸임강사,

2002년3월~현재 극동정보대학 전산정보처리과 초빙전임강사

관심분야 : Micro Display HDTV, CAD, Internet Applications

송 호 정

1994년 배재대학교 물리학과 졸업(학사)

1996년 청주대학교 전자공학과 졸업(공학석사)

2001년 충북대학교 컴퓨터공학과 박사수료

2000년3월~현재 극동정보대학 전자통신과 겸임강사

관심분야 : VLSI 설계, High-level Synthesis, Genetic Algorithm