한국시뮬레이션학회 '00춘계학술대회 논문집 2000.4.15 영남대학교

사용빈도 높은 웹정보의 자동 검색

최준권*, 정연모*, 이학준**
경희대학교 전자공학과*, 엠아이넷주식회사**

AUTOMATIC RETRIEVAL OF FREQUENTLY USED WEB INFORMATION

Jun-Kwon Choi*, Yunmo Chung*, Hark-Jun Lee**

Dept. of Electronics Engineering, Kyung Hee University*
MINET Corp.**,

요 약

본 논문에서는 자주 찾는 인터넷의 웹 정보를 자동으로 검색하고 서버에 저장하여 이를 필요에 따라 서비스 해주는 시스템개발하였다. 푸쉬(Push) 기술과 웹 로봇 에이전트 기술을 이용하여 특정 웹사이트의 URL을 저장하여 관리하면서 이들을 매일 접속하여 새로운 정보나 변화가 있으면 시스템 서버(system server)의 데이터베이스 (database)에 저장하여 관리하며 저장된 정보를 미리 예약된 사용자에게 전자메일을 통해서 자동 전송하는 시스템을 개발하였다.

1. 서 론

20세기말의 컴퓨터 기술의 발달과 통신 기술의 발달은 이 두 가지를 통합한 인터넷(internet)의 등장을 초래하였으며 이를 이용한 모든 부분의 엄청난 변화일으켰다. 그중에서도 신속하게 새로운 정보를 얻고자 할 때 인터넷을 많이 이용하게 되었다. 학교나 회사, 또는 관공서에서 여러 웹사이트(web site)에서 정기적으로 정보를 얻을 경우 지금까지는 일반적으로 직원이 지정된 웹사이트들에 매일 접속해서 새로운 정보를 확인해 왔다. 이것은 인적인 면이나 시간적인 면에서 낭비를 초래한다.

본 논문에서는 위와 같이 자주 사용되는 인터넷 웹사이트의 정보를 자동으로 검색하여 서버에 저장하고 이를 필요에 따라 서비스해주는 시스템 개발 하였다. 푸쉬(Push) 기술과 웹 로봇 에이전트 기술을 이용하여 특정 웹사이트의 URL을 저장하여 관리하면서 매일 접속하여 이들을 새로운 정보나 변화가 있으면 시스템 서버(system server)의 데이터베이스 (database)에 저장하여 관리하며 저장된 정보를 미리 예약된 사용자에게 전자메일을 통해서 자동 전송하는 시스템을 개발하였다.

2. 푸쉬(Push) 기술

2.1 개요

대부분의 사람들은 인터넷의 컨텐츠(Contents)를 당겨음(pulling)으로써 원하는 정보를 얻는다. 즉, 사용자가 요청(링크에 대한 마우스 클릭)을 해야 인터넷의 컨텐츠가 사용자에게 보여지게 된다.

그러나 푸쉬(push)의 경우, 클라이언트가 사전에 지정한 컨텐츠가 서버에 준비되었을 때, 서버는 그것을 클라이언트에게 강제적으로 전달, 즉

푸쉬한다. 이 후, 사용자는 자신이 원할 때 이미 자동적으로 전달되어 있던 컨텐츠를 볼 수 있게 된다.

이러한 방식의 사용을 위한 기술을 "푸쉬 기술(Push technology)"이라 부르며, 실제로는 기존 풀링 방식의 단점을 보완하기 위한 반작용으로써 만들어진 기술이다.

2.2 푸쉬 기술의 특징

일반적인 Web 사용 방식과 클라이언트/서버 어플리케이션의 방식은 주로 정보를 필요로 하는 사용자가 정보에 접근하여 유효한 정보를 끌어가는 "PULL 방식"을 이용하는데 반해. PUSH 방식은 해당 사용자에게 유효하다고 판단되는 정보가 발생되면, 사용자에게 그 정보를 서버측에서 클라이언트로 밀어주는 방식을 택하고 있다. 따라서 현재 우리가 보통 사용하고 있는 "Web Navigation 방식"이나 "클라이언트/서버 방식"이 오래 전 일상 생활에서 쓰이던 방식이라면, PUSH 방식은 현재 일상적인 생활 중에 발생하는 수많은 정보 중에서 유효한 정보들만을 선별하여 받아들이는 보편적인 방식이며, 그 동안 컴퓨터와 네트워크의 한계로 인해 발생된 근대적인 정보 전달 체계로부터 21세기에 걸맞는 전달 체계로 일보 전진을 하게 된 것이라 할 수 있다.

푸쉬 기술을 사용함으로써 보다 능동적인 정보 서비스가 가능하며, 정보 수집 시간을 단축 가능하며, 중요 정보를 적기에 제공받지 못해서 발생될 기회 손실을 최소화할 수 있으며, 네트워크 자원의 절감이 가능하다.

2.3 푸쉬 기술의 구조

푸쉬 기술의 구현에는 크게 두가지 방식이 있다. 그 중의 하나는 클라이언트측의 전송 요구가 있은 후에 서버측에서 전송하는 기존의 "Pull 방식"의 확장형이 있으며, 또 다른 하나는 클라이언트측의 전송 요구 없이도 사전 등록 상황에 따라 서버측에서 지능적으로 판단하여 클라이언트로 전송해 주는 "Real PUSH 방식"이 있다.

본 연구에서는 "Pull 방식"을 채택하여 연구를 수행할 예정이다. "Pull 방식"은 푸쉬 클라이언트의 요구에 의해서 정보의 전송이 이루어지며 그에 대한 동작 구조는 그림 1 과 같다.

푸쉬 기술의 구현을 위해서는 크게 서버 소프트웨어와 클라이언트 소프트웨어, 그리고 사용자 정보와 웹 정보 컨텐츠를 저장, 관리할 데이터베이스로 구성된다. 이중 서버 소프트웨어와 클라이언트 소프트웨어의 기능은 [표 1]과 같다.

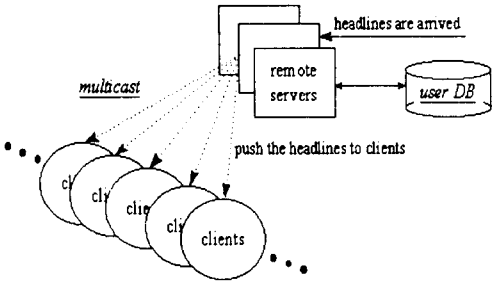

그림 1. 푸쉬 클라이언트의 동작

표 1 푸쉬에서 서버 S/W와 클라이언트 S/W의 기능

구 분	기 능
Server S/W	- 외부정보의 자동수신 - 사용자의 등록 및 관리 - 전송되어 온 정보의 자동분류 저장 - 사용자가 원하는 정보의 주기적 갱신 - 새로운 정보만 갱신하는 기능 - Log파일 생성 프로그램 및 Log 파일 - 분석프로그램 개발 - 서버와 클라이언트 S/W간에 데이터를 전송하는 기능
Client S/W	- 서버로부터 주기적으로 정보를 가져오는 기능 - 자체 인터넷 브라우저 채용 - 수신한 데이터를 보여주는 자체 프로그램 내장 - MS Explorer, Netscape Navigator 자동 연결 기능 - 해당 회사의 인터넷 홈페이지로 이동 기능 - 여러가지 접속기능(모뎀, 전용선) 제공

3. 웹 로봇 에이전트 기술

3.1 개요

로봇 에이전트란 웹을 순회하며 각 홈페이지들의 정보를 수집하는 프로그램이다. 웹서버에 접속하여 데이타(HTML 화일)를 가져오는(fetch)하는 기능적인 측면으로만 봐서는 웹브라우저와 같

은 기능을 하는 셈이다. 단지 웹브라우저는 가져온 데이타를 예쁘게 화면에 보여주고 화면에 하이퍼링크가 있고 사용자가 링크를 클릭하면 대항 홈페이지가 또 보이는 기능이 있는 것이고, 로봇 에이전트는 예쁘게 보여주는 대신 HTML을 분석하고 URL 부분을 추출하여 다른 URL로 접근하게 하는 기능이 있을 뿐이다. 자동적으로 홈페이지를 찾아다니므로 로봇 에이전트를 이용한다면 사람이 할 수 있는 일이지만 귀찮은 다양한 일을 할 수 있다.

로봇 에이전트는 웹페이지를 돌아다니면서 할 수 있는 다양한 일들을 자동적으로 해 주는데 의미가 있다. 현재 로봇 에이전트를 이용한 검색엔진도 마찬가지로 사람이 일일이 찾아다니면서 검색을 위한 인덱스를 만드는 것이 아니라 로봇 에이전트가 자동적으로 돌아다니면서 원하는 정보를 수집하는 것이다.

그림 2는 웹 로봇 에이전트의 동작 구조를 보여준다.

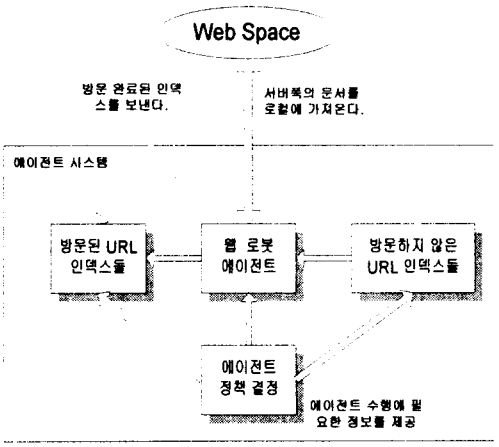

그림 2. 웹 로봇 에이전트 동작

4. 웹 자동 검색 시스템의 구조 및 구현

4.1 전체 시스템 구조

앞 장의 각 기술들을 사용하여 구현하게 될 시스템은 [그림 3]과 같이 개인 PC와 서버 그리고 사이트들을 관리하는 데이터 베이스로 구성된다.

서버의 기능은 데이터베이스에 등록된 사이트들을 주기적으로 체크하며, 사이트의 내용이 변경되었을 때에는 데이터베이스에 사이트의 내용을 등록한 뒤에 사용자의 PC에 사이트의 내용이 갱신되었다는 메시지를 통보하게 된다.

개인 PC에서는 원하는 사이트를 쉽게 서버의 데이터베이스에 등록하고, 갱신여부를 통보 받기 위해 서버와 통신할 수 있는 소켓프로그램이 설치되어야한다.

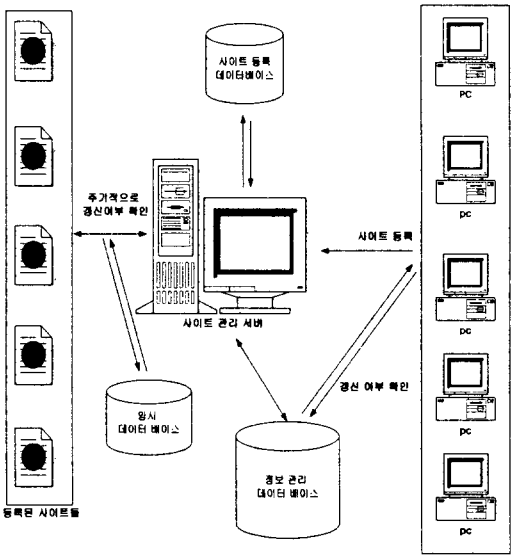

그림 3 전체적인 시스템 구성도

데이터 베이스에는 [표 2]와 같은 테이블로 구성된다. VARCHAR2는 오라클 8i에서 가변길이 문자열을 저장하기 위한 데이터 타입으로 최대 4000바이트까지 저장할 수 있다. NVARCHAR2는 VARCHAR2와 유사한 타입으로 다양한 국가 언어를 지원한다. 이밖에 DATE는 날짜와 시간을 나타내는 데이터 타입이고 CLOB는 구조화되지 않은 문자 데이터를 최대 4GB까지 저장할 수 있는 데이터 타입이다.

표 2 데이터베이스 테이블

사이트 이름	사이트 주소	날짜	시간	contents 제목	contents 내용
NVAR CHAR2	VARC HAR2	DATE	DATE	NVARCH AR2	NVARC HAR2

데이터 베이스 테이블에 대해 구체적으로 설명하면 사이트 이름에는 정보를 가져오게 되는 사이트 이름이 들어간다. 사이트 주소 위치에는 정보를 가져온 사이트의 위치 정보가 수록되고 날짜와 시간란에는 정보를 가져온 당시의 시간을 기록하게 된다. contents 제목에는 가져온 정보의 내용을 개괄적으로 알 수 있는 정보가 수록된다. 이 정보는 게시판이나 등록물의 제목이 될 것이

다. 마지막으로 contents 내용에는 사용자가 필요로 하는 갱신 정보가 수록된다.

전체 시스템 구성은 다음과 같은 구조를 가진다.

1)사이트 관리 서버 프로그램

- 임시 데이터 베이스에 액서스하여 검색 목록 사이트의 내용과 비교, 분석하고 갱신 여부를 판단한다.
- 사이트 등록 데이터 베이스에 액서스하여 정해진 사이트를 일정 시간 간격으로 조사하고 클라이언트에서 새로운 사이트 등록 요청이 들어왔을 때 이를 받아들여 사이트 등록 데이터 베이스를 수정한다.
- 갱신된 내용이 있을 경우 관련 내용을 정해진 데이터 베이스 테이블에 따라 정보 관리 데이터 베이스에 저장한다.

2)클라이언트 프로그램

- 정보 관리 데이터 베이스에 엑서스하기 위한 로그인 기능을 갖는다.
- 정보 관리 데이터 베이스의 정보를 호출하기 위한 정해진 규약의 sql문을 전송한다.
- 정보 관리 데이터 베이스로부터 최신 정보를 얻기 위해 가장 최근에 데이터 베이스에 엑서스한 시간을 자동으로 기록하고, 이를 기초해서 새로운 자료를 조사한다.
- 사이트 관리 서버 프로그램에 접속하여 새로운 사이트를 등록하는 기능을 갖는다.
- 초기 프로그램 설치시 검색을 시작할 초기 시간 설정 기능을 갖는다.
- 데이터 베이스로부터 가져온 정보를 로컬 디스크에 정리, 저장하고 관리한다.

3)정보 관리 데이터 베이스

- 갱신된 사이트 정보를 정해진 테이블 규격에 맞게 저장한다.
- 각 사용자의 아이디와 비밀번호를 관리한다.

4)임시 데이터 베이스

- 각 사이트의 가장 최근 정보를 임시로 저장한다.
- 저장된 정보는 각 사이트의 현재 정보와 비교해 현재의 정보가 갱신된 정보인지 비교, 분석하는데 사용된다.

5)사이트 등록 데이터 베이스

- 서버 관리 프로그램이 특정 사이트를 매번 조사하도록 사이트 목록을 제공한다.

4.2 클라이언트 소프트웨어

초기 화면으로 나오는 로그인 과정을 성공적으로 마치면 그림 4와 같은 화면이 나타난다. [정보제목]창에 있는 목록을 클릭하면 [상세보기]창에 정보의 자세한 내용이 나타나고 그 상태에서 삭제 버튼을 누르면 선택되어 있는 목록이 삭제된다.

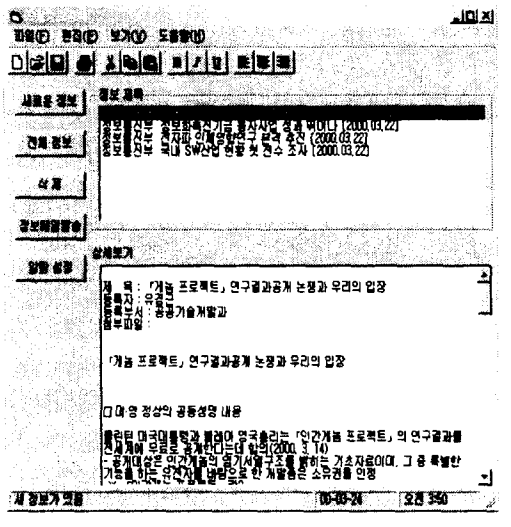

그림 4. 클라이언트 소프트웨어

5. 결론 및 향후 개발 과제

본 논문에서는 특정 웹사이트의 URL을 저장하여 관리하면서 이들을 매일 접속하여 새로운 정보나 변화가 있으면 시스템 서버(system server)의 데이터베이스 (database)에 저장하여 관리하며 저장된 정보를 미리 예약된 사용자에게 전자메일을 통해서 자동 전송하는 시스템을 개발하였다. 향후 사용자의 일정 관리 기능과 대용량 멀티미디어 파일이나, 그래픽 파일 등을 서비스 할 수 있도록 할 예정이다.

참 고 문 헌

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

[2] 박정훈, 조현성, 이강찬, 이규철, "인터네트 정보 발견 시스템의 개발 및 구현", 정보과학회논문

[3] Joon Ho Lee, "Combining Multiple Evidence from Different Properties of Weighting Schemes", Cornell University

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

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

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