2002년도 대한전기학회 하계학술대회 논문집 2002.7.10-12

Web-Picker를 이용한 템포럴 메세징 시스템

이미란°, 조동섭
이화여자대학교 컴퓨터학과

Temporal Messaging System using Web-picker

Mi-Ran Lee°, Dong-Sub Cho
Dept. of Computer Science and Engineering, Ewha Womans University

Abstract - 오늘날 인터넷이 보편화되면서 수많은 정보들이 빠른 시간 내에 갱신되고 있지만, 사용자들은 필요로 하는 정보가 언제 갱신되는지 알지 못해서 동일 웹 페이지를 반복적으로 접근한다. 이렇게 정보의 갱신을 확인하기 위하여 동일 웹사이트를 여러 번 방문하는 일은 너무나 번거롭고 비효율적이다. 본 논문에서는 이러한 문제점을 해결하기 위하여 웹픽커(Web-Picker)를 이용한 템포럴 메세징 시스템을 제안하고자 한다. 사용자가 등록한 웹 페이지들을 웹픽커는 주기적으로 접속하여 검색하고, 웹 페이지에서 사용자가 필요로 하는 정보가 갱신된 경우 그 정보를 사용자에게 e-mail로 자동으로 전송해주는 시스템이다.

1. 서 론

1990년대 중반에 일어나기 시작한 인터넷 열풍은 정보의 자원을 보다 확산시키는 계기가 되었다. 특히, 사람들 사이의 메시지를 주고받을 수 있는 e-mail 시스템은 기존의 업무연락 수단인 우편제도에 많은 영향을 주었다. 인터넷은 웹 서버를 중심으로 발전하면서 웹 기반의 응용 서비스가 계속적으로 개발되었고, 사용자의 다양한 욕구가 그 추진 원동력이 되고 있다. 인터넷의 메시지는 HTTP(Hyper Text Transfer Protocol)를 중심으로 전달되고 있어서 대부분의 정보는 웹 페이지 단위로 저작되고 관리되고 있다. 일부 대형 데이터베이스를 사용하는 서비스는 웹 데이터 베이스를 운영한다. 사용자가 원하는 정보는 최종적으로 웹 페이지의 형식으로 전달되어지므로 클라이언트인 사용자는 특정의 정보를 일정한 형식으로 받아보게 된다.

사용자의 웹 사용의 형태를 분석해보면 단일 정보를 주기적으로 접근하는 경우가 많다. 일부 불특정 정보의 서핑도 가능하지만 시간에 따라 변하는 정보를 주기적으로 확인해야 하는 일이 대부분이다. 정보의 갱신을 확인하기 위해서 주기적으로 지정된 웹사이트에 접속하여 정보를 검색하는 일은 너무 번거롭고 비효율적이다. 웹 정보의 자동 검색과 관리를 위한 새로운 프로그램을 서비스함으로써 사용자의 편의와 정보 욕구를 더욱 증대시킬 필요가 있다.

이러한 문제점을 해결하기 위하여 본 논문에서는 웹픽커(Web-Picker)라는 응용 서비스를 제안하여 이를 실제적으로 응용할 수 있는 개발 프로그램을 만들어 보았다. 웹픽커를 이용한 템포럴 메세징 시스템은 웹사이트에서 검색되어 축적된 정보는 사용자에게 e-mail로 자동으로 전송해주는 시스템이다. 기존의 HTTP 웹 서버를 그대로 사용하며 자체적으로 웹픽커 서버를 운영하도록 하여 이를 실현하였다.

웹픽커 서비스는 시간에 따라 변하는 데이터를 관리하는데 더욱 효과적이다. 웹에서 수집한 정보를 시간 축상에서 재정리하여 사용자가 원하는 형태로 변환하여 가공하여 전송하도록 하였다. 사용자들의 정보 요구는 사용자가 미리 지정해 놓은 정보만을 전송할 수 있으며, 일반 e-mail 방식으로 정보를 전송하는 과정에서 정보의 발신을 자동으로 처리할 수 있다. 사용자도 별도의 정보 관리용 데이터베이스를 사용하지 않고 Outlook Express와 같은 e-mail 클라이언트 프로그램을 사용하여 정보를 효율적으로 관리할 수 있다.

논문의 순서는 다음과 같다. 먼저 2장에서 기존의 연구들을 기술하고, 3장에서는 웹픽커 서비스를 위한 시스템과 구현 결과에 대해 설명한다. 마지막으로 4장에서는 웹픽커 서비스의 향후과제로 본 논문을 맺는다.

2. 웹 문서 수집 및 처리

2.1 웹 문서 수집(Robot Agent)

로봇이 웹 상의 HTML(Hypertext Markup Language) 문서들을 수집한다. 로봇은 HTTP를 통한 웹 서버와 통신을 가지고 있으며 HTML 문서를 처리할 수 있는 능력을 가지고 있다. 예를 들어 URL만을 따로 뽑아 내거나, 문서상의 모든 태그(Tag)를 떼어내는 일도 할 수 있다. URL(Uniform Resource Locator)만을 별도로 뽑아내면 이 URL들을 가지고 로봇은 다음 웹 서버로 항해를 계속 할 수가 있다. 이렇게 로봇이 웹을 돌아다니다 보면 이미 방문했던 곳을 다시 방문하는 일이 발생하게 되는데 특별한 일이 아니면 대부분 다시 방문하는 일은 하지 않는다. 따라서 다시 방문하는 것을 막기 위해 방문한 URL들의 리스트를 별도로 가지고 있어야 한다. 웹사이트에 방문하기 전에 막기 위해 방문한 URL들의 리스트를 별도로 가지고 있어야 한다. 웹사이트에 방문하기 전에 URL 리스트를 통하여 방문했는지를 확인한 다음 방문했던 곳이면 방문하지 않는다[5].

일반적으로 검색엔진에서의 소프트웨어 로봇은 통계적 분석, 유지관리, 미러링, 리소스 탐사 등의 목적으로 이용되고 있다.

탐색 방법과 탐색 주기에 따라 웹 문서 수집기의 성능이 달라질 수 있는데, 일반적으로 주어진 URL 주소 집합을 이용해서 너비 우선 탐색과 같은 탐색 과정을 수행하며, 탐색 주기는 검색 시스템의 도메인에 따라 하루에 한번 또는 2~3일에 한번씩 탐색하도록 한다. URL 주소 관리 시 탐색 여부를 확인하는 필드를 두어 주기적으로 자료를 갱신할 수 있도록 하고 웹 문서 탐색기는 네트워크 부하를 줄이기 위해 사용자가 줄어드는 시간에 작동하도록 한다[4].

2.2 정보 검색 에이전트

2.1과 같이 로봇에서 수집된 문서는 정보검색 에이전트에게 전달된다. 정보 검색 에이전트는 각 문서들에 대해서 분석을 수행하여 색인 단어를 추출한다. 추출된 색인 단어와 문서는 별도로 저장된다. 사용자가 질의한 쿼리는 정보 검색 에이전트에게 전달된다. 정보 검색 에이전트는 이미 수집된 문서들에 대해 질의 내용이 얼마나 적합한지를 평가한다. 평가한 결과 중에서 가장 좋은 문서만을 추려서 그 문서들의 정보를 전송한다. 정보 검색

이 논문은 2002년도 두뇌한국21사업에 의하여 지원되었음.

에이전트는 몇 개의 문서를 골라야 하는지를 알 수 없기 때문에 사용자는 검색하고자 하는 문서의 최대 개수를 지정해 주어야 한다[1].

3. Web-Picker 서비스 시스템

3.1 Web-Picker 서비스의 개념 및 필요성

본 논문에서 제안하는 웹픽커를 사용한 템포럴 메세징 시스템은 사용자가 등록해 놓은 웹 페이지를 주기적으로 접속하여 웹 페이지의 정보가 갱신되었는지 확인하고, 만약 웹 페이지의 정보가 갱신되었을 경우, 웹 페이지에서 검색하여 축적한 정보를 사용자에게 e-mail을 통해 전송해주는 시스템이다.

이 시스템을 사용하면 사용자는 웹 페이지의 URL들과 키워드만을 등록하여 놓으면 된다. 그러면 정보가 갱신되었을 경우 웹픽커 서비스가 e-mail을 통해 그 정보를 알려주어 사용자의 불필요한 시간낭비를 막을 수 있게 한다. 그리고 주기적으로 정보의 변화를 정리해서 문서화함으로써 웹픽커 서비스는 시간에 따라 변하는 정보의 내용을 검색하는데 효과적이다. 또 일반 e-mail 방식으로 정보를 전송하므로, 사용자도 별도의 정보 관리용 데이터베이스를 사용하지 않고 e-mail 클라이언트 프로그램을 사용하여 정보를 효율적으로 관리할 수 있다.

3.2 Web-Picker 서비스 처리 단계

웹픽커 서비스를 위하여 자체적으로 웹픽커 서버를 운영하였고, 웹픽커 서비스 프로그램은 기존의 웹 서버들을 HTTP를 사용하여 접속하였다.

그림 1은 웹픽커 서비스의 전체적인 처리 과정을 보여주고 있다.

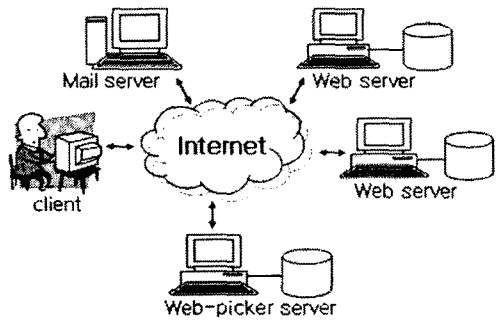

그림 1. Web-Picker 서비스 처리 과정

웹픽커 서비스의 전체적인 처리 과정은 우선, 사용자가 등록해놓은 웹 서버에 접속하여 웹 페이지의 정보를 가져온다. 이 때 가져온 문서들의 정보가 갱신되었을 경우, 그 정보를 웹픽커 서버의 데이터 베이스에 등록시킨다. 등록된 웹 페이지는 웹픽커의 프로세싱 단계에서 사용자가 미리 지정해 놓은 정보만을 추출하는 과정을 통해 문서화된다. 이렇게 문서화된 정보는 사용자의 e-mail로 보내지고, 사용자는 자신의 메일을 확인함으로써 정보가 갱신되었다는 것을 알 수 있게 되고, 또한 갱신된 정보의 내용이 무엇인지도 쉽게 알 수 있다.

위에서 설명한 웹픽커를 사용한 템포럴 메세징 시스템은 크게 웹픽킹(Web Picking) 단계와 웹픽커의 메일 처리 단계로 나눠진다.

3.2.1 Web-Picking 단계

웹픽킹 단계는 웹픽커 서비스 프로그램이 웹서버에 접속하여 웹 페이지의 정보를 데이터베이스에 등록하기까지의 과정을 말한다.

우선 접속해야하는 웹서버의 주소를 알기 위하여 사용자가 자주 정보를 확인하는 웹 페이지의 URL들을 입력받고, 또 웹 페이지의 정보 검색을 위하여 사용자가 필요로 하는 정보의 키워드들을 입력받는다.

사용자가 입력해 놓은 URL의 웹 서버에 HTTP를 사용하여 접속하고, 접속한 웹서버에서 등록되어 있는 웹 페이지들의 정보를 가져온다. 가져온 웹 페이지의 정보와 이전에 등록해 놓은 웹 페이지의 정보를 비교한다. 비교한 결과 웹 페이지의 정보가 갱신된 경우는 그 웹 페이지의 정보를 웹픽커 서버의 데이터베이스에 등록시키고, 웹 페이지의 정보가 갱신되지 않은 경우는 그 웹 페이지의 정보를 버린다.

이때 웹 서버의 과다한 부하를 줄이기 위하여 웹픽커 서비스 프로그램은 웹 페이지를 가져올 때에만 연결을 유지하고 웹 페이지를 가져온 이후에는 접속을 끊는다. 그리고 일정 시간동안 기다렸다가 다시 접속하여 웹 페이지를 가져와서 비교하는 위의 과정을 반복한다.

위에서 설명한 웹픽킹 단계를 그림 2에서 flowchart로 나타내었다.

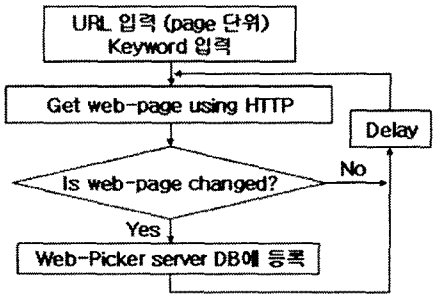

그림 2. Web-Picking 단계 flowchart

3.2.2 Web-Picker Mail 처리 단계

웹픽커 메일 처리 단계는 데이터베이스에 등록된 웹 페이지의 정보를 검색하여 추출한 내용을 사용자에게 메일로 보내기까지의 과정을 말한다.

그림 3에서는 웹픽커 메일 처리 단계를 flowchart로 나타내었다.

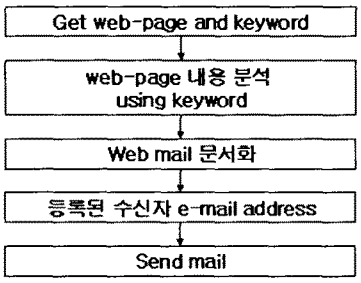

그림 3. Mail 처리 단계 flowchart

먼저 웹픽킹 단계에서 사용자로부터 입력받은 키워드와 데이터베이스에 등록해 놓은 웹 페이지의 정보를 가져온다. 이렇게 가져온 웹 페이지의 정보 내용을 분석하는데, 이때 키워드를 사용해서 내용을 분석한다. 즉 웹 페이지 내용 중에 키워드의 정보가 있는지 검색해서 키워드에 해당하는 정보가 있으면 그 정보의 내용들만 추출한다. 정보의 내용은 e-mail의 본문 내용으로 보내질 것이므로 web mail 문서화 작업을 한다. 이렇게 만들

어진 문서는 사용자가 등록해 놓은 e-mail address로 보내지고, 보내진 e-mail을 사용자가 확인함으로써 키워드에 해당하는 정보가 갱신된 것과 갱신된 정보의 내용이 무엇인지 알 수 있다.

3.3 구현 결과 및 평가

웹 페이지를 가져오기 위하여 웹 서버에 HTTP로 접속하도록 구현하였고, 가져온 웹 페이지는 이전에 가져온 웹 페이지와 비교하여 정보가 갱신된 경우에만 저장하도록 구현하였다. 이렇게 가져온 웹 페이지에서 키워드에 해당하는 내용이 있는지 정보를 검색하는 알고리즘은 그림 4에서 C 언어로 표현하였다.

```
FILE *compare1, *compare2;
compare1 = fopen("compare.txt", "r"): //키워드 파
compare2 = fopen(outf_name, "r"):    //웹 페이지

while(fscanf(compare1, "%s", str1)!=EOF)
{ length1=strlen(str1);
  while (fscanf(compare2, "%s", str2)!=EOF)
  { length2=strlen(str2);
    for (i=0; i<length2; i++)
    { if ((strncmp(&str2[i], str1, length1))=
      //키워드와 웹 페이지 내용이 일치하는지 검색
      { fscanf(compare2, "%s", str2);
        fprintf(out, "%s => %s", str1, str2);
      //일치하는 경우 정보의 내용 출력
      }
    .....
}
```

그림 4. 웹 페이지에서 정보를 검색하는 알고리즘

실제로 구현한 웹픽커 서비스 프로그램을 테스트하기 위한 실험의 예로 주가에 대한 정보 검색을 해보았다.

웹픽커 서비스가 접속할 웹 페이지의 URL로 자주 정보의 내용이 갱신되는 http://www.daum.net/ 과 http:// kr.financc.yahoo.com/ 를 입력하였다. 그리고 정보 검색을 위한 키워드에는 KOSDAQ, 코스닥, Nasdaq, 나스닥을 입력하였다.

그림 5는 위의 정보를 입력한 웹픽커 서비스를 실행하였을 때, 사용자의 e-mail에 보내진 내용을 캡쳐한 화면이다.

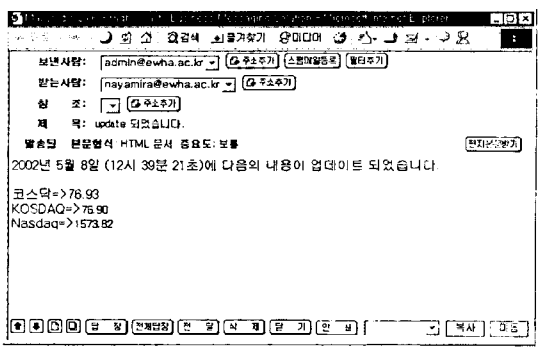

그림 5. Web-Picker 서비스가 전송한 e-mail

그림 5에 보이는 것처럼 웹 페이지가 갱신될 때마다 웹 페이지의 정보 중에서 키워드의 정보들만 추출되어 나타났고, 시간에 따라 정보가 어떻게 변하는지 알 수 있도록 정보가 갱신된 시간이 나타났다.

테스트 결과 웹 페이지를 가져와서 데이터베이스로 등록하고, 또 키워드에 해당하는 정보를 검색하는 과정이 완전하게 동작하고 있고, e-mail로 보내진 정보 검색의 내용 또한 만족할 만한 수준으로 나타나고 있다.

4. 결 론

제안한 웹픽커 서비스는 시간에 따라 변화하는 웹 정보를 효율적으로 관리하는 서비스이다. 사용자의 주기적인 웹 서버 접근의 어려움을 해소해주고, 수집된 정보의 처리 등 효과적인 자료의 표현을 도와줄 수 있다. 실험적으로 구축된 웹픽커 서비스로 알 수 있듯이 웹픽커 서비스는 완전하게 동작할 수 있고, 웹 상에서 관리되는 웹 페이지의 모든 정보는 웹픽커 서비스 시스템으로 접근 가능하다. 웹 데이터베이스를 사용한 대형 웹 서비스도 마찬가지로 본 웹픽커 서비스를 사용하면 사용자의 접속 방식이 결정된 후, 자동적으로 자료를 모으고 처리하여 e-mail로 전송할 수 있다. 다만, 대형의 데이터를 e-mail로는 보낼 수는 없겠지만 일정 수까지는 통계적 자료처리 수준에서 결과를 전송할 수 있을 것이다. 향후 그래픽 전용 처리 모듈과 결합되고, 다수의 사용자를 관리하는 모듈을 추가하면, 다양한 서비스를 개발할 수 있으리라 본다.

[참 고 문 헌]

[1] Gabriel L. Somlo, Adele E. Howe, "Incremental clustering for profile maintenance in information gathering web agents," In Proceedings of the fifth ACM international conference on Autonomous agents, pp.262-269, 2001.

[2] Edmund S. Yu, Ping C. Koo, Elizabeth D. Liddy, "Evolving intelligent text-based agents," In Proceedings of the fourth ACM international conference on Autonomous agents, pp.388-395, 2000.

[3] Koster Martijin, "Robot in the Web: threat or treat," April, 1995.

[4] M.Amin, D.Ballard, "Defining new markets for intelligent agents," Vol. 2 No. 4, pp.29-35, 2000.

[5] 남건우, 이형우, 최창원, 김태윤, "사용자 인터렉션이 가능한 다중 에이전트 기반 전문분야 검색 엔진 설계," 정보과학회 학술발표논문집, Vol. 25 No. 1, pp.255-257, 1998.

[6] 성낙윤, 백철경, 조민규, "소프트웨어 에이전트 모형개발에 관한 연구," 경성대학교 논문집, Vol. 19 No. 2, pp.441-447, 1998.

[7] 성백균, "사례 기반 추론을 이용한 사용자 인터페이스 에이전트에 관한 연구," 충주대학교 논문집, Vol. 34 No. 2, pp.325-340, 1999.