웹 페이지 분석을 위한 Web-Picker 설계 및 구현

이미란° 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과
{nayamira°, dscho}@ewha.ac.kr

Design and Implementation Web-Picker for web page analysis

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

요 약

1990년대 중반에 일어나기 시작한 인터넷 열풍은 웹을 통한 인터넷의 확산으로, 웹 기반 서비스의 발전을 가져왔다. 현재 대부분의 인터넷 서비스는 HTTP를 사용한 일정한 형식의 웹 페이지로, 사용자가 최종적으로 받아보게 되는 형식은 HTML의 태그로써 나타내어진다. 어떠한 태그를 어떻게 사용하였느냐에 따라 그 웹 페이지는 사용자가 사용하기 편리할 수도 있고, 사용하는데 불편할 수도 있다. 따라서 웹사이트의 개발자는 사용자가 편리함과 친숙성을 느낄 수 있도록 웹사이트를 개발해야 한다.

본 논문에서는 이를 해결하기 위해 웹 페이지 분석을 위한 Web-Picker를 제안하고자 한다. Web-Picker를 이용하면 사용자들이 자주 방문하는 웹 페이지의 태그를 분석할 수 있고, 이렇게 분석한 정보를 통해 새로운 웹사이트를 개발하는 개발자들은 사용자가 친숙성과 편리함을 느낄 수 있도록 웹사이트를 개발 할 수 있다.

1. 서 론

웹을 중심으로 인터넷이 발전하면서 웹 기반의 응용 서비스가 계속적으로 개발되고 있고, 사용자의 다양한 욕구가 추진 원동력이 되고 있다. 인터넷 메시지는 HTTP(Hyper Text Transfer Protocol)를 중심으로 전달되고 있고, 대부분의 정보는 웹 페이지 단위로 저작되고 관리되고 있다. 사용자가 원하는 정보는 최종적으로 웹 페이지의 형식으로 전달되어지므로 클라이언트인 사용자는 대게 정보를 일정한 형식으로 받아보게 된다.

이러한 일정한 형식은 HTML(Hyper Text Markup Language)의 태그(Tag)로써 나타내어진다. 어떤 태그를 어떻게 사용하였느냐에 따라 사용자가 웹사이트를 이용하는데 편리함을 느낄 수도 있고, 불편함을 느낄 수도 있다. 따라서 웹사이트의 개발자는 사용자가 이용하는데 편리함을 느끼도록 웹 페이지를 개발할 필요가 있다.

이를 해결하기 위해 본 논문에서는 웹 페이지 분석을 위한 Web-Picker를 제안하고자 한다. Web-Picker는 여러 웹 페이지들의 주소를 입력받아 해당 서버에 접속하여 웹 페이지를 가져온다. 가져온 웹 페이지를 Web-Picker는 읽어들여, 웹 페이지에서 사용된 각각의 태그에 대한 사용 빈도 수를 카운트한다. 이렇게 분석된 웹 페이지와 각각의 태그에 대한 사용 빈도 수는 데이터베이스에 저장된다. 저장된 정보들을 이용하여 사용자들이 자주 방문하는 웹 페이지의 태그 특징들을 알아 낼 수 있다. 또 개발자가 새로 개발하는 웹 페이지의 태그 특징들을, 사용자들이 자주 찾는 웹 페이지의 태그 특징들과 비교·분석하여, 새로 개발한 웹 페이지를 사용하는데 편리함과 친숙성을 느끼게 할 수도 있다.

이 논문은 2003년도 두뇌한국21사업에 의하여 지원되었음.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 연구들을 기술하고, 3장에서는 웹 페이지 분석을 위한 Web-Picker의 설계 방법과 구현 결과에 대해 설명한다. 마지막으로 4장에서는 Web-Picker 서비스의 향후과제로 본 논문을 맺는다.

2. 웹 문서 수집 및 정보 검색

2.1 웹 문서 수집(Robot Agent)

로봇이 웹 상의 HTML 문서들을 수집한다. 로봇은 HTTP를 통한 웹 서버와 통신을 가지고 있으며 HTML 문서를 처리할 수 있는 능력을 가지고 있다. 예를 들어 URL(Uniform Resource Locator)만을 따로 뽑아 내거나, 문서상의 모든 태그를 떼어내는 일도 할 수 있다. URL만을 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 URL 리스트를 통하여 방문했는지를 확인한 다음 방문했던 곳이면 방문하지 않는다.

때때로 로봇이 동작할 곳의 지역을 한정시킬 필요가 있다. 예를 들어 한국에 있는 웹사이트 중에서만 검색을 해보고 싶다면 로봇을 한국 이외의 웹 서버에는 방문하지 못하도록 하면 된다. 또한 로봇이 방문할 수 있는 영역을 설정함으로써 불필요한 검색을 막을 수도 있다.

웹 상에는 수많은 문서들이 있기 때문에 로봇에게 수집할 문서의 최대 개수를 제한하지 않는다면 로봇은 무한정 웹을 돌아다니게 될 것이다. 따라서 로봇이 수집할 문서의 개수를 한정시킬 필요가 있다[4].

2.2 정보 검색 에이전트

2.1과 같이 로봇에서 수집된 문서는 정보검색 에이전트에게 전달된다. 정보검색 에이전트는 각 문서들에 대해서 형태로 분석을 수행하여 색인 단어를 추출한다. 추출된 색인 단어와 문서는 별도로 저장된다.

사용자가 질의한 쿼리는 정보검색 에이전트에게 전달된다. 정보검색 에이전트는 이미 수집된 문서들에 대해 질의 내용에 얼마나 적합한지를 평가한다. 평가한 결과 중에서 가장 좋은 문서만을 추려서 그 문서들의 정보를 전송한다. 정보검색 에이전트는 몇 개의 문서를 골라야 하는지를 알 수 없기 때문에 사용자는 검색하고자 하는 문서의 최대 개수를 지정해주어야 한다[4].

3. 웹 페이지 분석을 위한 Web-Picker

3.1 Web-Picker의 개념 및 필요성

본 논문에서 제안하는 웹 페이지 분석을 위한 Web-Picker는 미리 입력해놓은 여러 웹 페이지의 주소를 가지고 해당 서버에 접속하여 HTTP를 사용하여 웹 페이지의 정보를 가져온다. 여러 웹 서버에서 각기 다른 웹 페이지의 내용을 한번에 가져올 수도 있다. 가져온 페이지의 정보를 분석하여 웹 페이지에서 사용된 태그에 대한 각각의 사용 빈도 수를 카운트한다. 이렇게 분석된 웹 페이지와 태그에 대한 사용 빈도 수를 데이터베이스에 저장하고, 저장된 데이터들을 이용하여 사용자가 자주 찾는 웹 페이지들의 태그 정보들을 알아낼 수도 있고, 또 새로운 웹 페이지를 개발할 때에 사용자에게 익숙한 태그들을 사용하여 웹 페이지를 개발할 수도 있다.

3.2 Web-Picker 처리 단계

Web-Picker의 전체적인 처리 과정은 그림 1과 같다.

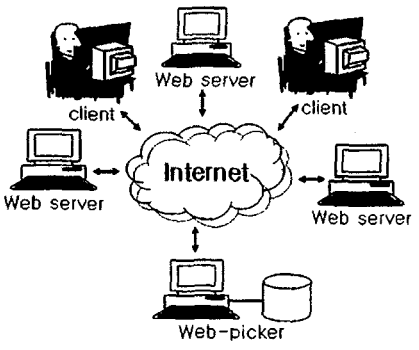

그림 1 Web-Picker 시스템 구성도

전체적인 처리과정으로 우선 Web-Picker는 등록해놓은 웹 페이지의 주소에 따라 웹 서버에 접속하여 웹 페이지의 정보를 가져온다. 이 때 가져온 웹 페이지들의 태그를 분석하여, 각각의 태그가 사용된 사용 빈도 수를 카운트한다. 태그 분석이 끝난 웹 페이지와 태그 각각의 사용 빈도 수는 데이터베이스에 저장된다.

앞에서 설명한 내용은 크게 Web Picking 단계와 Tag Analysis 단계로 나누어진다.

3.2.1 Web Picking 단계

Web Picking 단계에서는 웹 서버에 접속하여 웹 페이지의 정보를 가져오기까지의 과정을 말한다. Web Picking 처리 과정은 그림 2와 같다.

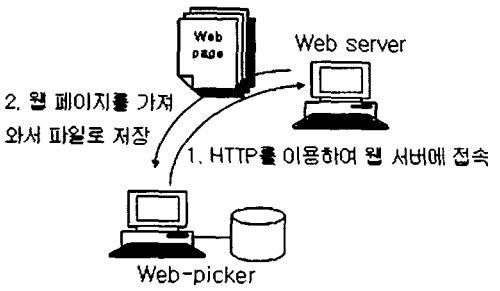

그림 2 Web Picking 처리 과정

우선 접속해야하는 웹 서버의 주소를 알기 위하여 웹 페이지의 URL들을 입력받는다. 입력해 놓은 URL의 웹 서버에 HTTP를 사용하여 접속하고, 접속한 웹 서버에서 등록되어 있는 웹 페이지의 정보를 가져온다. 가져온 웹 페이지의 정보는 임시적으로 파일로 저장되고, 더 이상 입력해 놓은 URL이 없을 때까지 계속해서 웹 서버에 접속하여 웹 페이지의 정보를 가져온다.

3.2.2 Tag Analysis 단계

Tag Analysis 단계에서는 Web Picking 단계에서 가져온 웹 페이지를 읽어들여, 웹 페이지에서 사용된 태그를 분석한다. 이때 HTML에서 사용되는 모든 태그들은 미리 입력되어져 있고, Web-Picker는 읽어들인 웹 페이지에서 미리 입력된 각각의 태그들이 사용될 때마다 카운트를 하나씩 증가시킨다. 이렇게 분석된 웹 페이지를 데이터베이스에 저장하고, Web Picking 단계에서 임시적으로 저장해두었던 웹 페이지 파일은 삭제한다. 웹 페이지를 데이터베이스에 저장할 때, 각각의 태그에 대한 사용 빈도 수도 함께 데이터베이스에 저장한다.

Tag Analysis 처리 과정은 그림 3과 같다.

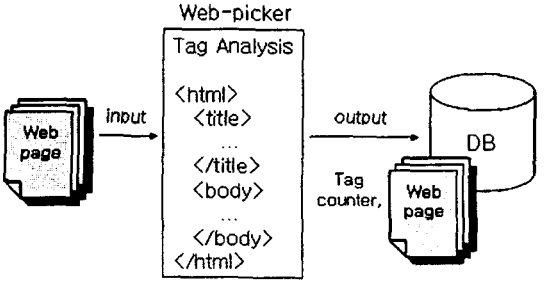

그림 3 Tag Analysis 처리 과정

3.3 구현 결과 및 평가

웹 페이지들의 URL을 미리 입력받도록 구현하였고, 이렇게 입력받은 웹 페이지들을 가져오기 위하여 각각의 웹 서버에 HTTP로 접속하도록 하였다. 가져온 웹 페이지는 임시적으로 파일로 저장하고, 태그 분석이 끝난 후 태그 사용 빈도 수와 함께 MS-SQL 서버에 저장되도록 하였다.

실제로 구현한 Web-Picker를 테스트하기 위하여 웹 페이지의 URL로 http://www.ewha.ac.kr을 입력하고 실행하였다. 아래에 있는 그림 4는 입력시킨 URL인 이화여자대학교의 홈페이지 화면이고, 그림 5는 해당 웹 페이지의 태그가 분석되어 데이터베이스에 입력된 결과를 나타낸 화면이다. 아래에 있는 태그뿐만 아니라 여러 태그들이 있지만, 화면 관계상 몇 가지 태그들만 보이도록 하였다.

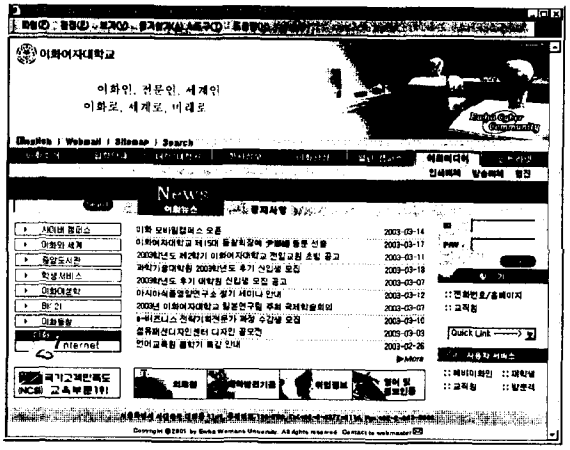

그림 4 이화여자대학교의 홈페이지

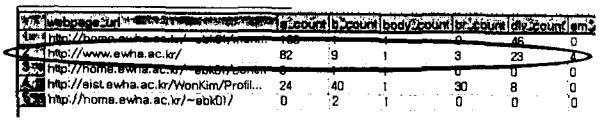

그림 5 이화여자대학교 홈페이지를 분석한 결과화면

테스트 결과 실제로 여러 개의 HTML 문서를 가져올 수 있었고, 가져온 웹 페이지들이 각각의 태그 사용 빈도 수에 따라 분석되는 것을 볼 수 있었다. 또 이렇게 분석된 내용을 데이터베이스로 저장하는 과정까지 모두 만족할 만한 수준으로 나타났다.

4. 결론

제안한 웹 페이지 분석을 위한 Web-Picker는 웹 페이지의 태그를 분석하는 프로그램이다. 인기 있는 페이지의 태그 사용 현황을 손쉽게 알 수 있고, 또 여러 사이트에서 어떤 태그가 주로 이용되고 있는지를 분석해 볼 수도 있다.

웹 페이지에서 얻어낼 수 있는 정보는 태그 정보뿐만 아니라, 웹 페이지에서 주로 사용하는 색깔이나 자주 이용되는 프레임 형식 및 비율 등도 알아낼 수 있다. 이러한 정보들을 효율적으로 이용한다면, 웹사이트를 개발하는 개발자들은 사용자들이 사용하기에 편리하고 친숙한 웹 페이지를 개발 할 수 있을 것이다.

향후 웹 페이지 분석 방법이 태그의 사용 빈도 수에서 그치지 않고, 데이터 마이닝을 이용하여 웹 페이지의 패턴을 분석하는 것 등의 좀 더 다양한 접근을 해 볼 생각이다.

참고문헌

[1] 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.

[2] 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.

[3] Greg, R., "Searching the Hidden Internet", Database, Vol. 20, No. 2, 1997.

[4] 성낙운, 백철경, 조민규, "소프트웨어 에이전트 모형 개발에 관한 연구," 경성대학교 논문집, Vol.19, No.2, pp.441-447, 1998.

[5] 윤효근, 이상용, "바이오 인포메틱스를 이용한 웹 페이지 분석 기법에 관한 연구," 한국정보과학회 2001가을 학술발표논문집, VOL.28, NO.2, pp.97-99, 2001