2003년도 대한전기학회 하계학술대회 논문집 2003.7.21-23

웹 환경에서의 이미지 수집을 위한 Web Image Picker 설계 및 구현

이미란, 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과

Design and Implementation of Web Image Picker for image collection

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

Abstract - 웹을 중심으로 인터넷이 발전하면서 웹 기반의 응용 서비스가 계속적으로 개발되고 있고, 대부분의 인터넷 정보들은 웹 페이지 단위로 저작되고 관리되고 있다. 웹 페이지는 일정한 형태인 HTML의 태그로써 나타내어지고, 텍스트뿐만 아니라 그림, 소리, 동영상 등의 여러 가지 미디어를 사용하여 사용자에게 다양한 정보를 제공하고 있다. 본 논문에서는 웹 페이지에서 사용되고 있는 다양한 미디어 중에서도 특히 이미지를 수집하기 위한 Web Image Picker를 제안하고자 한다. 이는 여러 웹 페이지들의 주소를 입력받아 해당 웹 서버에 접속하여 웹 페이지를 가져오고, 가져온 웹 페이지에서 사용된 이미지 태그를 분석한다. 분석된 이미지 정보를 통해 이미지의 사용 횟수를 알 수 있고, 또한 이미지 파일들을 자동으로 수집할 수 있다.

1. 서 론

오늘날 인터넷 열풍으로 인하여 정보의 자원이 보다 확산되고 있다. 특히 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)을 중심으로 전달되고 있고, 대부분의 정보는 웹 페이지 단위로 저작되고 관리되고 있다. 사용자가 원하는 정보는 최종적으로 웹 페이지의 형식으로 전달되어지므로 클라이언트인 사용자는 정보를 일정한 형식으로 받아보게 된다. 웹 페이지에서 대개 사용하는 형식은 HTML(Hyper Text Markup Language)의 태그(Tag)로써 나타내어진다. 이러한 일정한 형식을 가지는 웹 페이지는 텍스트뿐만 아니라 그림, 소리, 동영상 등의 여러 가지 미디어를 사용하여 사용자에게 다양한 정보를 제공하고 있다.

본 논문에서는 웹 페이지에서 사용되고 있는 다양한 미디어 중에서도 특히 이미지를 수집하기 위한 Web Image Picker를 제안하고, 이를 실제적으로 응용할 수 있는 개발 프로그램을 만들어 보았다. Web Image Picker는 여러 웹 페이지들의 주소를 입력받아 해당 서버에 접속하여 웹 페이지를 가져온다. 가져온 웹 페이지를 읽어들여 웹 페이지에서 사용된 이미지의 태그를 분석하고, 이렇게 분석한 내용을 통하여 이미지의 사용 횟수와 웹 페이지 상에서 사용된 이미지를 자동으로 저장할 수 있다. Web Image Picker를 통해 웹 페이지에서 사용되고 있는 다양한 이미지를 수집할 수 있고, 또한 웹 페이지에서 이미지가 사용되는 빈도수도 알 수 있다.

본 논문의 순서는 다음과 같다. 먼저 2장은 기존연구로써 웹 문서 수집과 이미지 파일에 대해 기술하고, 3장에서는 웹 환경에서의 이미지 수집을 위한 Web Image Picker의 설계 방법과 구현 결과에 대해 설명한다. 마지막으로 4장에서는 Web Image Picker의 향후과제로 본 논문을 맺는다.

이 논문은 2003년도 두뇌한국21사업에 의하여 지원되었음.

2. 웹 문서 수집과 이미지 파일

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 것을 막기 위해 방문한 URL들의 리스트를 별도로 가지고 있어야 한다. 웹 사이트에 방문하기 전에 URL 리스트를 통하여 방문했는지를 확인한 다음 방문했던 곳이면 다시 방문하지 않는다.

때때로 로봇이 동작할 곳의 지역을 한정시킬 필요가 있다. 예를 들어 한국에 있는 웹 사이트 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].

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[5].

2.2 이미지(Image)

이미지 파일을 크게 분류하는 방식은 그림을 표현하는 방식에 따라 두 가지로 나누어진다. 첫 번째로 비트맵(Bitmap) 방식이 있다. 비트맵 방식은 픽셀이라는 다수의 사각형 입자로 이미지를 표현하는 방식으로 각각의 픽셀이 이미지를 나타내기 위한 고유의 색상값과 좌표를 가지고 이 픽셀들이 하나의 이미지를 구성하게 된다. 따라서 이런 비트맵 방식은 이미지를 계속 확대하면 모자이크 식으로 그림이 나타나는 것을 볼 수 있다. 이러한 픽셀들이 일정한 단위 넓이에 얼마나 존재하느냐에 따라 이미지의 해상도(Resolution)가 결정된다. 픽셀의 수가 많을수록 해상도가 높아지고 이미지가 선명해진다.

두 번째는 벡터(Vector) 방식 이미지가 있다. 벡터 방식은 수학적인 연산을 따르는 직선, 곡선, 다각형, 타원 등을 이용해 이미지를 표현하는 것으로 각각의 직선과 곡선은 이미지를 나타내기 위한 수학적인 좌표와 각도로 연결되며 그 내부를 고유의 색상으로 채우게 된다.

따라서 벡터 이미지는 비트맵 이미지와 달리 확대하거나 축소하는 경우에도 형태나 색상에 아무런 변화가 일어나지 않으며, 출력되는 모니터나 프린터의 해상도에만 영향을 받는다[6].

이러한 그림 파일의 포맷은 40여 가지가 넘을 정도로 다양하다. 이런 포맷들은 각각의 특징에 따라 작업의 성격에 맞는 형식을 사용하게 된다. 주로 사용되는 포맷으로는 jpg, jpge, gif, png, bmp 등이 있다.

3. Web Image Picker

3.1 Web Image Picker의 개념 및 시스템 구성

본 논문에서 제안하는 웹 환경에서의 이미지 수집을 위한 Web Image Picker는 웹 페이지에서 사용되는 다양한 이미지를 수집한다. 우선 미리 입력해 놓은 여러 웹 페이지의 주소를 가지고 해당 서버에 접속하여 HTTP를 사용하여 웹 페이지의 정보를 가져온다. 여러 웹 서버에서 각기 다른 웹 페이지의 내용을 한번에 가져올 수 있다. 가져온 페이지의 정보를 분석하여 웹 페이지에서 사용된 이미지에 대한 태그를 알아내고, 이미지의 사용 횟수와 웹 페이지 상에서 이미지 태그를 분석함으로써 이미지가 저장된 웹 서버와 그 밖의 이미지에 대한 정보들을 분석한다. 이렇게 분석된 정보를 가지고, 웹 페이지에서 사용된 이미지에 해당하는 웹 서버에 접속하여 이미지 파일을 가져와서 저장한다. 이 과정을 각각의 이미지마다 반복함으로써 웹 페이지에서 사용된 모든 이미지를 자동으로 수집할 수 있다.

Web Image Picker를 통하여 각각의 웹 페이지에서 사용된 이미지의 사용 빈도 수를 알 수 있고, 또한 웹 페이지에서 사용되고 있는 많은 이미지를 손쉽게 수집할 수 있다. Web Image Picker의 전체적인 시스템 구성은 그림 1과 같다.

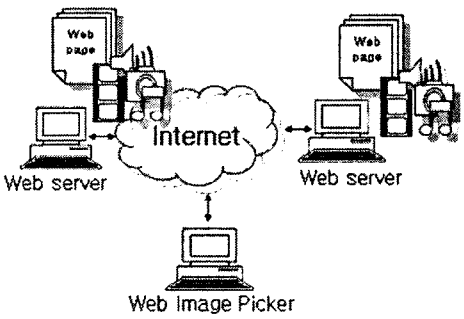

그림 1. Web Image Picker 시스템 구성

3.2 Web Image Picker 처리 단계

Web Image Picker는 크게 Web Page Picking 단계, Tag Analysis 단계, Web Image Picking 단계로 나누어진다.

3.2.1 Web Page Picking 단계

Web Page Picking 단계에서는 웹 서버에 접속하여 웹 페이지의 정보를 가져오기까지의 과정을 말한다. 우선 접속해야 하는 웹 서버의 주소를 알기 위하여 웹 페이지의 URL을 입력받는다. 입력받은 URL에 해당하는 웹 서버에 HTTP를 사용하여 접속하고, 접속한 웹 서버에서 등록되어 있는 웹 페이지의 정보를 가져온다. 가져온 웹 페이지의 정보는 파일로 저장되고, 더 이상 입력해 놓은 URL이 없을 때까지 계속해서 웹 서버에 접속하여 웹 페이지의 정보를 가져온다.

Web Page Picking 의 처리 과정은 그림 2와 같다.

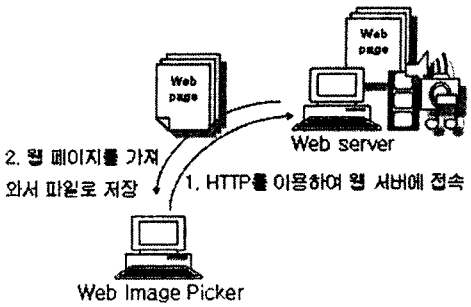

그림 2. Web Page Picking 처리 과정

3.2.2 Tag Analysis 단계

Tag Analysis 단계에서는 Web Page Picking 단계에서 가져온 웹 페이지를 읽어들여, 웹 페이지에서 사용된 태그를 분석한다. HTML에서 사용하는 이미지에 대한 태그를 분석하여, 이미지가 저장되어 있는 웹 서버와 해당 이미지 경로, 이미지 이름 등의 정보를 알아낼 수 있다. 이때 이미지에 대한 태그 카운터를 두어 웹 페이지에서 이미지 태그가 사용될 때마다 카운터를 하나씩 증가시킨다. 이미지 카운터를 통해 분석하고 있는 웹 페이지에서 이미지가 몇 번 사용되었는지 알 수 있다.

이렇게 웹 페이지에서 사용된 이미지 태그에 대한 분석 정보를 이용하여 Web Image Picking 단계에서 웹 서버로부터 이미지 정보를 가져올 수 있다.

3.2.3 Web Image Picking 단계

Web Image Picking 단계에서는 Tag Analysis 단계에서 분석한 이미지에 대한 정보를 가지고, 웹 서버에 접속하여 해당 이미지를 가져온다. 이미지가 저장되어 있는 웹 서버에 접속하고, Tag Analysis 단계에서 분석한 정보인 이미지 경로와 이미지 이름을 통하여 해당 웹 서버로부터 이미지를 가져올 수 있다. 이렇게 가져온 이미지는 Web Image Picker가 실행되는 위치에 저장된다. 이같은 과정을 각각의 이미지마다 반복함으로써 웹 페이지에서 사용된 모든 이미지를 수집할 수 있다.

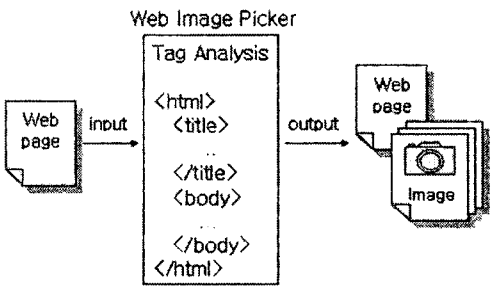

그림 3. Tag Analysis, Image Picking 과정

Tag Analysis 단계와 Web Image Picking 단계의 처리 과정은 그림 3과 같다.

3.3 구현 결과 및 평가

웹 페이지를 가져오기 위하여 웹 페이지들의 URL을 미리 입력받도록 구현하였고, 이렇게 입력받은 웹 페이지들을 가져오기 위하여 웹 서버에 HTTP로 접속하도록 하였다. 가져온 웹 페이지에서 사용된 이미지 태그에 대한 분석을 하고, 이미지가 저장된 웹 서버에 접속해서 각각의 이미지를 가져와서 저장하도록 구현하였다. 이렇게 웹 페이지에서 이미지 태그를 분석하여 해당 이미지

를 파일로 저장하는 알고리즘은 C언어로 구현하였다.

실제로 구현한 Web Image Picker를 테스트하기 위하여 웹 페이지의 URL로 http://www.naver.com을 입력하고 실행하였다. 그림 4은 입력시킨 URL인 네이버의 홈페이지 화면이고, 그림 5은 네이버 홈페이지의 이미지 파일들이 저장된 결과를 보여준다.

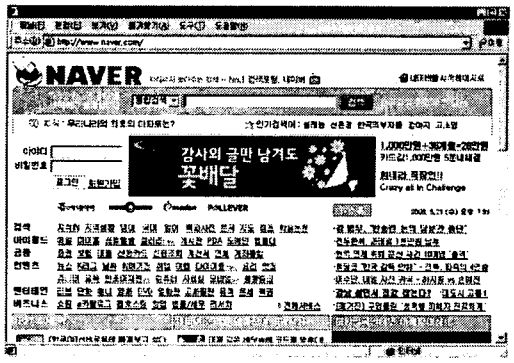

그림 4. 네이버 홈페이지

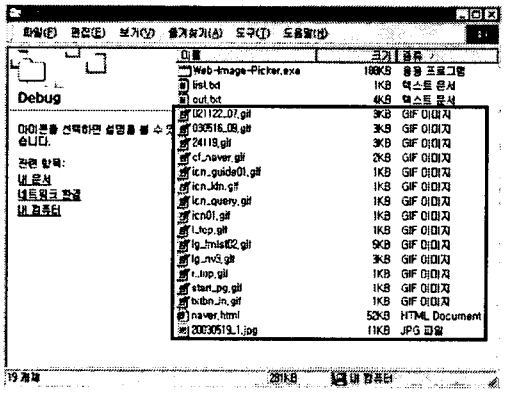

그림 5. 이미지 파일들이 저장된 결과화면

위에서 사각형으로 묶어놓은 파일들은 네이버 홈페이지에서 사용된 이미지들과 웹 페이지를 저장한 것이다. 이때 가져온 이미지는 원래 이미지 파일명 그대로 사용하였다. 그 외의 파일들 중 Web- Image-Picker.exe는 실제로 구현한 프로그램의 실행 파일이고, list.txt는 미리 URL을 입력시켜 놓은 파일이다. 그리고 out.txt는 해당 웹 페이지에서 이미지가 몇 번 사용되었는지 카운트한 결과를 출력해놓은 파일이다.

4. 결 론

제안한 웹 환경에서의 이미지 수집을 위한 Web Image Picker는 웹 페이지에서 사용된 다양한 이미지를 수집하는 프로그램이다. 웹 페이지에서 사용되고 있는 다양한 이미지를 손쉽게 수집할 수 있고, 또 웹 페이지에서 이미지가 사용되는 빈도 수도 알 수 있다. 실험적으로 구축된 Web Image Picker에서 알 수 있듯이 Web Image Picker는 완전하게 동작할 수 있고, 웹 상에서 관리되는 웹 페이지의 모든 이미지 정보를 수집할 수 있다. 향후 Web Image Picker가 웹 페이지에서의 이미지의 수집에서 그치지 않고, 좀 더 다양한 미디어 파일들을 수집할 수 있도록 프로그램을 확장시킬 것이고, 또한 웹 페이지의 태그를 분석함으로써 웹 페이지의 구조적인 정보, 다른 페이지로의 이동 경로 등도 알아낼 것이다.

[참 고 문 헌]

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

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

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

[6] http://compedu.inue.ac.kr/~chlee56/

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