웹 환경에서의 분산 미디어 분석을 위한 Web Media Picker 설계 및 구현

이미란, 이민수, 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과

Design and Implementation of Web Media Picker for Distributed Media Analysis

Mi-Ran Lee, Min-Soo Lee, Dong-Sub Cho
Dept. of Computer Science and Engineering, EIST, Ewha Womans University
{nayamira°, mlee, dscho}@ewha.ac.kr

요 약

웹을 중심으로 인터넷이 발전하면서 웹 기반의 응용 서비스가 계속적으로 개발되고 있고, 사용자의 다양한 욕구가 추진 원동력이 되고 있다. 현재 대부분의 인터넷 서비스는 텍스트뿐만 아니라 그림, 소리, 동영상 등 여러 가지 다양한 미디어를 사용하고 있다. 본 논문에서는 이렇게 다양한 미디어가 어디서, 어떻게 사용되었는지를 분석하기 위하여 웹 환경에서의 분산 미디어 분석을 위한 Web Media Picker를 제안하고자 한다. Web Media Picker를 이용하면 웹 페이지에서 사용된 각각의 태그를 분석할 수 있고, 이렇게 분석한 정보를 통하여 각각의 미디어의 사용 횟수와 미디어가 웹 페이지 상에서 사용된 방법에 대하여 알 수 있다.

1. 서론

1990년대 중반에 일어나기 시작한 인터넷 열풍은 웹을 통한 인터넷의 확산으로, 웹 기반 서비스의 발전을 가져왔다. 인터넷 메시지는 HTTP(Hyper Text Transfer Protocol)을 중심으로 전달되고 있고, 대부분의 정보는 웹 페이지 단위로 저작되고 관리되고 있다. 이러한 웹 페이지는 텍스트뿐만 아니라 그림, 소리, 동영상 등 여러 가지 미디어를 사용하여 사용자에게 다양한 정보를 제공하고 있다.

본 논문에서는 이렇게 다양한 미디어가 어디서 어떻게 사용되었는지를 분석하기 위하여 웹 환경에서의 분산 미디어 분석을 위한 Web Media Picker를 제안하고자 한다. 위에서 언급한 것처럼, 사용자가 원하는 정보는 최종적으로 웹 페이지의 형식으로 전달되어지므로 클라이언트인 사용자는 대개 정보를 일정한 형식으로 받아보게 된다. 이러한 일정한 형식은 HTML(Hyper Text Markup Language)의 태그(Tag)로써 나타내어진다. Web Media Picker는 여러 웹 페이지들의 주소를 입력받아 해당 서버에 접속하여 웹 페이지를 가져온다. 가져온 웹 페이지를 읽어들여 웹 페이지에서 사용된 각각의 태그를 분석하고, 이렇게 분석한 내용을 통하여 각각의 미디어의 사용 횟수와 미디어가 웹 페이지 상에서 사용된 방법에 대하여 알 수 있다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 멀티미디어의 종류들을 기술하고, 3장에서는 웹 환경에서의 분산 미디어 분석을 위한 Web Media Picker의 설계 방법과 구현 결과에 대해 설명한다. 마지막으로 4장에서는 Web Media Picker 서비스의 향후 과제로 본 논문을 맺는다.

이 논문은 2003년도 두뇌한국21사업에 의하여 지원되었음.

2. Web Media Picker를 위한 멀티미디어 파일의 분석

멀티미디어에는 여러 가지 미디어가 사용되며, 여기에서는 Web Media Picker를 위한 미디어 각각의 특성을 간략히 소개한다. [4]

2.1 텍스트

웹 페이지에 등장하는 여러 가지 미디어 중에서도 가장 기본적인 미디어는 글자이다. 그러나 단순한 의미의 전달만이 아니라 글자의 모양이나 크기, 글자로

이루어진 문장과 같은 종합적인 의미로 텍스트라는 단어를 사용한다. 텍스트를 사용하지 않고 웹 페이지를 제작할 수는 있으나 중요한 의미를 텍스트를 사용하지 않고 전달하는 것은 쉬운 일이 아니다.

텍스트는 다른 미디어에 비해 상대적으로 적은 정보량으로 많은 내용을 보여 줄 수 있는 효율적인 미디어다. 주로 사용되는 포맷으로는 txt, doc, html, hwp, bak 등이 있다.

2.2 그림

그림 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 색상 값과 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 벡터방식(Vector) 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.

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.

2.3 소리

텍스트나 그림 정보는 모니터 화면으로 전달되는데 비해 소리는 스피커를 통해 음파 형태로 우리의 귀를 통해 듣게 된다. 따라서 소리를 사용하기 위해서는 소리를 발생하는 전용의 하드웨어를 필요로 한다. 우리가 단순히 소리라고 하는 것도 여러 가지 다른 형태가 있다. 글자로 표현할 수 있는 사람의 음성과 글자로는 정확히 표현할 수 없는 천둥 치는 소리와 같은 자연의 음, 또는 음악 분야에서 악보로 표시되는 음표 정보는 우리 귀에는 모두 음파로 전달되나 표현하는 방법이나 의미의 전달이 서로 다르다. 이러한 모든 개념을 포괄하여 사운드라 한다. 사운드 중에서 음표가 아닌 음파로 표시된 정보는 오디오라는 단어로 설명하기도 한다. 주로 사용되는 포맷으로는 mid, wav, aif, au, mp3, ra 등이 있다.

2.4 동영상(비디오와 애니메이션)

애니메이션은 한 장 한 장을 그려서 만든 연속된 그림이며 만화영화를 생각하면 이해가 쉽다. 이에 비해 비디오는 그려서 만드는 것이 아니라 비디오 카메라를 사용하여 실제 상황을 촬영한 것을 말한다. 애니메이션이나 비디오는 연속된 동작이 자연스럽게 보이도록 하기 위해서는 최소한 초당 15 프레임 이상을 보여주어야 한다. 일반적으로 영화의 경우에는 초당 24 프레임, 텔레비전의 경우에는 전송방식에 따라 차이가 있어 우리 나라의 경우 초당 30 프레임, 유럽의 일부 국가에서는 25 프레임을 사용한다.

애니메이션이나 비디오 모두 최종 결과물에는 사운드를 포함하게 된다. 따라서 사운드 처리에 발생하는 문제는 비디오의 경우에 똑같이 발생한다. 이 때 비디오와 사운드의 특성이 다르기 때문에 하나의 파일 내부에도 따로 기록된다. 그러나 결과를 화면과 스피커를 통해 사용자에게 전달될 경우 말하는 사람의 입모양과 소리가 정확히 맞아야 하며 이를 동기화 문제라고 한다.

여러 가지 형태의 동영상 파일 포맷 등이 있지만, 지금은 몇 가지 정도로 정리되어 가고 있다. 주로 사용되는 포맷으로는 avi, mov, mpg, mpeg, rm, ram, dat 등이 있다.

3. 미디어 분석을 위한 Web Media Picker

3.1 Web Media Picker의 개념 및 처리 단계

본 논문에서 제안하는 웹 환경에서의 분산 미디어 분석을 위한 Web Media 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 한 번에 가져올 수 있다. 가져온 페이지의 정보를 분석하여 웹 페이지에서 사용된 미디어에 대한 태그를 알아내고, 각각 미디어의 사용 횟수와 미디어가 웹 페이지 상에서 어떻게 사용되었는지 사용 방법에 대하여 분석한다. 이렇게 분석된 웹 페이지와 미디어에 대한정

보를 데이터베이스에 저장하여, 여러 가지 웹 페이지에서 다양한 미디어가 어떻게 사용되었는지 분석 할 수 있다. Web Media Picker의 전체적인 시스템 구성은 그림 1과 같다.

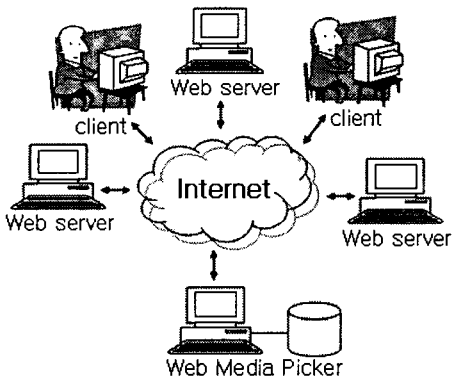


그림 1. Web Media Picker 시스템 구성도

위에서 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.

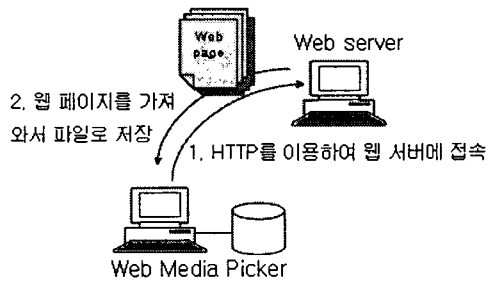


그림 2. Web Picking 처리 과정

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 사용하는 미디어에 대한 태그 카운터를 두어 웹 페이지에서 해당 태그가 사용될 때마다 카운터를 하나씩 증가시키고, 또한 미디어의 사용 방법을 분석한다. 이렇게 분석된 웹 페이지는 데이터베이스에 저장하고, Web Picking 단계에서 웹 페이지를 임시적으로 저장해두었던 파일은 삭제한다. 웹 페이지를 데이터베이스에 저장할 때, 각각 미디어의 사용 횟수와 미디어가 웹 페이지 상에서 사용된 사용 방법에 대한 정보도 함께 저장한다.

Tag Analysis 처리 과정은 그림 3과 같다.

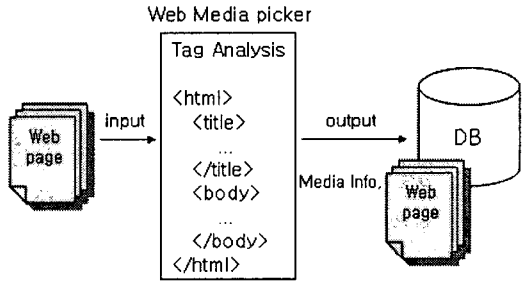


그림 3. Tag Analysis 처리 과정

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 페이지는 임시적으로 파일로 저장하고, 분석이 끝난 후 미디어 사용에 관련된 태그 정보와 함께 MS-SQL 서버에 저장되도록 구현하였다.

실제로 구현한 Web Media Picker를 테스트하기 위하여 웹 페이지의 URL로 http://www.ewha.ac.kr을 입력하고 실행하였다. 그림 4는 입력시킨 URL인 이화여자대학교의 홈페이지 화면이고, 그림 5는 해당 웹 페이지의 미디어 관련 태그 정보가 분석되어 데이터베이스에 입력된 결과를 보여준 화면이다.

테스트 결과 실제로 여러 개의 HTML 문서를 가져올 수 있었고, 가져온 웹 페이지들이 사용된 미디어 관련 태그에 따라 분석되는 것을 볼 수 있었다. 또한 미디어 사용에 대해 분석된 태그 정보가 정리되어 데이터베이스에 저장되는 과정까지 모두 만족할 만한 수준으로 나타났다.

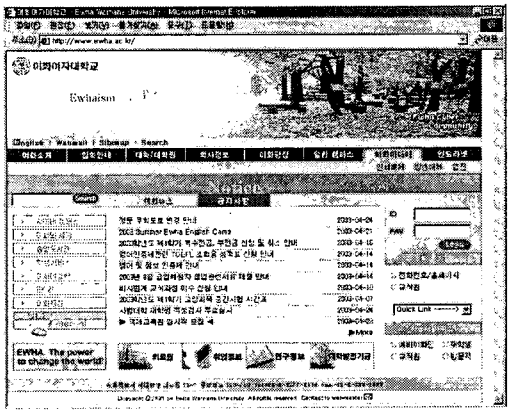
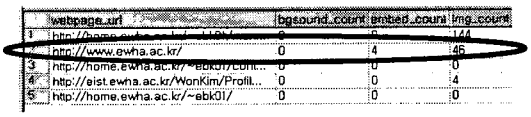


그림 4. 이화여자대학교 홈페이지



	webpage_url	bgsound_count	embed_count	img_count
1	http://home.ewha.ac.kr/~ebk01/cont...	0	0	144
2	http://www.ewha.ac.kr/	0	4	46
3	http://home.ewha.ac.kr/~ebk01/cont...	0	0	0
4	http://eist.ewha.ac.kr/WonKim/Profil...	0	0	4
5	http://home.ewha.ac.kr/~ebk01/	0	0	0

그림 5. 미디어 관련 태그 분석 결과화면

4. 결론

제안한 웹 환경에서의 분산 미디어 분석을 위한 Web Media Picker는 웹 페이지에서 여러 가지 미디어의 사용 정보를 분석하는 프로그램이다. 다양한 미디어의 태그 사용 횟수를 손쉽게 알 수 있고, 여러 사이트에서 미디어가 주로 이용되고 있는지, 또 어떤 방법으로 이용되고 있는지를 분석해 볼 수도 있다.

향후 미디어의 분석 방법이 웹 페이지에서 미디어의 사용 횟수와 사용 방법에서 그치지 않고, 좀 더 다양한 분석 방법을 접목시킬 것이다. 또한 웹 페이지에서 얻어낼 수 있는 정보는 미디어 정보뿐만이 아니다. 웹 페이지에 담겨있는 다른 여러 가지 정보들의 분석을 통하여 웹 페이지의 구조적인 정보, 다른 페이지로의 이동 경로 등도 알아낼 것이다.

[참고문헌]

[1] Tom Huang, Sharad Mehrotra, Kannan Ramchandran, "Multimedia Analysis and Retrieval System (MARS) Project," In Proc of 33rd Annual Clinic on Library Application of Data Processing-Digital Image Access and Retrieval, 1996.

[2] M. Adiba, R. Lozano, H. Martin, F. Mocellin, "Management of multimedia data using an Object-Oriented Database System," DEXA Workshop, pp.106-111, 1997.

[3] 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.

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

[5] 성낙운, 백철경, 조민규, "분산 멀티미디어 문서 그룹화를 위한 개념적 클러스터링," 컴퓨터산업교육기술학회 논문지, VOL.04, NO.01, pp.23-30, 2003

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