웹페이지의 분석과 활용을 위한 Crawler 구현

서승기*, 김진욱*, 조명진*
*고려대학교 전기전자전파공학부
e-mail: euney@korea.ac.kr

Implementation of Crawler for Webpage Analysis and Utilization

Seung-Gi Seo*, Jin-Wook Kim* and Myeongjin Cho*
*School of Electrical Engineering, Korea University

요 약

본 논문에서는 웹페이지의 내용을 분석하고, 또 분석한 정보를 바탕으로 웹페이지의 내용을 활용할 수 있게 해주는 Crawler의 개발에 대해 기술한다. 또한 단순한 정보의 수집뿐만 아니라 Crawler를 통하여 수집된 정보를 활용해 완전한 독립적인 프로그램을 생성할 수 있는 위젯 기능에 대해서도 자세히 기술하고자 한다.

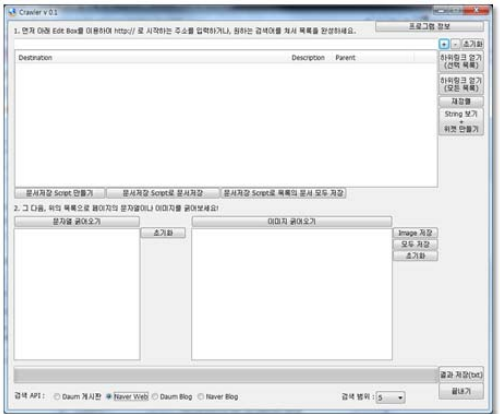
1. 서론

웹에는 수많은 정보들이 있다. 너무 많은 정보들이 있어서 오히려 정보의 바다에서 원하는 정보를 추출하는 것이 까다로운 일 중 하나가 되었다. 유사어 등으로 많은 검색의 오류가 있고 링크에서 링크로 이어지는 정보들도 많이 있다. 이에 사용자가 원하는 정보와 그와 관련된 정보를 웹으로부터 추출해주는 기술이 등장하고 있다. 그 대표적인 기술이 바로 Crawling이다[1]. Crawling은 무수히 많은 컴퓨터에 분산된 문서로부터 정보를 수집하는 기술을 말한다.

Crawling 기술을 기반으로 웹의 정보를 가져오는 Web Crawler는 웹상의 HTML로 기술된 정보를 파싱하여 분석하고 다양한 정보들을 가져올 수 있도록 설계된 프로그램이다. 기본적으로 검색어나 단어, 혹은 웹페이지 주소를 이용하여 웹페이지들을 분석한다. 분석된 웹페이지를 바탕으로 페이지내에 포함되어 있는 링크들을 바탕으로 하위의 Page를 무한대로 추출할 수 있는 것이 특징이다. 또한 웹페이지에 있는 정보(텍스트, 이미지)들을 통해 단어의 빈도수, 이미지 등을 얻어올 수 있다.

이에 본 논문에서는 Crawling 기술을 기반으로 하는 Crawler의 구현과 Crawler를 통해서 분석된 자료를 바탕으로 독립적인 프로그램을 생성할 수 있는 위젯 기능에 대해서 설명하고자 한다. Crawler는 그림 1과 같이 구성되어 있으며 프로그램을 구현을 위해 C/C++ 프로그래밍 언어를 사용하였고 Microsoft사에서 제공하는 MFC (Microsoft Foundation Class) Library를 사용하였다.[2]

논문의 구성은 다음과 같다. 2장에서는 프로그램이 지원하는 기능에 대하여 알아보고 3장에서는 기존의 Crawler를 활용하는 방안인 프로그램이 추출한 정보를 바탕으로 위젯을 생성하는 방법에 대해 설명한다. 마지막으로 4장에서 결론 및 향후 연구 방향에 대하여 논의한다.

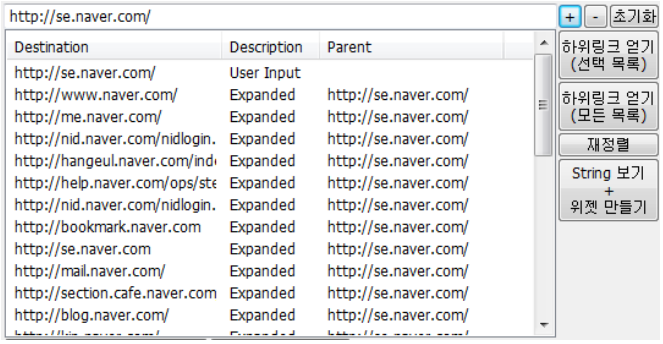


(그림 1) Crawler

2. 웹페이지 분석 지원 기능

2.1 분석할 페이지 확장

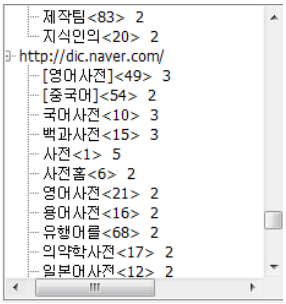
그림 2는 상위 페이지에서 파생될 하위 페이지들을 쉽게 얻을 수 있는 기능을 보여주고 있다. 각각의 웹페이지들은 마우스 더블클릭을 통하여 하위페이지의 링크들 얻을 수 있다. 또한 별도의 메뉴로 얻어진 전체 페이지 목록에 대해서 하위 페이지들을 가져올 수도 있다.



(그림 2) 하위 페이지 확장

2.2 페이지의 단어 분석 기능

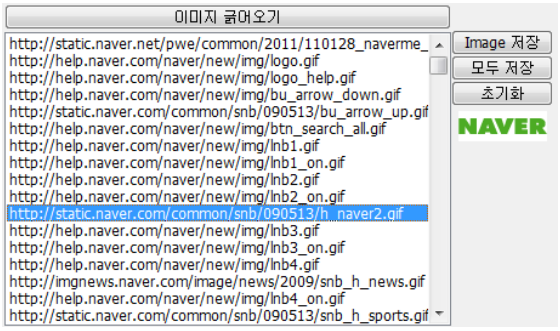
그림 3은 페이지의 전체적인 경향을 파악할 수 있는 단어 분석 기능을 보여준다. 단어의 Index와 출현 횟수를 볼 수 있으며 기본적으로 2번 이상 같은 단어가 반복적으로 출현할 때만 출력하게 구현되어있다.



(그림 3) 단어 분석

2.3 페이지의 모든 이미지 가져오기

그림 4에서 볼 수 있는 것처럼 웹페이지에는 이미지 링크 정보를 가지고 있다. 이 링크를 통해서 얻어진 이미지는 Crawler 디렉토리 아래에 일괄적으로 저장하거나 Crawler에서 제공하는 미리보기 기능을 통해서 하나씩만 저장할 수도 있다.



(그림 4) 페이지의 이미지 가져오기

2.4 분석 결과 저장

Crawler는 웹페이지 분석 결과들을 txt형식으로 저장할 수 있는 기능을 제공한다. 이 기능을 통하여 웹페이지 간의 연결 관계, 또는 웹 리소스간의 연결 관계를 간단하게 파악할 수 있다.

3. 분석 결과의 활용

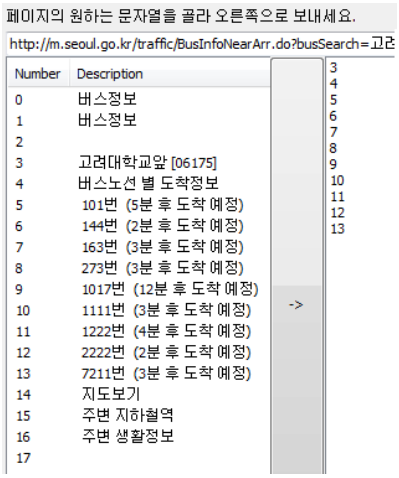
3.1 위젯의 생성

HTML DOM 구조를 따라서 파싱을 하게 되면, 정보를 가지는 페이지를 단순한 단위로 쪼갤 수 있다. Crawler의 위젯 기능은 이 점에 착안한 기능으로, 웹페이지를 분해하여 다시 재조립하는 기능이다.

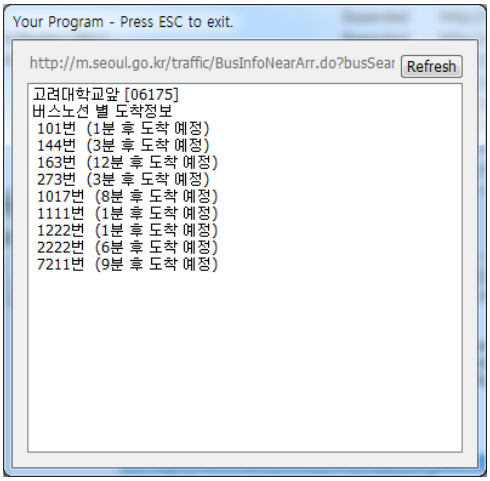
위젯은 그림 5와 같이 추출된 웹 페이지의 문자열이 번호와 함께 출력되며 원하는 정보를 선택하고 위젯 만들기 기능을 이용하여 위젯을 생성할 수 있게 된다.

3.2 위젯 이용

생성된 위젯은 하나의 완전한 프로그램으로 동작할 수 있다. 기본적으로 HTML의 DOM 구조에 번호를 붙여, 선택한 번호의 텍스트만 출력시키는 원리이다. 완성된 위젯은 그림 6과 같으며 그림 6은 그림 5에서 생성한 위젯을 실제로 실행시킨 모습이다.



(그림 5) 위젯의 생성



(그림 6) 완성된 위젯

4. 결론 및 향후 연구 방향

본 논문에서는 웹페이지를 분석하고 그 결과를 바탕으로 위젯을 생성할 수 있는 Crawler를 설계하고 구현하였다. 향후 분석기능을 그래프처럼 시각화하고 위젯을 사용자가 보기 편하도록 테마를 설정할 수 있게 하는 등 UI적인 부분을 향상시킬 예정이다.

참고문헌

[1] Pinkerton, B. (1994). Finding what people want: Experiences with the WebCrawler. In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

[2] Microsoft Developer Network, http://msdn.microsoft.com/