Proceedings of KIIS Spring Conference 2008 Vol. 18, No. 1.

웹크롤러의 비표준 링크에 관한 링크 추출 방안

A Method of Link Extraction on Non-standard Links in Web Crawling

정준영[1], 장문수[2], 강선미[3]

[1] 서울시 성북구 서경대학교 소프트웨어학과
E-mail: grounder@naver.com
[2] 서울시 성북구 서경대학교 소프트웨어학과
E-mail: cosmos@skuniv.ac.kr
[2] 서울시 성북구 서경대학교 컴퓨터과학과
E-mail: smkang@skuniv.ac.kr

요 약

웹크롤러는 웹페이지 내의 URL링크를 추적하여 다른 문서를 수집한다. 국내의 상당수 웹사이트는 웹 표준에 맞지 않는 링크방식으로 웹문서를 연결하고 있다. 일반적인 웹크롤러는 링크의 비표준적인 사용을 가정하지 않기 때문에 이러한 문서는 수집할 수 없다. 비표준적인 링크가 가능한 것은 사용자의 실수에 강인한 마크업 언어인 HTML에 자바스크립트 기능이 추가되면서 자바스크립트의 변칙적인 사용이 허용되었기 때문이다. 본 논문에서는 230여개의 웹사이트를 조사하여 기존 웹크롤러에서 해결하지 못한 링크 추출 문제를 찾아내고, 이를 수집하기 위한 알고리즘을 제안한다. 또한 자바스크립트 문제 해결을 위한 무거운 자바스크립트 엔진을 대신하여 필요한 기능만으로 구성된 모듈을 사용함으로써 효율적인 문서 수집기 모델을 제안한다.

Key Words : Web crawler, Link abstraction, URL link, Non-standard link, Javascript

1. 서 론

웹(Web)을 이용한 정보전달을 하기 위한 수요가 점차 늘어나고 있다. 웹 페이지(Web Page) 제작 권고안 준수도 그만큼 중요해지고 있다.

일반적으로 하나의 웹사이트는 다수의 웹문서들로 구성된다. 웹 크롤러(Web Crawler)는 사이트내의 한 문서를 선택하고, 선택한 문서로부터 같은 사이트에 속하는 문서들을 수집하기 시작한다. 일반적인 웹크롤러는 HTML페이지에 있는 하이퍼링크만으로 수집 경로를 형성한다.

기본적으로 웹크롤러는 HTML 권고안에 맞게 작성된 문서를 대상으로 한다. 그래서 표준에 맞지 않게 작성된 웹 사이트의 수집에는 어려움이 있다.

국내의 대부분의 웹 사이트는 마이크로소프트(Microsoft)사의 인터넷 익스플로러(Internet Explorer)라는 브라우저에서 문제없이 작동하는 것을 목표로 만들어 졌다.[1] 문제는 특정 웹 브라우저에서만 접근 가능한 웹 페이지에서는 변칙적인 방법을 이용해 웹 페이지를 작성하였기 때문에 수집이 용이치 않다.

본 논문에서는 웹크롤러에서 수집하기 힘든 문서연결 방식을 가진 웹사이트를 조사하고 다양한 링크방법에 대처할 수 있는 링크 추출 방안을 제시한다.

2. 연구 배경

<> 웹문서를 수집하기 위해 URL(Universal Resource Locator)을 이용해 접근하는데, 일반적인 웹 페이지는 특수한 목적을 가지고 있지 않다면 URL을 HTML 문서의 앵커태그에 링크를 연결한다. W3C의 HTML권고안[2]을 보면 웹 문서의 하이퍼링크(Hyperlink)를 하는 방법을 알 수 있다.

웹의 서비스가 다양해지고 웹사이트의 방문자가 다양한 기능을 요구함에 따라 자바 스크립트(Java Script)를 이용하여 링크를 연결하는 경우도 있다. 특수한 목적을 가지지 않은 보통의 웹크롤러는 이런 자바 스크립트를 처리할 수 있는 엔진(Engine)이 없기 때문에 자바 스크립트를 이용하는 웹사이트의 자료를 전부가지고 올 수는 없다. 또한 이를 처리하기위해

해당 엔진을 이용하면 시스템에 부하가 늘어난다. 또한 자바 스크립트를 이용하여 링크를 연결할 때에는 다양한 태그에 링크를 걸 수 있는데 이는 웹크롤링에 장애물이 되고 있다.

3. 링크 추출기 성능향상

웹크롤러는 인터넷상의 월드 와이드 웹(World Wide Web)을 자동으로 이동하면서 문서를 수집하여 데이터베이스화하는 도구이다. 다른 말로는 웹스파이더(Web Spider), 웹로봇(Web Robot)로 쓰인다.

본 논문에서 제안하는 웹크롤러는 이전에 제안되었던 시스템을 사용한다.[3] 이 시스템은 링크 추출 모듈, 링크 필터링 모듈, 문서수집 모듈로 구성된다.

링크 추출 모듈은 수집 대상 웹페이지에서 다른 페이지로의 링크를 추출하는 모듈이다. 본 논문에서 제안하는 알고리즘을 이용하는 모듈로 기존의 웹크롤러에서 처리가 힘들었던 부분을 해결하기위해 알고리즘을 새로 수정하였다.

링크 필터링 모듈은 웹페이지에서 추출된 링크를 수집하기 용이한 URL로 바꿔주는 역할을 한다. 링크 추출 모듈에서 수집된 링크 주소만으로 웹페이지를 접근할 수 있는 링크도 있지만, 상대주소나 자바스크립트를 이용한 링크는 문서수집 모듈에서 바로 사용할 수 없기 때문에 링크 필터링 모듈을 이용해야한다.

문서수집 모듈은 링크 필터링 과정을 마친 URL들의 중복여부를 판단하여 로컬에 각각의 페이지를 저장하는 역할을 한다.

3.1 링크 종류 조사

일반적인 웹크롤러에서는 그림 1에 나온 기본적인 링크에서 하이퍼링크 주소를 추출한다. 본 논문에서는 그림 1의 형태외의 링크를 추출하기위해 자바스크립트 함수분석과 분석결과 링크로 판단된 함수들을 추출한다.

```
<a href="하이퍼링크 주소“>표시 텍스트</a>
```

그림 1. 기본 하이퍼링크 예

링크 추출을 어렵게하는 자바스크립트를 사용한 하이퍼링크는 특정 태그에 국한되지 않고 표시하고자 하는 태그의 자바스크립트 이벤트 속성(Attribute)을 이용한다. 이러한 특징은 웹크롤링 과정에서 링크 추출을 방해하는 요인이 된다. 또한 기본적인 링크방법에 주소를 쓰지 않고 '#' 텍스트를 입력하여 웹브라우저에서 소스를 보지 않으면 해당 링크의 주소를 보이지 않게 하는 경우도 있다. 그림 2는 이런 변칙적인 링크 방법들의 예이다.

```
<a href="#“ onClick="javaScript:자바스크립트
함수명(매개변수)">표시 텍스트</a>

<HTML태그  onclick="자바스크립트
함수명(매개변수);">표시 텍스트</HTML태그>
```

그림 2. 변칙적인 링크 예

본 논문에서는 웹사이트별로 링크방법의 유형을 조사하기 위하여 사용자들의 방문이 많은 쇼핑몰, 가격비교, 도서, 기업협회, 인물정보, 학회, 각종 커뮤니티 등 260여개의 웹사이트를 분석하였다. 한 개 이상의 링크가 자바스크립트를 사용하는 웹사이트를 조사한 결과 54개의 사이트를 찾을 수 있었다.

자바스크립트를 사용한 웹사이트 54개에서 링크를 걸은 태그를 조사하였다. 표 1은 각 웹사이트의 페이지 소스를 분석한 결과 태그별로 어떠한 태그에 많은 사용빈도를 나타내는지를 보여주는 표이다.

표 1. 링크방법 종류별 사이트 갯수

태그	속성	사이트 수
<A>	href	9개
	onClick	39개
<SPAN>	onClick	2개
<TD>	onClick	4개

자바스크립트로 페이지이동을 하는 방법은 3가지 방법을 이용하고 있었다. 각각 자바스크립트의 location 객체를 이용하여 위치를 이동하는 방법, windows 객체를 이용하여 새로운 창을 여는 방법, document객체의 form을 이용하여 페이지를 바꾸는 방법이다.

대부분의 방법은 사용자에게 이동하려는 페이지 정보를 보이지 않게 하기 위한 방법이었다. 자바스크립트를 사용하여 링크하는 방법은 다양하였다. 하지만 웹크롤러가 링크에 관련된 함수를 분석하기 위해 자바스크립트 엔진 전체를 이용하지 않아도 된다고 판단되었다.

3.2 링크 추출 알고리즘 제안

이전의 웹크롤러들의 링크 추출 방법은 앵커 태그의 하이퍼링크 주소를 추출하는 방식이다. 그 방법은 웹 페이지에서 최대한 많은 링크를 추출하지 못하거나, 아예 링크자체를 수집하지 못하였다. 본 논문에서 제안하는 추출기는 다양한 형식의 링크들을 대부분 추출할 수 있도록 구현하였다. 그림 4는 본 논문에서 제안하

는 링크 추출 알고리즘의 흐름도이다.

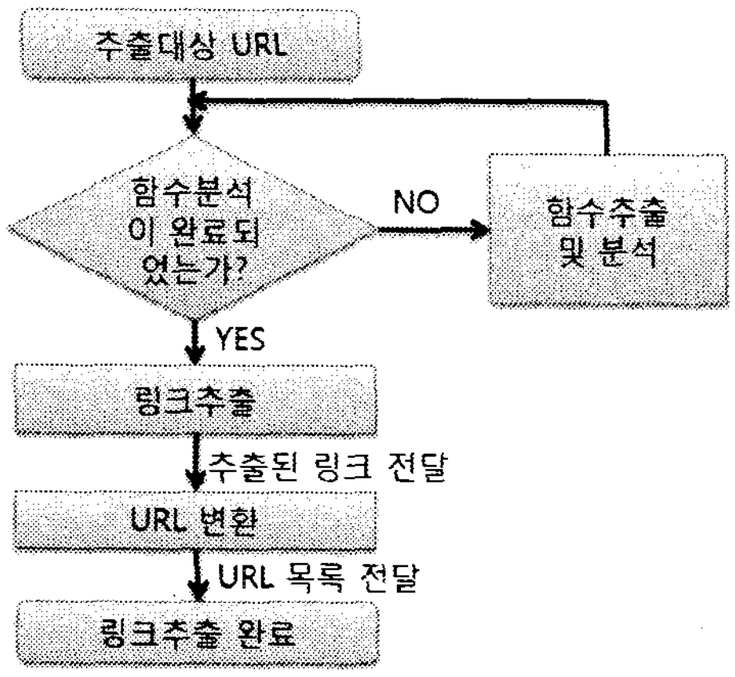

그림 4. 제안 링크 추출 알고리즘 흐름도

추출 알고리즘에 따라, 추출대상 URL을 시작으로 분석을 시작한다. 추출대상 URL에 해당하는 웹 페이지를 수집한다. 수집된 웹 페이지가 속한 웹사이트에서 자바스크립트 함수 분석 여부를 먼저 확인한다.

함수의 분석이 되지 않았을 경우, 해당 웹페이지에서 자바스크립트 코드를 모두 추출한다. 자바스크립트 코드는 두 가지 형태로 저장이 된다. 첫 번째로는 페이지 내부에 코드가 존재하는 경우이다. 두 번째로는 별도의 파일을 두어 필요한 코드를 별도의 파일로 수집하는 경우이다. 그림 5는 이런 두 가지 방식으로 자바스크립트 코드를 표시하는 예를 보여준다.

자바스크립트 코드를 추출한 후, 해당 함수별로 분석이 필요하다. 링크에 사용된 모든 함수를 찾아내는 과정이다. 3.2절에서 조사한 결과를 보면 링크를 위해 사용된 함수 내부에는 3가지 방식으로 웹페이지를 이동한다. 자바스크립트의 “windows”객체의 “open” 메서드를 이용해 창을 띄우는 것, “location”객체의 ”href“ 값을 바꿔 페이지 이동하는 것 그리고 ”form”객체를 생성하여 “submit”메서드를 실행해 데이터를 전달하는 방법이다. 세가지 객체를 이용하는 함수를 찾는다면 그 함수는 링크에 필요한 함수이다.

분석된 함수를 기반으로 앞 3.4절에서 언급한 URL패턴 스크립트에 링크 변환 방법을 입력하면된다.

링크 추출 과정은 기본적으로 HTML 의 ”A” 엘리먼트의 “href”속성 값을 모두 수집하는 것과 앞 단계에서 분석된 함수중 링크에 관련된 함수가 사용된 엘리먼트의 속성값들을 수집하는 것이다. 이런 과정을 거치면 정상적으로 링크를 걸지 않은 방법들을 모두 해결할 수 있다.

```
<html>
 <head>
 <script type="text/javascript" src="/js/www.js?" />
 <script>
<!--
function move(Value){
 ...
 window.location = url;
}
-->
 </script>
 </head>
 <body> ..... </body>
</html>
```

그림 5. URL 패턴 스크립트

URL 변환과정은 URL 패턴 스크립트를 이용해 앞 단계에서 추출된 링크들을 기본 URL로 변환하는 과정이다. 이 과정을 완료하면 최종적으로 수집을 하기위한 URL들을 모두 얻을 수 있다.

4. 실험 및 결과

웹크롤러는 자바를 이용해 웹 페이지를 액세스하는 방법을 사용했다. 실험 대상 웹사이트는 조사대상 260개 사이트 중 자바스크립트를 이용한 링크를 통하지 않으면 해당 정보를 볼 수 없는 5개의 웹사이트를 선정하여 실험을 하였다.

각 웹사이트 별로 100개의 중복되지 않은 임의의 웹 페이지에서 HTML표준 링크와 자바스크립트를 이용한 모든 링크를 수집하였다.

표 2 는 10개 웹사이트에 대하여 각 사이트별로 100개의 대상 웹페이지를 대상으로 링크를 추출한 결과이다. 본 실험에서 선정한 웹 사이트들은 대부분 제품, 인물, 기업에 관한 정보를 가지고 있는 사이트이다.

표 2. 사이트별 링크 추출 결과

웹사이트	순수 링크	변칙 링크
가격비교1	1830	20905
가격비교2		
인물정보1		
기업정보1		
기업정보2		

5. 결론

본 논문에서는 웹크롤링 과정에서 링크 추출의 성능향상을 위한 알고리즘을 제안 하였다. 프로그램 제작에 비해 비교적 자유로운 웹 사

이트 제작은 많은 정보를 사용자들에게 전달하는 역할을 했지만, 사용자들의 눈에 보이지 않는다는 이유로 HTML 문서 작성 표준에 대해 무감각해왔다는 사실을 알 수 있다.

웹 표준이 지켜지지 않은 웹사이트가 점차 늘어남에 따라, 표준에 맞게 제작된 웹크롤러는 제 역할을 할 수 없었다. 본 논문에서는 자바스크립트를 이용하여 제작된 웹 사이트들을 웹 크롤링하는 기법을 제안했다.

참 고 문 헌

[1] 서응교, 김홍기, 서길수, "국제 표준을 통해 살펴본 한국 기업과 공공기관의 웹사이트 접근성의 현황과 개선 방안", 한국경영정보학회 9회 하계통합학술대회

[2] W3C Dave Raggett, Amaud Le Hors, Ian Jacobs, "HTML 4.01 Specification" (http://www.w3.org/TR/REC-html40/struct/links.html)

[3] 장문수, 정준영, "", 한국경영정보학회 9회 하계통합학술대회