웹 사이트의 웹봇 접근성 평가 프로그램

윤원기, 박용희, 김성환, 김민주, 김석일
충북대학교 소프트웨어학과
e-mail : ywk0716@chungbuk.ac.kr

Web Bot Accessibility Evaluation Program for A Web Site

Won-Ki Yun, Yong-Hwi Park, Seong-Hwan Kim, Min-Ju Kim, Suk-Il Kim
Dept. of Software Engineering, ChungBuk National University

요 약

최근 DDOS 공격이 활발해짐에 따라 많은 사이트에서 검색 페이지에 노출되지 않도록 웹봇을 차단한 경우가 생기게 되었다. 하지만 이는 사이트 사용자나 사이트 측 모두에게 악영향을 미치므로 웹 사이트의 웹봇 접근성을 평가하는 프로그램을 제작하게 되었다. 본 프로그램은 Robots.txt를 분석하거나 Robots 메타 태그 등을 분석하여 웹 사이트의 웹봇 접근성을 평가하는 프로그램이다. 이 프로그램으로 평가한 결과를 피드백의 근거로 삼아 더 나은 검색 결과를 기대할 수 있다.

1. 서론

최근 각종 웹 사이트에 대한 DDOS 공격이 활발해짐에 따라 각 정부 지자체와 공공기관의 웹 사이트를 포함한 다양한 웹 사이트에서 검색 사이트의 검색 결과에 웹 사이트가 노출되지 않도록 하거나 검색을 차단하는 경우가 나타나게 되었다. 하지만 이 경우, 웹 사이트를 방문하는 사용자는 웹 사이트를 방문하기 힘들어지게 되고 웹 사이트를 운영하는 기관은 사용자의 방문이 줄어들게 되므로 웹 사이트의 검색 차단은 양측 모두에게 악영향을 미치게 된다. 본 논문에서는 검색 사이트의 검색 결과를 수집하는 프로그램의 일종인 웹봇이 특정 웹 사이트에 접근 가능하고 정보를 수집할 수 있는지에 대해 평가하기 위한 목적의 프로그램의 설계와 구현에 대해 다룬다.

2. 관련 연구

웹봇은 프로그램의 일종으로 웹 크롤러(Web Crawler) 라고도 불린다. 웹봇은 보통 웹 사이트를 돌아다니며 정보를 수집하는 용도로 사용되거나 검색 사이트에서 검색 결과를 나타내기 위해 웹 사이트의 인덱스와 검색 정보를 수집하는데 사용된다. 일반적으로 웹봇의 성능을 향상시키기 위한 다양한 연구가 존재하는데, 웹 크롤러의 신뢰성 향상에 대한 연구 [1] 나 효과적인 웹 크롤링[2] 과 같은 연구가 대표적이다. 그러나 이 경우는 웹봇 자체의 성능에만 관심을 둔 경우이고 웹 봇의 웹 사이트 접근성에 대한 연구는 거의 찾아볼 수 없다. 웹봇의 접근성이 아닌 일반적인 웹 접근성을 평가하는 평가 도구인 K-wah 의 경우, 한 번에 여러 웹 페이지를 평가할 수 있고 자동으로 접근성을 평가해주지만 웹봇의 접근성을 평가할 수 없고 자동으로 평가할 수 있는 일부분의 검사만 수행할 수 있다는 문제점과 실시간으로 렌더링된 웹 페이지가 아닌 HTML 소스코드만으로 평가를 하는 문제가 있다.

3. 기존의 문제점과 동작 원리

3.1 기존의 문제점

여러 웹 사이트에서 웹봇을 차단하거나 웹봇이 접근하기 힘들게 하는 등 다양한 문제점이 있음에도 불구하고, 이에 대한 평가를 할 수 있는 종합적인 도구 등이 존재하지 않거나 매우 미흡하였다. 또한 웹 표준들을 지키지 않은 사이트가 많이 존재하여 이로 하여금 웹 사이트에 대한 웹봇의 접근성 평가를 할 수 있는 도구를 만들 필요성이 생기게 되었다.

3.2 웹봇의 동작 원리

문제점을 해결하기 위해서는 먼저 웹봇이 어떻게 동작하는지를 먼저 알아야 한다. 웹봇마다 조금씩 차이는 있겠지만 대부분 [그림 1] 과 같은 방식으로 동작한다.

(그림 1) 웹봇의 동작 순서

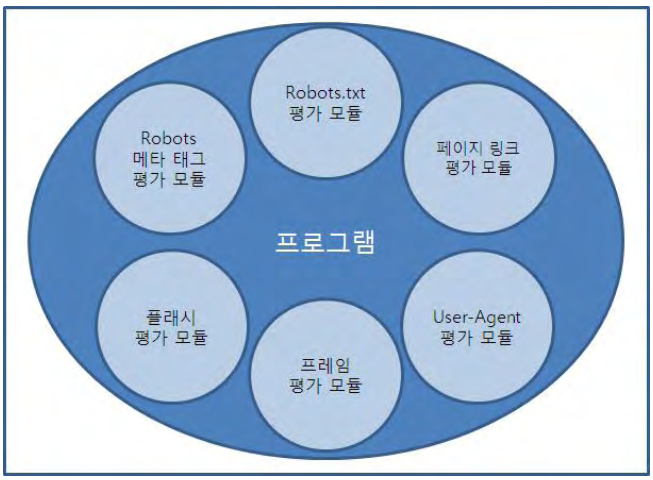
먼저, 웹봇은 웹 사이트의 최상위 경로에 위치한 Robots.txt 파일을 확인한다. Robots.txt 파일은 웹봇의 User-Agent 별로 접근 가능한 경로와 접근 불가능한 경로를 표시해주어 웹봇의 특정 경로의 접근을 막는 기능을 한다. Robots.txt를 확인한 후 웹봇은 웹 사이트의 메인 페이지에 접속하여 Robots 메타 태그를 확인한다. Robots 메타 태그는 웹 페이지의 HTML 소스에 존재하는 meta 태그 중 하나로 각각의 웹 페이지에 대해 웹봇의 개별적으로 작동하도록 사용된다[3]. Robots 메타 태그를 확인한 후 웹봇은 페이지에서 정보를 수집하고 페이지의 링크를 수집하여 현재 페이지에서 다른 페이지의 경로를 얻어와 다시 해당 페이지들로 이동하여 Robots 메타태그를 확인하고 정보를 수집하는 방식으로 동작하게 된다.

4. 구현 및 실험

4.1 프로그램 요구사항과 개발 환경

프로그램은 Windows XP 서비스팩 2 이상의 운영체제와 1GB 이상의 메모리카드, 100MB 이상의 여유 공간을 갖는 하드디스크를 가진 PC에서 동작한다. 개발 도구는 MS Visual Studio 2013을 사용하였고 개발 언어는 C# Windows Form Application을 사용하였고 프로그램의 실행에는 .Net Framework 4.0 이상의 버전이 필요하다.

4.2 프로그램 구성



(그림 2) 프로그램의 구성

(그림 2)와 같이 프로그램은 6개의 평가 모듈로 구성된다. 각 평가 모듈은 독립적인 평가 기능을 제공한다.

(그림 3)은 프로그램의 실행화면을 보여준다.

<그림 3> 프로그램 실행화면

4.2.1 Robots.txt 평가 모듈

이 평가 모듈에서는 자동으로 웹 사이트 최상단의 Robots.txt를 분석하고, 사용자로 하여금 특정 웹봇에 대해 웹 사이트에서 특정 경로의 접근을 허용 또는 차단하였는가를 평가할 수 있는 기능을 지원한다.

4.2.2 Robots 메타 태그 평가 도구

이 평가 모듈에서는 자동으로 웹 사이트 페이지의 Robots 메타 태그를 분석하여 사용자에게 현재 페이지의 Robots 메타 태그 정보를 알려주어 평가할 수 있도록 도움을 주는 기능을 지원한다.

4.2.3 페이지 링크 평가 도구

웹 사이트에서는 다른 페이지로 이동하기 위하여 일반적으로 링크를 제공한다. 보통 웹봇은 링크의 주소를 수집하여 웹 사이트를 탐색하지만 종종 이 주소가 자바스크립트로 이루어져 있는 경우가 있다. 이 경우, 해당 자바스크립트를 실행하기 전에는 이동될 페이지의 경로를 알 수 없게 되므로 웹봇이 웹 사이트를 탐색하는데 큰 어려움을 주게 된다. 이 평가 모듈에서는 현재 페이지의 링크들을 수집하여 자바스크립트로 이루어진 링크들을 찾아내고, 이를 화면에 표시하여 사용자로 하여금 해당 링크의 중요도를 직접 판단할 수 있게 하여 평가할 때 도움을 주는 기능을 지원한다.

4.2.4 플래시 평가 도구

플래시는 웹 사이트에서 여러 멀티미디어를 제공하기 위해서 사용되는 외부 플러그인이다. 일반적으로 플래시는 HTML로는 나타내기 힘든 여러 멀티미

디어 정보를 웹 사이트의 이용자들에게 제공하기 위해 사용된다. 하지만 플래시는 일종의 독립된 파일이기 때문에 웹봇으로 내용을 읽는 것이 거의 불가능하다. 따라서 이 평가 모듈에서는 현재 페이지에서 플래시를 찾아내고 이를 화면에 표시하여 사용자가 해당 플래시가 현재 페이지에서 얼마나 큰 부분을 차지하며, 웹봇의 정보 수집에 어느 정도 영향을 미치는지 평가할 수 있는 기능을 지원해준다.

4.2.5 프레임 평가 도구

프레임은 웹 사이트 페이지 내부에서 다른 페이지를 보여주기 위하여 사용되는 HTML 태그 중 하나이다. 일반적으로는 화면의 일부만을 프레임으로 사용하지만, 몇몇 웹 사이트의 경우 화면의 중요부분 또는 전체를 프레임 페이지로 사용하는 경우가 있다. 물론 거의 모든 웹봇이 프레임 안의 내용에 접근하여 웹 페이지를 평가 가능하지만, 페이지의 URL과 제목이 변하지 않게 되어 일반적인 웹 접근성에 위반되는 문제가 생기게 된다. 이 평가 모듈에서는 웹 페이지의 소스코드를 분석하여 프레임을 찾아내 사용자에게 나타내주며, 사용자가 해당 프레임을 확인하여 평가할 수 있도록 하는 기능을 지원한다.

4.2.6 User-Agent 평가 도구

User-Agent 란 사용자가 사용하는 웹 브라우저의 이름을 나타내는데 사용된다. 각각의 웹봇은 보통 고유한 User-Agent를 지니고 있는데, 일부 웹 사이트의 경우, 특정 웹봇의 User-Agent를 차단하여 웹봇이 해당 사이트에 접근하는 것을 막는 문제가 있다. 이 경우, 웹 사이트의 내용을 제대로 가져올 수 없어 웹 사이트를 검색 결과에 나타나게 할 수 없는 문제가 생기게 된다. 이 평가 모듈에서는 현재 페이지에 대해 유명한 웹봇의 User-Agent를 선택하거나 사용자가 직접 입력한 User-Agent를 입력하여 평가 페이지에 접속함으로서 웹 사이트의 User-Agent 차단 여부를 평가할 수 있도록 하는 기능을 지원한다.

4.3 프로그램 실험

일반적으로 자주 사용되는 웹 사이트와 다른 웹 사이트와는 다른 결과를 갖는 웹 사이트를 예로 실험을 해보았다.

<표 1> 네이버(http://www.naver.com) 평가 결과

평가 항목	평가 결과
Robots.txt	합격 (Robots.txt 없음)
Meta 태그	합격 (Robot Meta 없음)
Link	합격
Flash	합격
Frame	합격
User-Agent	합격

<표 1>은 현재 대한민국에서 가장 많이 사용되는 웹 사이트인 네이버를 프로그램으로 평가한 결과이다. 결과가 모두 합격인 것으로 보아 해당 웹 사이트는 웹봇이 접근하는데 어려움이 없는 사이트라는 것을 쉽게 알 수 있다.

<표 2> KBS(http://www.kbs.co.kr) 평가 결과

평가 항목	평가 결과
Robots.txt	불합격 (모든 하위 경로 비허용)
Meta 태그	합격 (Robot Meta 없음)
Link	합격 (자바스크립트 없음)
Flash	합격 (플래시 없음)
Frame	합격
User-Agent	합격

<표 2>는 대한민국 공영 방송사인 KBS의 홈페이지를 평가한 결과이다. 다른 항목은 합격이지만 Robots.txt 항목을 보면 모든 하위 경로 탐색 비 허용으로 인하여 웹봇이 웹 사이트의 하위 경로로 접근할 수 없도록 만들어 놓았다. 이런 경우, 웹 사이트의 메인 주소는 검색 결과에 나타나게 되지만, 해당 사이트 내부의 정보는 검색 사이트를 통한 검색으로 나타나지 않는 문제가 생기게 된다.

<표 3> 한화생명(http://www.hanwhalife.co.kr) 평가 결과

평가 항목	평가 결과
Robots.txt	합격 (Robots.txt 없음)
Meta 태그	합격 (Robot Meta 없음)
Link	합격 (자바스크립트 없음)
Flash	합격 (플래시 없음)
Frame	불합격 (페이지 전체가 프레임)
User-Agent	불합격 (차단됨)

<표 3>은 대한민국의 보험사중 하나인 한화생명의 홈페이지를 평가한 결과이다. 상위 4개 항목은 문제없이 합격하였지만 하위 2개 항목은 불합격인 것을 알 수 있다. 프레임 평가의 경우, 프레임 셋을 사용하여 페이지를 나누어 놓았기 때문에 웹 접근성을 위반하였다. User-Agent 평가의 경우, 주요 검색 사이트인 구글을 포함한 여러 웹봇의 접근을 차단해 버리는 문제가 발생하였다. 해당 사이트는 검색 사이트에서 검색 시 사이트의 URL 까지는 표시되지만, 메인화면을 비롯한 모든 페이지의 정보를 검색을 통

해서는 알 수 없게 되리란 걸 알 수 있다.

5. 결론

최근 스마트폰을 비롯한 전자기기가 대중적으로 사용되고 있는 현재 웹봇을 이용한 검색은 우리 삶에 큰 영향을 끼치고 있다. 하지만 웹봇에 관련된 연구들은 대부분 웹봇 자체의 성능 향상에만 초점을 맞추는 경우가 많으며 많은 웹 사이트에서 웹봇의 접근성을 지키지 않은 경우가 존재한다. 본 논문에서 소개한 프로그램으로 어느 정도 평가를 할 수는 있겠지만 이 프로그램의 경우, 앞서 언급한 K-wah와 달리 여러 웹 페이지를 동시에 평가할 수 없고 자동적으로 평가가 불가능하다는 문제와, 프로그램 자체는 프로그램 그 자체로는 단순한 평가 그 이상의 일을 할 수 없는 엄연한 단점이 존재한다. 하지만 사이트의 관리자 또는 사용자가 자신이 관리 또는 사용하는 사이트를 직접 평가함으로서 스스로 자신의 웹 사이트의 웹봇의 접근성을 높임으로서 자연스럽게 효과적인 검색 정보 수집이 가능하게 하여 좀 더 나은 검색 결과가 나타나는 긍정적인 효과를 기대할 수 있겠다.

감사의 글

이 연구는 NIPA의 2014년도 서울어코드사업의 지원을 받아 수행되었습니다.

참고문헌

[1] Viv Cothey, “Web crawling reliability”, 2004, Available http://onlinelibrary.wiley.com/doi/10.1002/asi.20078/pdf (accessed on AUG. 13, 2004)

[2] Carlos Castillo, “Effective web crawling“, 2005, Available http://dl.acm.org/citation.cfm?id=1067287 (accessed on JUN. 1, 2005)

[3] M. Carl Drott, Indexing aids at corporate websites: the use of robots.txt and META Tags, Information Processing and Management: an International Journal, v.38 n.2, p.209-219, March 2002

[4] Danny Sullivan, “How Search Engines Work”, 2002, Available http://www.uniroma2.it/didattica/prog_web/deposito/search_engine.pdf (accessed on OCT. 14, 2002)