심층 웹 문서 수집을 위한 크롤링 알고리즘 설계

원동현 · 강윤정 · 박혁규*

원광대학교

Crawling Algorithm Design for Deep Web Document Collection

Dong-Hyun Won · Yun-Jeong Kang · Hyuk-Gyu Park*

Wonkwang University

E-mail : dhwon79@wku.ac.kr / yjkang66@wku.ac.kr / hgpark7@wku.ac.kr

요 약

웹 기술이 발전함에 따라 웹은 사용자의 요구에 맞는 맞춤 정보들을 제공하게 된다. 클릭 이벤트나 사용자의 질의어에 따라 정보가 제공되며 검색엔진으로는 검색이 어려운 정보가 제공되는 웹 서비스를 심층웹이라 한다. 이러한 심층웹은 표면웹보다 많은 정보를 포함하고 있지만, 방문 당시의 정보를 수집하는 일반적인 크롤링으로는 정보 수집이 어렵다. 심층웹은 javascript와 같은 스크립트언어를 브라우저에서 실행함으로 서버의 정보를 사용자에게 제공한다. 본 논문에서는 심층웹 수집을 위해 스크립트를 분석하여 동적으로 변화되는 웹사이트의 탐색 및 정보 수집이 가능한 알고리즘을 제안한다. 본 논문에서는 실험을 위해 질병관리청의 게시판의 스크립트를 분석하였다.

ABSTRACT

With the development of web technology, the web provides customized information that meets the needs of users. Information is provided according to the input form and the user's query, and a web service that provides information that is difficult to search with a search engine is called an in-depth web. These deep webs contain more information than surface webs, but it is difficult to collect information with general crawling, which collects information at the time of the visit. The deep web provides users with information on the server by running script languages such as javascript in their browsers. In this paper, we propose an algorithm capable of exploring dynamically changing websites and collecting information by analyzing scripts for deep web collection. In this paper, the script of the bulletin board of the Korea Centers for Disease Control and Prevention was analyzed for experiments.

키워드

web, deep web, asynchronous javascript, archiving

Ⅰ. 서 론

정적인 웹페이지는 크롤러가 방문하는 순간 이미 모든 페이지가 완성된 형태로 제공 되어 웹 크롤러가 다음 링크를 참조하여 페이지를 이동하거나 정보를 저장하는 것이 가능하였으나, AJAX[1]와 같은 웹 기술의 발전으로 웹 페이지에 이용자가 방문 하면 그 다음에 관련 정보를 데이터베이스에서 로드하여 페이지를 생성하는 방법이 주로 이용하고 있다. 이와 같은 페이지는 내용이 동적으로 생성되며 웹 크롤러가 방문한 순간에 웹 페이지에는 일부 정보만 존재한다. 이처럼 페이지가 동적으로 생성되거나 또는 방문한 페이지에 검색폼을 이용해야만 페이지의 정보를 불러 올 수 있는 웹 페이지를 심층웹[2] 이라고 하며, 이러한 심층웹에서는 일반적인 크롤러로 해당 웹 페이지의 정보를 수집하기 어렵다. 그러나 심층웹이 가진 정보는 정적으로 구성되는 표면웹보다 약 450~550배 이상의 정보를 가지고 있을 것으로 추산하고 있으며, 따라서 이들 심층웹에 있는 정보들을 크롤러를 이용해 자동으로 수집하는 방안이 필요하다. netcraft [3]의 2022년 9월 기준 조사에 따르면 그림 1과 같

* corresponding author

이 전 세계에 1,129,251,133 개의 사이트가 존재하고 있다. 각각의 웹사이트가 제공하는 콘텐츠 등의 정보를 고려해본다면 웹에서 제공하는 정보의 양은 가름하기 어려울 만큼 많아지고 있다.

본 논문에서는 심층웹 수집을 위해 스크립트를 분석하여 동적으로 변화되는 웹사이트의 탐색 및 정보 수집이 가능한 알고리즘을 제안한다. 본 논문에서는 실험을 위해 질병관리청의 게시판의 스크립트를 분석하였다.

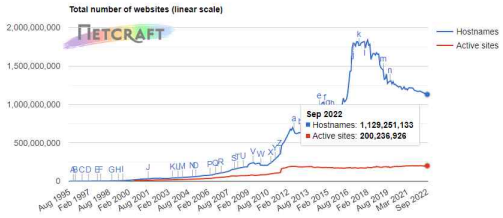

그림 1. Total number of websites

II. 관련 연구

크롤러는 사이트 방문이나 크롤러 실행 방식에 따라 집중 크롤러(Focused Crawler), 증분 크롤러(Incremental Crawler), 분산 크롤러(Distributed Crawler), 병렬 크롤러(Parallel Crawler)로 구분지어 지며 [4] 심층웹 크롤러로는 Deepbot, HiWE, Incremental Web Crawler가 있다.[5]

증분 크롤러 (Incremental Crawler) 방식은 크롤링 중인 콘텐츠 원본에 지정되어 있는 웹문서를 크롤링하여 마지막 크롤링 이후 수집되지 않은 정보를 수집하여 기존 크롤링 정보에 변경 내용만 추가하는 방법이다.

분산 크롤러 (Distributed Crawler) 방식은 웹문서를 수집하는 과정에 시간이 많이 소요되는 문제를 해결하는 방안으로 분산시스템 기반 크롤러로 크롤러가 동작하는 다수의 서버가 동시에 웹을 수집하고 중심역할을 하는 서버를 두고 각 크롤러 서버를 관리하여 웹문서를 수집한다.

병렬 크롤러 (Parallel Crawler) 방식은 웹의 규모가 커짐에 따라 하나의 프로세스 또는 스레드만으로 전체 웹 페이지를 수집하기에는 어려움을 해결하는 방안으로 개발되었다. 여러 개의 프로세스 및 스레드를 이용하여 대량의 웹 페이지들을 빠르게 수집한다.

Deepbot은 미니 웹브라우저라는 클라이언트 스크립트 실행 도구를 내장하여 서버와의 세션 유지 및 클라이언트의 스크립트 실행이 가능하여지도록 하여 데이터를 수집한다.

HiWE은 웹 질의어 인터페이스에 숨겨진 데이터를 추출하기 위해 입력폼을 분석하고 입력폼에 값을 전달하는 방식으로 심층웹을 수집한다.

Incremental Web Crawler는 심층웹이 제공하는 정보의 변화에 즉시 반영된 결과를 저장하기 위한 크롤러로 크롤러가 웹페이지를 방문하는 방문주기를 확률적으로 정하고 웹페이지 변화 주기를 계산하여 재방문하여 정보를 수집한다.

III. 심층웹 수집 알고리즘 설계

본 논문에서는 심층웹 수집알고리즘을 설계하기 위해 질병관리청(https://kdca.go.kr/) 홈페이지를 분석하였다. 질병과 관련된 긴급한 정보를 비동기식 정보를 제공함으로 일반적인 크롤러가 즉각 웹 검색 엔진에 결과를 반영하기 어렵지만 심층웹 수집 알고리즘이 적용 가능하다면 동기식 주소를 반영할 수 있어서 활용도가 높기 때문이다. 일반적으로 중요한 정보를 게시하게 되는 알림・자료 메뉴의 게시판을 분석하였다.

표 1. 링크 분석 및 관리

게시판 이름	게시물 주소 변경
이달의건강소식	O
보도자료	X
홍보자료	O
공지사항	X
공고/공시	X
채용공고	X
법령・지침・서식	O

표 1과 같이 게시물 클릭 시 링크의 주소가 변경되어 고유 주소가 있는 경우 크롤러는 해당 정보를 수집하여 검색 결과를 보여줄 수 있다. 하지만 게시물 클릭 시 주소가 변경되지 않을 때는 같은 주소에서 서로 다른 정보를 제공하게 되므로 크롤러는 정확한 정보를 제공해주기 어렵다.

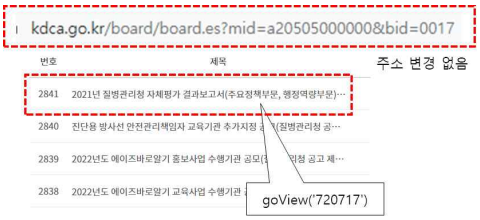

그림 2. 질병관리청 알림・자료 게시판

그림 2와 같이 게시물을 클릭하는 경우 주소창에는 아무런 변화 없이 웹페이지의 내용이 변경된다. 링크를 분석해보니 게시물 링크에 그림 3과 같이 스크립트 기반 페이지 이동 링크가 있었다. 그림 2와 같은 주소에서 크롬 콘솔창을 열고 에서 goView('720663') 명령어를 실행하자 해당 게시물에 직접 접근이 가능하였다.

```
onclick="goView('720663'); return false;
간용 방사선 안전관리책임자 교육기관 추가지
관리청 공고 제2022-307호..">
```

그림 3. 페이지 이동 javascript 코드

게시물 하나당 하나의 주소를 가지는 경우라면 저장된 주소를 불러와 직접 게시물을 참고하는 것이 가능하지만 그림 3와 같이 주소가 명령어로 되어 있는 경우는 게시판 주소를 알아도 게시물을 불러오기 어렵다. 하지만 게시판의 주소를 알고 있다면 스크립트 명령어를 해당 게시판에서 실행함으로 게시물을 보여주는 것이 가능하다. 이러한 성질을 이용하여 본 논문에서는 그림 4와 같이 심층웹 수집 알고리즘을 설계 하였다.

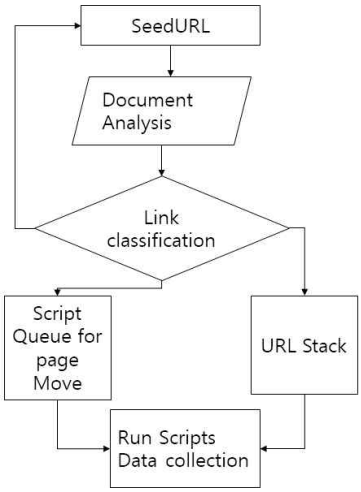

그림 3. 심층웹 수집 알고리즘

먼저 게시판 주소와 같이 정적인 주소들을 수집하여 Page URL로 Page URL Stack에 저장한다. 저장된 페이지로 이동하여 페이지 이동 관련 스크립트를 추출하고 추출된 스크립트는 스크립트 Queue에 저장한다. 모든 스크립트를 저장한다. 저장된 스크립트는 입력 순서대로 실행된다. 스크립트는 내장된 브라우저에서 수집 Page URL Stack에서 추출한 페이지가 로드되면 페이지 이동 스크립트들을 실행하게 되고, 스크립트 실행으로 페이지 내용이 변경되면 변경된 문서를 저장한다. 스크립트 Queue가 모두 실행되면 수집Page Url Stack에서 다음 URL을 가져오게 되며 동일한 과정을 수집 Page URL Stack에 정보가 없을 때까지 반복한다.

스크립트가 실제로 실행되었는지 표 2와 같이 판단한다. 스크립트는 실행 오류 및 응답 지연 또는 명령어 변경으로 인해 예상하지 못하는 결과를 보여줄 수 있음으로 표 2와 같이 평가하여 정확히 웹페이지가 수집되었는지 평가한다.

Ⅳ. 결 론

심층웹은 표면웹의 증가와는 비교가 되지 않을 만큼 빠른 증가를 보이고 있다. 사물인터넷의 발전 및 웹 콘텐츠 생산 주체의 다양화, 웹에 대한 이용자들의 요구 증가 등으로 심층웹의 증가 속도가 가속화되고 있다. 이러한 심층웹은 표면웹보다 약 450 ~ 550배 이상의 정보를 가지고 있는 것으로 추산된다. 이러한 정보들을 수집하기 위해 스크립트를 링크와 같이 활용하는 방법을 제안하였다. 스크립트를 링크로 활용하기 위해서는 사이트별로 스크립트에 대한 분석이 필요하지만, URL과 명령 스크립트를 함께 제공하면 심층웹에 대한 접근이 가능했다. 향후 연구에서는 심층웹 관련 스크립트들을 자동으로 수집 및 분석하여 관리자의 개입을 최소화할 수 있는 알고리즘을 연구하고자 한다.

References

[1] MDN Web Docs, AJAX[Internet]. Available : https://developer.mozilla.org/ko/docs/Web/Guide/AJAX.

[2] Ishan Pandya, Hitanshu Joshi, Biren, Patel, Harshil Joshi, “Threats That Deep Web Possess to Modern World,” *International Journal for Innovative Research in Science & Technology*, Jaipur, pp. 140-148, 2017.

[3] Netcraft. September 2022 Web Server Survey [Internet]. Available : https://news.netcraft.com/archives/category/web-server-survey/

[4] Rahul kumar, Anurag Jain, Chetan Agrawal, “SURVEY OF WEB CRAWLING ALGORITHMS”, Advances in Vision Computing: An International Journal. Vol. 3, No. 3, Sep. 2016.

[5] Desai Keyur, Devulapalli Virala, Agrawal Smita, Kathiria Preeti, Patel Atul, “Web Crawler : Review of Different Types of Web Crawler, Its Issues, Applications and Research Opportunities,” International Journal of Advanced Research in Computer Science, Vol. 8, Issue 3, pp. 1199-1202, Mar. 2017.