AN AJAX POWERED FORMS BASED APPROACH FOR CRAWLING THE DEEP WEB

F. Iftikhar, S. Gull, M. Shoaib*, S. Shoaib* and K. Karim*

Department of Computer Science, LCWU, Pakistan
*Department of CS & Engineering, University of Engineering & Technology, Lahore, Pakistan
Corresponding author: sh_sonia22@yahoo.com

ABSTRACT: The Deep Web is considered as a vital gap in search engine coverage. Large portion of the web is made up of deep web contents. Many systems and techniques have been proposed to surface the deep web content. Current techniques do not support crawl based web contents behind AJAX or JavaScript forms. The AJAX based forms are loaded dynamically at client side script due to which pre-cashing is not possible. Traditional crawlers are unable to crawl and index dynamic contents. A solution is presented in this paper to solve this problem. AJAX based forms have been analyzed to extract the relevant information from database. The information extracted is indexed for retrieval from database. Further work can be done to accommodate HTML or AJAX based information.

Keywords: AJAX, Web Crawler, Deep Web, Invisible Web

INTRODUCTION

Presently the major source of information is the World Wide Web (WWW). Using this source, information is shared among communities. The information is available in forms of video, audio, text and tables etc (Chang and Cho, 2006).The contents of deep web referred to that World Wide Web contents which are not related to surface web. A survey shows that the deep web is 500 times larger than the surface web (Kahttab et. al., 2009). It has been noticed that some of the important information might not be appeared in search results because of inability of search engines to crawl deep web. Search engines only indexed the surface web and leave the hidden or deep web behind (Brandman et. al., 1999). There are major issues which prevent to expose deep web in front of traditional crawler in the HTML and AJAX powered forms available on web page to access some specific information.

Virtual integration and surfacing techniques are used for crawling the web behind HTML forms. Domain specific and vertical search engines make use of virtual integration and standard search engines make use of surfacing (Faloutsos and Christodoulakis, 1987; Madhavan et. al., 2008). Both of these techniques leave behind the dynamically generated AJAX web pages (Ali et. al., 2008; Madhavan et. al., 2008, 2009). The entry point to the deep web is a form. When a crawler finds a form, it needs to guess the data to fill out the form (Raghavan and Molina, 2001; Barbosa and Freire, 2004,2005).

Today most web applications are AJAX based because it reduced the surfing effort of user and network traffic (Cristian et. al., 2008, 2009). The problem which is focused in this research is to crawl the web behind AJAX powered forms. JavaScript functions are used by AJAX. AJAX takes request from one page, sends it to other page and then shows results again on the first page (Iftikhar, 2010). Using the following URL in JavaScript function, the data are retrieved to surface the complete form through web crawler. This paper figures out a procedure for crawling and indexing data and URL's behind the AJAX powered forms.

MATERIALS AND METHODS

The following diagram illustrates the flow of working of proposed application.

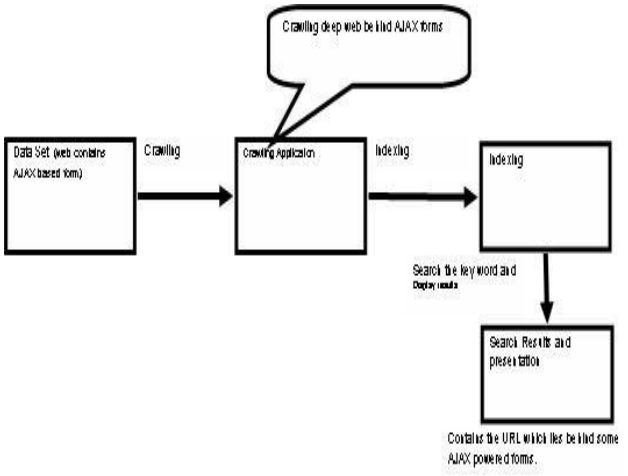

Fig. 1: The working of AJAX Crawler

The working of each component is giving as under:

Data Set (Web Contains AJAX Based Forms): A web page is made crawl-able and index-able in this step so that a user can access the information behind these forms without filling out the form.

Crawling Application: Crawlers application crawls behind AJAX forms by getting parameters from the URLs that are being called on different JavaScript events like "on Change".

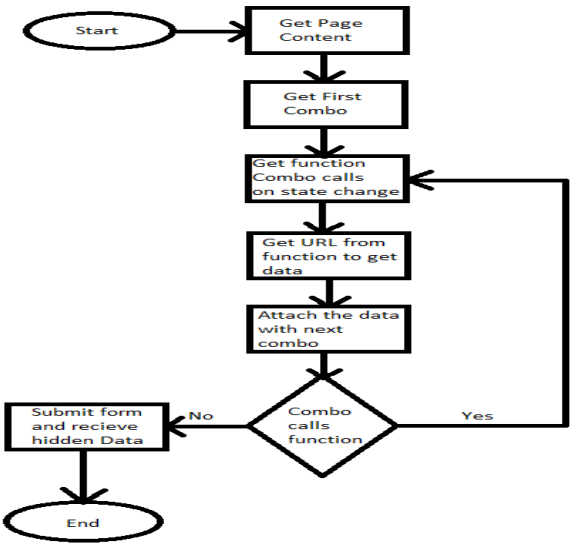

Figure 2: Crawling web or data behind AJAX powered form

Indexing: The proposed search engine's index is maintained by crawling the given URL, extract all links from it, save them in database with specific information and keyword.

Search Results and Presentation: The last step is how

to present search results to user. This is done by grouping the results in which many web pages contain the same keyword. When user make a query, query processor match it with index, find relevant match and display all them in result page.

Experimental Setup: A test web page which contains AJAX powered form and is crawlable. This web page is designed with some special properties defined by SEO to make it crawlable by AJAX crawler. The major properties are:

- Keep "ids" of all combos, values of all options, function name for "onChange" event in double quotes.
- The URL that the function calls to get the data for second combo should be in variable named URL and in double quotes

The test web page is designed with two combos. First combo contains the teacher names and second contains the courses, taught by the teachers. When user selects teacher from first combo, second combo dynamically populated with the courses he/she teaches. That courses data are coming from database. On selecting a course against a teacher the information of that course comes up. The focus of this experiment is to index the information behind this form.

RESULTS AND DISCUSSION

A test page was taken which contained the course information against the name of teacher. Crawler filled the form; crawl the information lie behind that form and then indexer saved the crawled information in database for future use. After crawling and indexing the data behind AJAX by the test crawler algorithm, a search engine was developed to check whether the data was crawled and indexed properly. The test search engine application contained a field and submit button. When the queries were passed to the search engine to retrieve the information, it fetched the URLs from database which contains the word that was in query string.

Several tests were made to test the crawler application's correctness. The list of list of keywords used for test the application was as under:

- Dr. Sonia
- Miss Mariam
- Computer Architecture
- Software Engineering

The results against these keywords are as follow:

Table 1: Search results against different search key words

Search Key Word	Total Number of URL Containing the Search Key Word.	Actual Number of URLs Containing the Search Key Word.
Dr. Sonia	2	2
Miss Mariam	3	3
Computer Architecture	1	1
Software Engineering	1	1

The results have proved that the technique to crawl the web behind AJAX forms has been implemented and working successfully.

This research has provided a solution to the problem of crawling the behind AJAX powered forms. There were many solutions to resolve this problem but all have their limitations. Some application developer has provided custom search engine or they expose web content to traditional search engine based on agreement. This was a manual solution and requires extra contribution from application developers. Some web developers have provided vertical search engine on their web site which was used to search specific information about their web site. Dynamic interface for users convenient and alternate static view for crawlers have also been provided. These solutions only discover the states and events of AJAX based web content and ignore the web content behind AJAX forms while the given in this paper has made it possible to crawl the web behind AJAX form by snipping the form entries, filling the form fields and getting the information behind that form.

CONCLUSION: In web forms, normally combo boxes are dynamically loaded and this is achieved by applying AJAX in the presented technique. The crawler application is able to fill the AJAX form and discover the information behind that form. It can be enhance by making a general technique for all kind of form items either HTML based or AJAX based.

REFERENCES

Ali, M., E. Bozda and A. V. Deursen. Crawling AJAX by Inferring User Interface State Changes, TUD-SERG. 408-437 (2008).

Brandman, O., J. Cho, H.G-Molina and N. Shivakumar. Crawler-Friendly Web Servers. Dept. of Computer Science, Stanford. 13-75 (1999).

Barbosa, L. and J. Freire. Siphoning Hidden-Web Data through Keyword-Based Interfaces. SBBD.309-321 (2004).

Barbosa, L. and J. Freire. Searching for Hidden-Web Databases, 8th International Workshops on Web and Databases. 44-45 (2005).

Chang, K.C.C. and J. Cho. Accessing the Deep Web:

From Search to Integration. SIGMOD, Chicago, Illinois, USA. 8: 27–29 (2006).

Cristian, D., F. Gianni, K. Donald, M. Reto, and C. Zhou. AJAX Search: Crawling, Indexing and Searching Web 2.0 Applications, VLDB. 1(2): 203-230 (2008)

Cristian, D., F. Gianni, K. Donald, M. Reto, and C. Zhou. AJAX Crawl: Making AJAX Applications Searchable. DOI 10.11.09/ICDE. 78-89 (2009).

Faloutsos, C. and S. Christodoulakis. Description and performance analysis of signature file methods, ACM TOOIS. 5(3): 237-257(1987).

Iftikhar F. Crawling the deep web behind AJAX powered forms, MSc. Thesis, Dept.of Computer Science, Lahore College for Women University, Lahore. (2010).

Kahttab, M. A., Y. Fouad and O. A. Rawash. Proposed Protocol to Solve Discovering Hidden Web Hosts Problem. IJCSNS. 9(8): 40-45(2009).

Madhavan, J., D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy. A Google's Deep-Web Crawl. PVLDB. 1(2):1241–1252 (2008).

Madhavan, J., L. Afanasiev, L. Antova and A. Halevy. Harnessing the Deep Web: Present and Future, CIDR. 394-405 (2009).

Raghavan, S. and H.G-Molina. Crawling the Hidden Web, VLDB.129-138 (2001).