International Journal of Computer Engineering & Technology (IJCET)
Volume 7, Issue 1, Jan-Feb 2016, pp. 36-44, Article ID: IJCET_07_01_005
Available online at
http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=7&IType=1
Journal Impact Factor (2016): 9.3590 (Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6367 and ISSN Online: 0976–6375
© IAEME Publication

STUDY OF DEEP WEB AND A NEW FORM BASED CRAWLING TECHNIQUE

Debraj Dey
Computer Science and Engineering Department,
MaulanaAbulKalam Azad University of Technology
Kolkata, West Bengal, India

Payel Das
Computer Science and Engineering Department,
Maulana Abul Kalam Azad University of Technology
Kolkata, West Bengal, India

ABSTRACT

The World Wide Web, abbreviated as WWW is global information medium interlinked with hypertext documents accessed via the internet. In a web browser a user can easily search the content by simply filling up a form. As the amount of information in the web is increasing drastically, the search result needs to be increased and it depends completely on the searching engine and the search engines are only as good as the web crawlers that serve up content for the result.

The paper gives an idea of a new hidden web crawling technique that is concerned with filling forms with meaningful values in order to get an appropriate search results.

Key word: WWW, Web Crawler, Deep Web, HTML, SQLI

Cite this Article: Chittineni Aruna and R. Siva Ram Pra. Study of Deep Web and A New Form Based Crawling Technique. *International Journal of Computer Engineering and Technology*, **7**(1), 2016, pp. 36-44.
http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=7&IType=1

1. INTRODUCTION

The Deep Web is vast, thousands of times larger than the visible internet, which is also called the surface web. But the Deep Web is not a place it simply accounts for all of the unindexed content online like banking data, administrative codes of government, corporations and universities. It’s like looking under the hood of the internet. The Deep Web cannot be accessed through traditional search engines like Goggle, Bing, Yahoo etc. because traditional search engines create their indices by crawling surface web pages [1]. This large amount of information in the Deep Web is only accessible through specific interfaces created by CGI or Common Gateway Interface and HTML forms or JavaScript. A software service called Tor, originally developed by US Military which is now an open source and publicly funded can also be used to access the Deep Web. It is important to understand that an ideal platform to accumulate all detailed information is nothing but the Deep Web which is worthwhile for journalists, government agencies and dissidents around the world.

In general there are two common approaches to access the Deep Web content [6]. One is to create a vertical search engine for specific domain, where semantic mapping is done in between individual data and mediator forms, and this technique is less accurate and certainly has some drawbacks. The other one is the surfacing technique, which is more common and effective approach. Our goal is to create a form based crawling technique and to equip a web crawler with appropriate input values in order to get accurate result.

The rest of the research paper is divided into six sections. In section II the characteristics and scale of the Deep Web is discussed. Then in section III our approach towards four categories of crawling technique is shown, including the architecture of form focused crawling, and followed by given experiment and its phases in section IV. Further in section V experiment results that are applied on people domain is shown. Finally, in section VI conclusion and future work is discussed followed by the references in section VII.

2. THE DEEP WEB

As time progresses the size of the Deep web is increasing exponentially and at a rate that defines quantification. In the verge of 20th century the WWW contained few documents and sites. Back then it was manageable task to post all the documents as static pages and the documents could easily be accessed through search engine by the help of web crawlers. But now the content of the web is accessed dynamically for maintaining a database. Since, the database is typically hidden from the traditional search engine; so, in order to access the content of the database sometimes we need to generate queries. The study by Bergman in the year 2001 suggests that the size of the Deep Web was 500 times larger than the Surface Web [1].But back then it was 29,254,370 websites available in the internet among which 75% were inactive [7]. According to University of California, Berkeley the Deep Web contains approximately 91,850 terabytes of data in 2003 [1]. A report published by IDC predicted that the total of all digital data created, replicated, and consumed will reach 6 zettabytes in 2014 [4]. By the end of 2014, NetCraft confirmed that the total number of websites increased to 1 billion [10]. Thus, according to the stats and opinion given by the experts suggest that

- 4.9% of the total web content is publically accessible.
- The Deep Web is nearly 4,000-5,000 times larger than the Surface Web.
- The Deep Web contains nearly 750 billion documents.

3. OUR PROPOSED APPROACH

3.1. Categories of Web Crawling

A web crawler is a program that visits webpages; forms etc. and reads their content and other required information like URL in order to create entries for the search engine index. In general there are mainly four categories of crawling technique which are:

3.1.1. Traditional Web Crawler

The traditional web crawlers do not distinguish between pages with and without forms [3]. The main components of traditional web crawler are frontier, fetcher, URL extractor, page filter and database. The architecture of traditional web crawler is shown in figure 1. First, the module gets a unique URL from a set of seed URLs which is later passed through frontier. Then the frontier sends the unique URL through fetcher, which fetches the content of the web. Then the content is passed through URL extractor. Later the extractor checks the pages to find new links. The new links are sent through HTML-page filler to eliminate uninteresting pages and finally the valid pages are stored in the database.

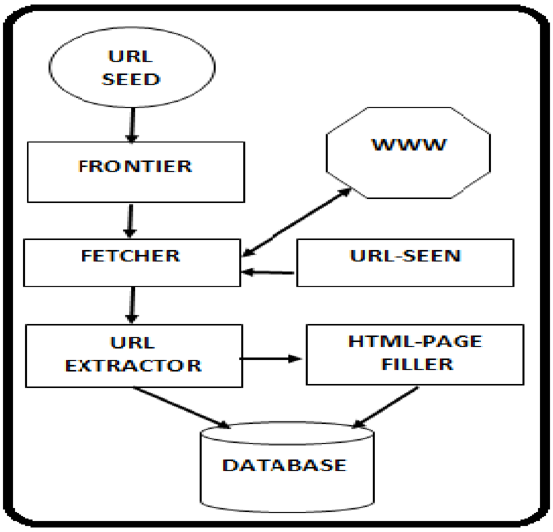

Figure1 Architecture of Traditional Web Crawler

3.1.2. Deep Web Crawler

There are millions of web pages which are indexed each and every day, but still a huge amount of information is hidden behind the web forms. So, in order to retrieve those information researchers are continuously developing the deep web crawler. Figure 2 shows the architecture of deep web crawler. The main components of deep web crawler are Input-Classifier, Domain Filter and HTML Analyzer. First, seed URLs, domain data and the user specific are sent as inputs of the Input-Classifier. The main function of Input-Classifier is to choose the HTML element and interact with the various inputs [8]. After that these data are sent through domain filter and later through HTML analyser to submit the forms to the web server. Finally, the Response-Analyser edits and also updates the content of the database.

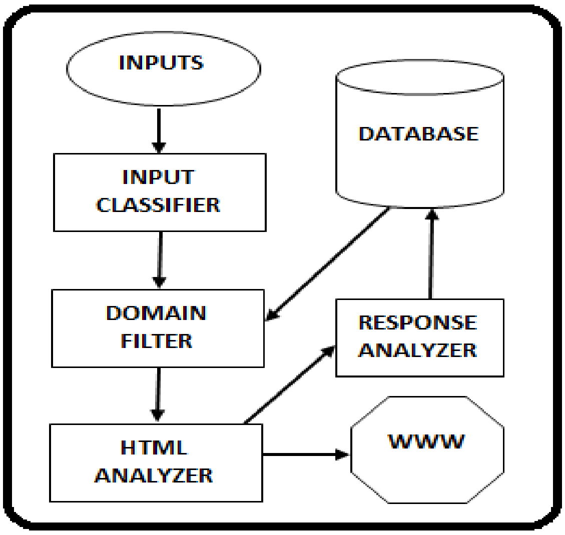

Figure2 Architecture of Deep Web Crawler

3.1.3. RIA Web Crawler

Over the last few decade, web application has become more interactive, responsive and user friendly. These applications are called Rich Internet Application (RIA).Certainly, accessing these applications are often time consuming because technologies like AJAX and DOM are used heavily in the RIA [5]. To reduce the time to crawl into RIA a good web crawler is required. The architecture of RIA Web Crawler is shown in Figure 3. First, JavaScript engine starts with a virtual browser and runs a JS engine, and then it retrieves the starting web page associated with an URL seed and loads it in the virtual browser. After that, the constructed DOM is passed to the DOM state to determine if this is the first time the DOM state is seen or not. To extract the JS events the DOM state is passed to the extractor. Then the JS event is passed to the strategy module. The strategy module decides which event is to execute. Then the chosen event is passed to the JS engine for the further execution. This process carries on until all the reachable DOM states are seen [9].

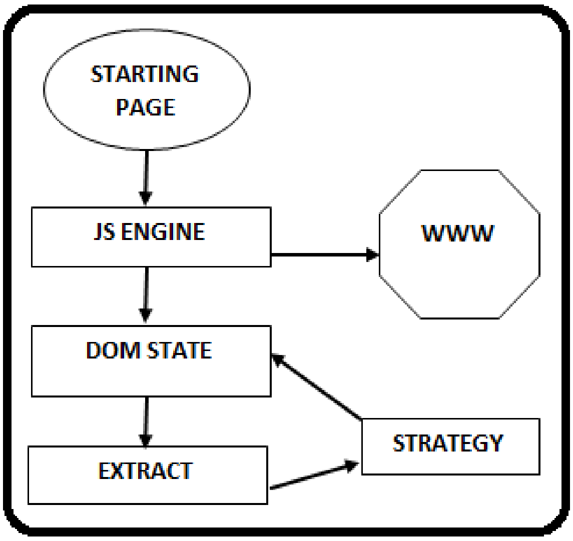

Figure 3 Architecture of RIA Web Crawler

3.1.4. Unified Model Web Crawler

The Unified Modeling Language (UML) is a general purpose graphic language which is usually used by software professionals for specifying, visualizing, constructing, and documenting the artifacts of a software intensive system. So, in Unified Model Web Crawler nod is calculated based on DOM and the URL. In this type of crawler redirecting the browser is a special client side event. For this UML uses three main models which are user model, object model and the dynamic model. The user model consists of use case diagram, object model represent by the class diagram and dynamic model is represent by the sequence diagram.

3.2. Search Interface

In this research paper we discussed two types of searching interface:

- Keyword based Search Interface.
- Multi-attribute Search Interface.

The information in the web database can be categorized in to categories which are structured database or unstructured database. The unstructured database contains plain text documents which are not well structured. It provides a simple keyword based search interface, where the user types the required keyword to fill text field. For example Fig 4 shows a typical keyword based search interface for people-search database.

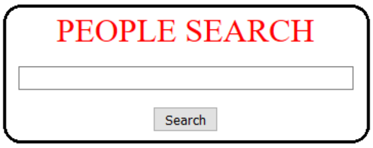

Figure4 Keyword based Search Interface

In contrast, structured database provide multi-attribute search interface which provides multiple query boxes depends upon the content. For example Fig 5 shows a multi-attribute search form interface for people-search database. Typically the given attributes are first name, last name, middle name, dob, email id, gender, location, pin code, marital status, contact number. Only first name, last name and gender are the mandatory fields; which means even though the user have not enough information about the person he is searching for, he will be able to see the search result. This is one of the advantages of multi-attribute search interface.

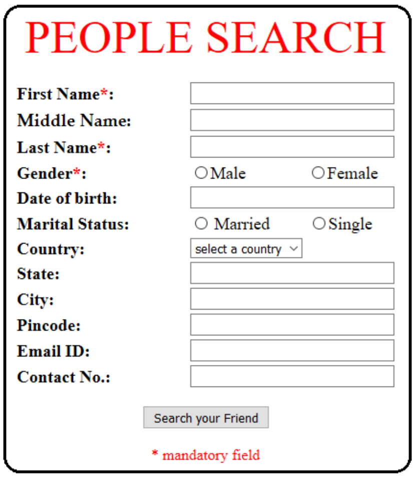

Figure 5 Multi-attribute Search Interface

3.3. Form Focused Crawler

This type of crawler combines the use of page classifier and link classifier. The page classifier concentrates on web pages and the link classifier works on the URLs. The frontier manager is used to select the next target links for crawling. Form Classifier is used to filter out useless forms. The new useful forms are added to the form database [2]. The architecture of Form Focused crawler is shown in fig. 6.

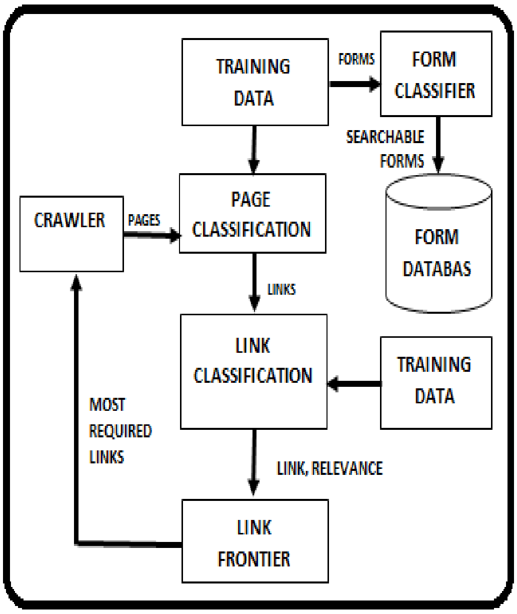

Figure6 Form focused Crawler

4. EXPERIMENTATION

Our challenge is to find the lost individual or relatives from our people-search database. The tools for this research include things like:

- Social Network
- Search Engines
- Public Records
- Online White Pages
- Birth, Marriage and Death indexes
- Refugee camp Records
- Non-Profit Organization Records

To determine the performance and accuracy of our technique, we test it on people domain. Our experiment will go through two important phases.

4.1. Collecting phase

This is the first phase of our experimentation. This phase consists of two sub phases:

4.1.1. Creating collection sites

In this sub phase we manually visited 6 websites at random which is shown in table I, and determined the attributes and aliases. The weight and threshold were also chosen manually, depending upon number of attributes.

Table I. Collection Websites

ATTRIBUTES	URL
Email search	https://www.peoplesmart.com/email
Phone No. search	http://www.spokeo.com/reverse-phone-lookup
Address search	http://www.yellowpages.com/whitepages/address
Name search	https://www.pipl.com/name
Pine code search	http://www.indiapost.gov.in/pincodesearch.aspx
Date of Birth search	https://www.dobsearch.com/

4.1.2. Creating Label table

After collection sites are being injected, all data of people are being determined and stored in labels table. The labels table contains label, alias and point which are shown in table II. The point depends upon the websites, if a website's search has more attributes in offering then the point will be much higher, and if the website's search option has less attributes then the point will be at the lower side.

Table II Lebel Table

LEBEL	ALIAS	MARK
First_name	'first_name of individual', name	70
Middle_name	'middle_name of individual'	75
Last_name	'last_name of individual'	70
Date of Birth	year, month, date	95
Location	country, state, city	75
Pin code		90
Gender	Male, female	95
Email		100
Contact no		95

4.2. Operating phase

In this phase we will determine the accuracy of our technique. First, we manually obtained a result by visiting the websites, second we determined the result which is obtained by the web crawler. Finally the results are being mapped by the help of matrix into five different stages.

- Set of associations between form field and domain attribute determined by web crawler = A
- Set of associations between form field and domain attribute determined manually = B
- Set of associations between forms and domains determined by web crawler= C
- Set of associations between forms and domains determined manually = D
- Set of submitted forms determined by web crawler = E

Here we used simple precision and recall method to determine the matrices. Five stages are as follows:

Stage 1: $[A\cap B] / [A]$ (PRECISION)

Stage 2: $[A \cap B] / [B]$ (RECALL)

Stage 3: [C ∩ D] / [C] (PRECISION)
Stage 4: [C ∩ D] / [D] (RECALL)
Stage 5: [E] / [C ∩ D] (PRECISION)

5. EXPERIMENTAL RESULTS

In this section we concentrate on the results of our experiment. We have considered six websites where the web crawler will crawl upon and extract the forms. The websites are shown in table III.

Table III Operating Websites

NAME	URL
Facebook	https://www.facebook.com
Google plus	https://www.plus.google.com
LinkedIn	https://www.linkedin.com
Pipl search	https://www.pipl.com
People search now	http://www.peoplesearchnow.com
Any who	http://www.anywho.com

We have gone through an ample number of experiments and tests to measure the accuracy and performance of our technique. The overall results are shown in table IV. Where X is denoted as collection dataset of collection websites and Y is denoted as operating dataset of operating websites. The global dataset is determined as (X+Y)*100.

Table IV Experimental Result

	X	Y	(X+Y)*100
FORM-DOMAIN ASSCIATION			
Precision	6/6 1.00	6/6 1.00	12/12 100
Recall	6/6 1.00	5/5 1.00	11/11 100
FIELD-ATTRIBUTE ASSOCIATION			
Precision	19/22 0.86	38/45 0.84	57/67 85
Recall	19/22 0.86	37/46 0.80	56/68 82
SUBMITTED FORMS			
Precision	6/6 1.00	6/6 1.00	12/12 100

6. CONCLUSION

This research paper represents a keyword based search form Interface and a multi-attribute search form interface for people database. Both forms are being created using simple HTML. SQL Injection technique is used to create the background database. The search results are up to the mark.

We also portrayed four different categories of crawling technique including Form Focused Web Crawling and given a fruitful source of information regarding each of the techniques. Further, the results of our experiment are quite convincing and some of them even reached 100 especially recall in associating forms and domains.

From the future point of view we can suggest that the crawling technique can be improved by adding more attributes to enhance the accuracy of the search. Handling forms by using JavaScript can be significantly improved in performance of hidden web crawling in the further future. Our proposed technique not only is applicable to people domain but it also can be applicable for different domains.

REFERENCES

[1] Michael Bergman, The Deep Web: Surfacing Hidden Value, in the journal of electronic Publication (JEP) volume 7, Issue 1, August, 2001.

[2] L. Barbosa, J Fieire, Searching for Hidden Web Database, in proceedings of WebDB page 1-6 2005

[3] S. Bal Gupta, Challenges in Designing a Hidden Web Crawler, in International Journal of Information Technology and System, 2(1), Jan-Jun 2013, pp 2277-9825

[4] IDC worldwide predictions 2014: Battles for dominance – and survival on the 3rd platform. http://www.idc.com/research/Predictions14/index.jsp, 2014.

[5] Seyed M. Mirtaheri, Mustafa EmreDinc¸t¨urk, Salman Hooshmand, Gregor V. Bochmann, Guy-Vincent Jourdan, Brief History of Web Crawlers, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario, Canada.

[6] Jaytrilok Choudhary and Devshri Roy. Priority Based Focused Web Crawler. *International Journal of Computer Engineering and Technology*, **4**(4), 2013, pp. 163-169.

[7] Houda El Bouhissi, Mimoun Malki and Djamila Berramdane. Applying Semantic Web Services. *International Journal of Computer Engineering and Technology*, **4**(2), 2013, pp. 108 - 113.

[8] A. Suganthy, G.S.Sumithra, J.Hindusha, A.Gayathri and S.Girija. Semantic Web Services and its Challenges. *International Journal of Computer Engineering and Technology,* **1**(2), 2010, pp. 26-37.

[9] J. Madhavan, D. Ko L. Kot, Google’s Deep Web Crawl, Proceedings of 1st International Conference on Very Large Data Bases(VLDB), August, Auckland, New Zealand, (2008),pp. 1241-1252

[10] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, accessing the deep web, Commun. ACM, 50(5), pp. 94–101, May 2007.

[11] Beena Mahar, C K Jha, A comparative study on web crawling for searching hidden web” in (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3)

[12] A. Mesbah, A. Van Deursen and S. Lenselink, Crawling Ajax based web applications through dynamic analysis of user interface state changes in ACM Transaction on the web- TWEB, volume 6, Issue 1,page 3, 2012.

[13] A report published by NetCraft on Web server survey January, 2015 http://www.netcraft.com/