Extraction of Query Interfaces for Domain-Specific Hidden Web Crawler

Nupur Gupta

Research Scholar in RGEC ,Meerut

Summary

Web databases are now permeative. Such a database can be retrieved via its query interface (only HTML query forms).Extracting HTML query forms is a major task in Deep Web. This task can be accomplished by two methods:

a) Positioned HTML forms on the web.

b) Recognizing domain-specific forms.

For positioning query forms (HTML forms) use HTML tags on the PIW (Publicly Indexable Web).Recognizing of query forms is essential because many of the forms are not the query forms. Non-query forms are used for access of data and data collection. This paper presents a novel approach for extracting web query interfaces using the query condition rules. Query conditions rules form by group label and form element in a query form. I have implemented the proposed novel approach in this paper.

Keywords:

Hidden Web database, query form extraction, domain-specific search.

1 .Introduction

Search engine like Google, Yahoo cannot search anything on the World Wide Web. These index only 10% of the entire web. And remaining 90% is called invisible web or Deep web. There are many websites through which user can access the desired information by wandering all the websites but it is very time-consuming job. So, it is required to build a consolidated query form through which users can get the desired information on the deep web, and also release the user from learn off the many websites. This task includes three main steps.

(i) Query form extraction
(ii) Matching and Mapping
(iii) Result integration.

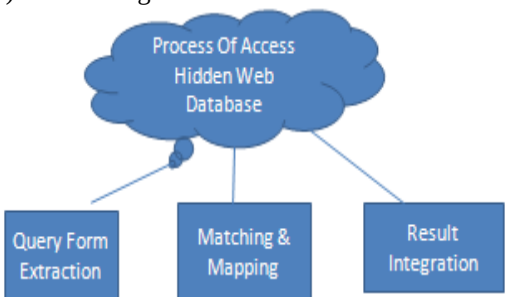

Fig 1: process of accessing hidden web database

In this paper, I explore the first task of accessing hidden web database. In this paper, I deduce the schema of each query form from query condition rules.

1.1 Deep Web

This is not the part of surface web and it cannot be searched by search engine. Infact, search engine search a very small portion of web and hidden web is much vast than surface web. This deep content is approximately 500 times vast than surface web. And this hidden information is a powerful research resource of all types of discipline area.

1.2 Searching Query Forms

There are billions of pages on the Surface Web and pages are freely added, deleted, or modified; and forms are sparsely distributed on the Surface Web [1], and Positioning of query forms is a challenging task. and recognizing of query forms is mandatory because some of forms are used for access of data and collection of data (e.g., login form, subscription forms, etc.).these are Non-Query Forms .These forms are detected and filtered by the crawling process. The techniques used for searching query forms are discussed later.

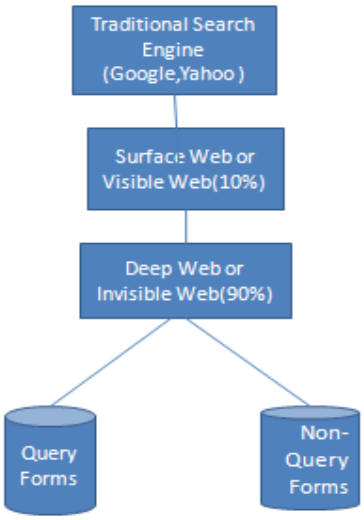

Fig 1.2.1: Location of Query Forms

Manuscript received February 5, 2016
Manuscript revised February 20, 2016

Fig 1.2.1 represents Traditional search engine (Google, Yahoo) search the data or information from the surface layer (10%of the entire web) and remaining 90% of the web is deep web or hidden web database which is extracted by the query forms. And Non-query forms are filtered out.

2. Literature Review

2.1 Crawler

Crawler is a small program that wanders around the internet and retrieves pages from links and hyperlinks. It is also known as web spider, worms or crawlers and is basically used for yielding up-to date data. It creates a replica of all traversed pages for later processing by a search engine that will list the downloaded pages to provide fast searches. A web crawler continues with a record of URL to visit, called the seeds or set of the URL. Crawler traverses these URLs, and recognizes all the hyperlinks in the page and maintains the record of all the traversed URLs. It is called crawl frontier. The main problem of crawler is that it needs a large amount of memory space and a huge network for maintaining a fraction of entire internet. A special kind of crawler (Focused Crawler) is required for eliminating such type of problem.

2.2 Focused Crawler

A focused crawler has as an objective to hunt for relevant pages for a predefined set of topics, Focused crawler avoids irrelevant links (unfocused pages) of the web and find the links that are related to the predefined set of boundary.
Advantages of focused crawler is remarkable saving in hardware and network resources and it also removes the problem of maintaining a large amount of memory space and huge network and helps keep the crawl more up-to-date

2.3 Form-Focused Crawler

Form focused crawler deals with the locating the forms on the PIW. Forms are distributed on the PIW, form focused crawler is used to locate the domain-specific forms on the PIW.
Form –focused crawler fabricated by the two techniques:

(i) Form Crawler (locate the forms on the PIW).
(ii) Focused Crawler(search from pre-defined set of boundary on the PIW).

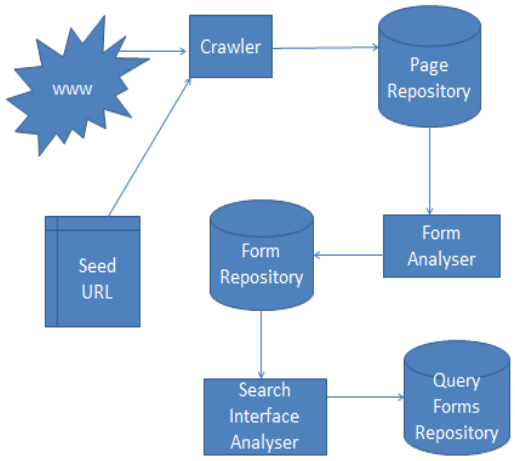

Fig: 2.1.1 A Form-Focused Crawler

The architecture of Form-Focused Crawler is depicted:
There are three fields used here:

(i) Page Repository
(ii) Form Analyzer
(iii) Form Repository
(iv) Search Interface Analyzer
(v) Query Form Repository

Crawler retrieves pages from the www (World Wide Web) and collects all the pages in a page repository database. Crawler also retrieves pages from some existing URL or seed URLs. These URLs are domain-specific. The Form-Analyzer classify the forms from all the pages then collects it in a Form Repository database, from form repository database all the query forms or search interfaces are analyzed by the Search Interface Analyzer and collects it in Query Forms Repository and useless forms (Non-Query Forms) are filtered out.

2.4 Hidden Web Crawler

Hidden Web is important for organizations having huge amount of information and maintain their data online, by using Web query search interface to their databases. Hidden Web crawler that can do independent search and download pages from the Hidden Web. Since the only way for retrieving the information from the hidden web site by the query interface. This crawler enables the techniques such as indexing, analysis and extracting of hidden web content. The mined data can be used to categorize and classify the hidden databases.

3. Query Condition Rules

A query interface has a set of query conditions that are semantically equivalent to each individual query (for a

particular domain). A query interface is not designed for a particular query form.

Query condition has a set of labels and form elements. There are four types of form elements (i) radio button, (ii) selection list, (iii) text box, (iv) check box in which user can either select from a preset values or enter any value.

A query interface has either a single attribute field or a multi-attribute field. A hidden web database categorize in two parts: (i) textual Database (ii) Structured Database

Textual Database: It contains a single field where user can enter a list of keywords and get the desired information.

Structured Database: It contains multi fields from where user can retrieve information for a particular domain.

For e.g.: Amazon.com website (Book) have both single attribute query interface and multi-attribute query interface, where user submit the query and search for desired result.

Fig: 3.1 A Single Attribute Query Interface

In fig: 3.1 A single textbox is given where user can write a list of keywords and search via this interface and all the related pages are retrieved. But there is a drawback of single attribute query interface that user can also get the some irrelevant pages also. The solution of this drawback is multi-attribute query interface.

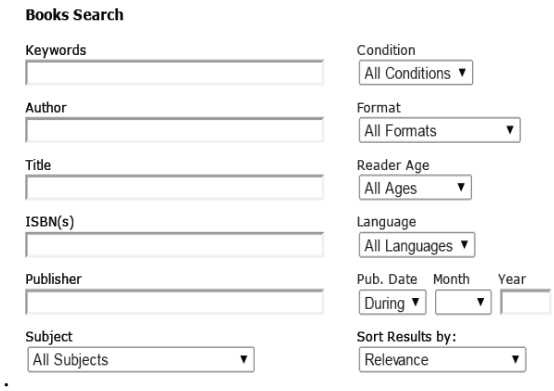

Fig: 3.2 A multi-attribute Query interface

In Fig : 3.2 A multi attribute field have more than one combination of Labels (Author, Title .etc..) and Form Elements(Textbox, Selection List etc….).Let user want to access the information about the particular book, user can enter author name , title and other details are shown in this figure. and user gets the desired information. The advantage of multi attribute query interface is that user get the specific information regarding their query .irrelevant pages are removed.

3.1 **Internal Representation of Query Forms**

Internal representation of form F includes the following information:

F= ({L1, L2, L3….Ln} ;{ E1, E2, E3….En}; S),

Where {L1, L2, L3…..Ln} is a set of n Labels, {E1, E2, E3, En} is a set of n form elements, S is Submit or Go button or from where information is processed by performing an action.

Query condition rules are used for searching the query forms or search interface on the hidden web database. It is defined by the <input type = ‘submit’, value = “Go”,”search”,”Find”>

4. Conclusion

This paper state the problem of query interface extraction. This paper proposed a novel approach of extracting query forms, consider the HTML forms (query forms) and extract the query forms by the HTML tag used for forms. Now, consider the query condition rules to search the relevant forms (query forms) and discard the Non-query forms. This approach has much improved quality of query form extraction.

References

[1] Mauricio C. Moraes, Carlos A. Heuser, Viviane P. Moreira, Denilson Barbosa, "Prequery Discovery of Domain-Specific Query Forms: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 8, pp. 1830-1848, Aug. 2013, doi:10.1109/TKDE.2012.111

[2] Nirpendra Narayan Das, Ela Kumar, “Hidden Web Query Technique for Extracting the Data Deep Web Data Base,” WCECS 2012 Vol I, October 24-26, 2012, San Francisco, USA.

[3] Jun Hong, Zhongtian He, David A.Bell,” Extracting Web Query Interfaces Based on Form Structures and Semantic Similarity”, IEEE DOI 10.1109/ICDE.2009.215.

[4] Bao-Hua Qiang, Jian-qing Xi, Ling Chen,”Effective Schema Extraction of Query Interface on the Deep Web”, IEEE DOI 10.1109/FSKD.2008.135.

[5] Ying Wang, Tao Peng,Wanli Zuo,Huifeng Zhu,”Schema Extraction of Deep Query Interface”,IEEE DOI 10.1109/WISM.2009.86.

[6] Yang Daowen, Liu Quan, Cui Zhiming, Fu Yuchen,”The Discovery and Extraction of Query Interface Based On Deep Web”, IEEE DOI 10.1109/WCSE.2009.129

[7] Wensheng Wu, AnHaiDoan, Clement Yu, and Weiyi Meng,”Modeling and Extracting Deep-Web Query Interfaces”SCI 251, pp.65-90 2009

[8] Luciano Barbosa, Juliana Freire,”Searching for Hidden-Web Databases” (WebDB 2005) June 16-17, 2005.

Nupur Gupta received the B.Tech degree, from Shobhit Institute of Engineering & Technology, Meerut in 2010. Affiliated from Uttar Pradesh Technical University, Lucknow and pursuing M.Tech degree, from Radha Govind Engineering College, Meerut, affiliated from Uttar Pradesh Technical University, Lucknow.