International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

# Automatic Retrieval System for Domain Specific Hidden Web Database

**Megha Saini, Dr. Mukesh Rawat**

Dr. A. P. J. Abdul Kalam Technical Unversity, Lucknow, Uttar Pradesh, India

**Abstract:** ***As the size of WWW is growing daily, lots of relevant information are available in the form of hypertext in WWW. Users can view publically indexible pages of any web portal and some information viewed after filling of the appropriate search interfaces. Vast information is available behind the search interfaces known as hidden data. Different search engines are using different techniques for fetching such deep information from the WWW and sent to the users according to their queries. The purpose of fetching such deep information is to provide a large set of relevant result to the user according to their search query. In this paper a new and innovative methodology suggested for automatic filling of search forms and then submit the form to the WWW and collect the response pages which will be send back after automatic submission of the form and then filter out relevant pages from these response pages according to the user query.***

**Keywords:** Search engine, crawlers, hidden deep web

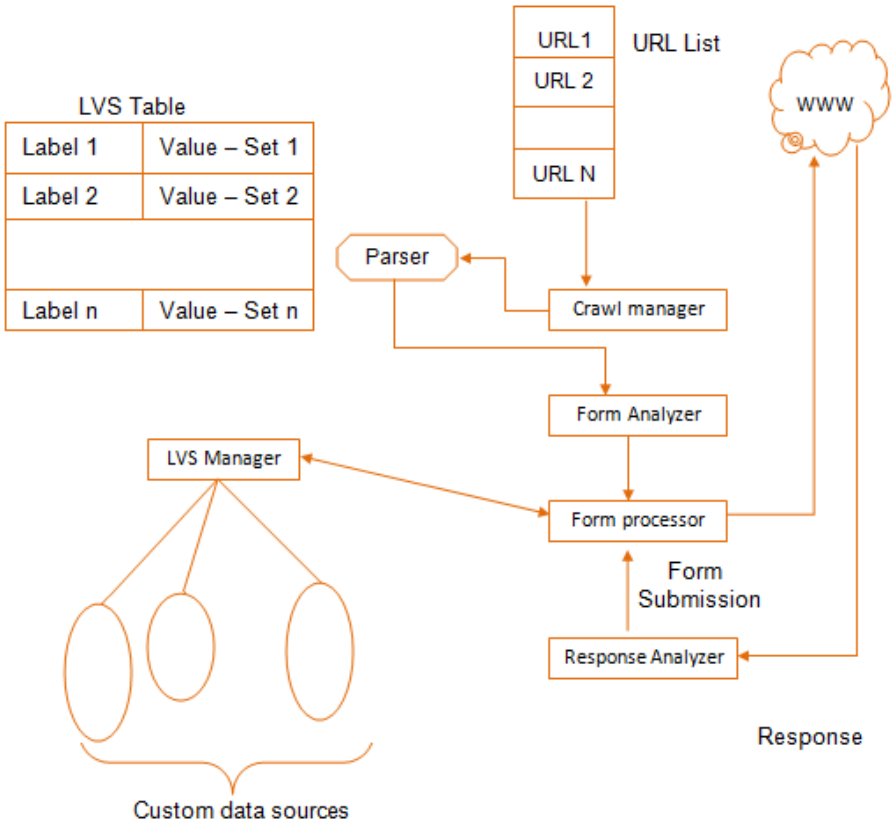
## 1. Introduction

Search Engines – A search engine is actually general class of program despite the term is generally used to exactly define system like Google, Bing and Yahoo that empower user to search for documents on the World Wide Web. Search engines are programs that search documents for detailed keywords and acknowledge the list of the documents where the keywords were found.

Crawlers – Traditionally, crawlers have only focused a section of the web called the publically indexible web (PIW). A crawler is a program that visits website and reads their page and other information in order to create enteries for a search engine index. PIW is the set of pages reachable exactly by following hypertext links neglecting search forms and pages that need certification.

Hidden deep web – The deep web are the parts of the world wide web whose content are not indexed by the standard search engines the contrary term for deep web is the surface web. Deep web is also known as the deep net or invisible web or the hidden web, are the parts of internet that are not considered part of the surface web.

## 2. HIWE Architecture



HIWE's Main Modules

- URL list: It Contains all URLs that crawler has discovered so far
- Parser: It abstract hypertext links from the crawler pages and adds them to the URL list
- Crawl Manager: It Commands the entire crawling process
- Form Analyzer, from processor, response Analyzer: Together implement the form processing and submission operations
- LVS table: HIWE's implementation of the task specific database
- LVS manager: Manages addition and access to the LVS table

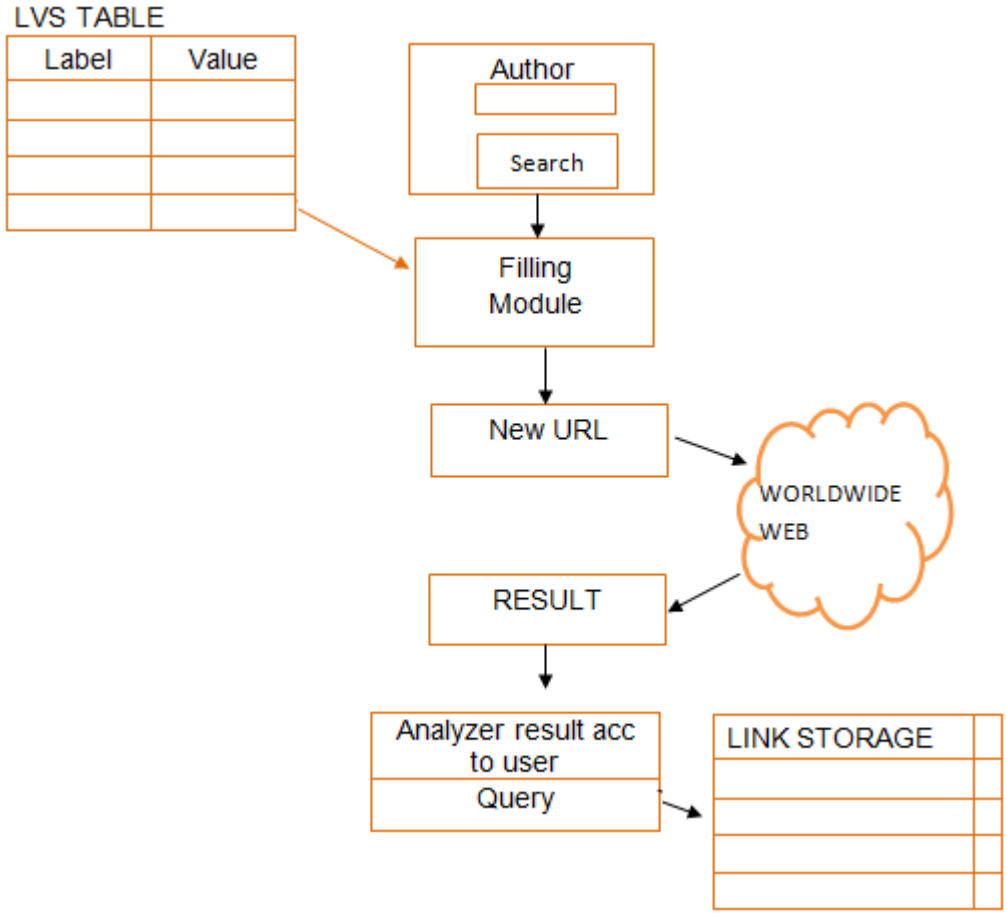
## 3. Methodology For Extracting Hidden Data

### 3.1 Form Representation

Crawler builds are internal form representation form F={E1,E2-----En}, Sn, M} where {E1-----En} set of n form element, M – meta information about the form for each element E1, S- submission information combined with the form, HIWE assembles a domain DOM (E1) and a label (E1). The label of the form element is the comparative information related with that element, most forms are related with some comparative text to help the user to understand of the element. The domain is the set of values which can be combined with the corresponding form element

### 3.2 Filling of forms and automatic submission

For every domain, the system tries to match its attributes with the fields of the form. By using the result of the first output of the previous step, the system determines that the form is relevant with respect to the query or not. If the form is relevant the crawler uses it to execute the queries defined in the domain. For each query, the module is filled using the LVS table and the new URL is generated which is sent to the World Wide Web. The World Wide Web generates the result of the query. The result of the www is analyzed according to the user query and matched with the keywords of the search query entered by the users. If it is relevant, the URL of the resultant page is directly stored in the link storage.




## 4. Conclusion

This is a new and efficient approach for automatic filling and submission of search interfaces to the WWW and get back the response pages as a result. In this paper, we addressed the problem of crawling and extracting content from the hidden web. We presented a simple model of a search form and the process of filling out of forms. In this technique the content of the response pages is matched with the keywords of the search query entered by the users without storing the page and if the page content is meaningful according to the user query then the URL of the resultant page is stored.