Selection of Relevant Fields of Searchable Form Based on Meta Keywords Frequencies

Anjali Chaudhary
Student,CS
Shobhit University,
Meerut-25011

Seema Kashyap
Assistant Professor, CS
Shobhit University,
Meerut-250110

Vinit Kumar
Student,CS
Shobhit University
Meerut-250110

Abstract- Crawling the hidden web is a very challenging problem due to the number of active web pages on the web is increasing exponentially. Hidden web contents are highly relevant to every information need for searching purpose. Hidden web contents are usually behind the searchable form, ignore by all traditional crawler so hidden web crawler used for this purpose. Because of hidden web contains a vast amount of data so it is necessary to crawl only specific hidden data. In this paper, a meta keywords frequencies based architecture has been proposed to carry this work.

Keywords: web, hidden web, form, Meta keywords.

I. INTRODUCTION

The Web is usually categorized in two parts, surface web and hidden web. Surface web that is usually crawled by traditional search engines and make searchable and easily accessible. The web resources that are not accessible by these traditional search engines called hidden web. So hidden web crawler is used to crawl the hidden web. According to the surveyed that the hidden web contains thousands terabytes of hidden information [2, 6]. Due to the enormous size of hidden web, it is very difficult to manage this large volume of hidden data. Usually crawlers ignore large amount of high quality data behind the searchable forms. So it is a major problem to extract only relevant information from the searchable forms. Hidden web crawlers are used for this purpose that enables indexing, analysis and mining of hidden web content [4]. This paper presents a Meta keywords frequencies based technique that select only relevant field of searchable form to extract relevant hidden web content.

II. Related work

The World Wide Web (WWW) is a huge repository of interlinked hypertext documents accessed via Internet. The World Wide Web consists of all public websites connected to Internet, Contain in text, videos, images and other multimedia information. Web is categorized in two parts, surface web and hidden web. Surface web consist resources that are usually crawled by traditional search engines. The web resources that are not accessible by these traditional search engines called hidden web or (invisible web).

Hidden web consist vast amount of relevant data that crawled using hidden web crawlers. This is a very challenging problem to crawl the hidden web behind the searchable forms. A task specific hidden web crawler or (HiWE) [4] address this problem by automatically filling out form, using label’s domain value pair. Since search forms are the entry-points into the hidden Web, HiWE is designed to automatically process, analyze and submit forms, using a set of (element, domain) pairs of forms and form submissions. A form element can be any one of the standard input objects such as selection lists, text boxes or radio buttons.

The basic actions of HiWE (fetching pages, parsing and extracting URLs and adding the URLs to a URL list) are similar to traditional web crawlers. However, HiWE also performs the following sequence of actions for each form on a page are, Form Analysis to build an internal representation based on the above model, Value assignment and ranking use approximate string matching between the form labels and the labels in the LVS table to generate a set of candidate value assignments.

It uses fuzzy aggregation functions to combine individual weights into weights for value assignments and use these weights for ranking the candidate assignments, Form Submission use the top “N" value assignments to repeatedly fill out and submit the form. Response Analysis and Navigation Analyze the response pages to check if the submission yielded valid search results. Use this feedback to tune the value assignments. Now it crawls the hypertext links in the response page to some pre-specified depth. A general architecture of a Task Specific Hidden Web Crawler or Hidden Web Exposer(shown in Figure1).

There are many other approaches that focus on the data extraction. Another technique [8] automatically generates agents to collect hidden web pages by filling forms.

Another method [10] extracts valuable information behind the HTML forms but do not include crawler. Model [12] for form filling process successfully extracts hidden web contents. It gives a detailed account of the architecture of the deep web crawler and describes various approaches for building domain and list of values pairs. Problem behind this strategy lies is that it is not automated and scalable.

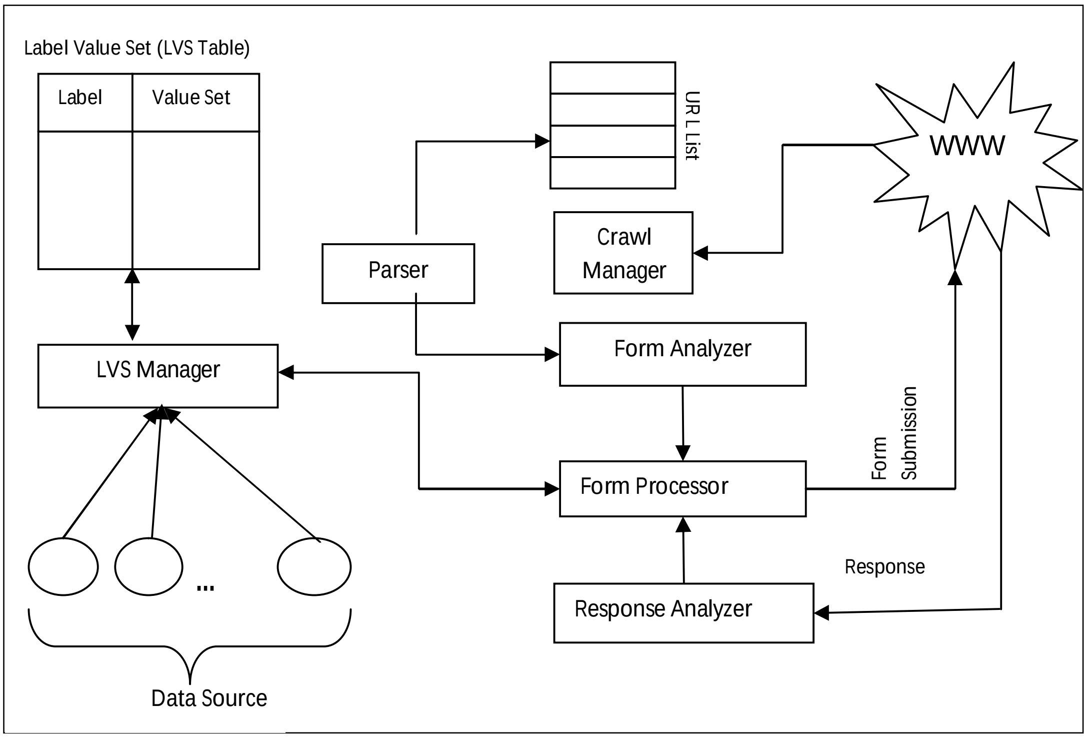

Figure 1: A general architecture of a task specific hidden web crawler or (HiWE)

III Proposed Work

The proposed technique selects relevant field of searchable forms to extract relevant data behind the searchable form. In this proposed work Meta keywords frequencies are used to select only relevant fields of searchable form to extract relevant hidden web content.

Meta tags have been chosen because search engines give more preference to these tags for searching purpose and to gather the information about a website. The content of Meta tags generally describes information about the HTML page which usually cannot be represented by any other HTML tags.

Architecture of Selection of Relevant Fields of searchable form based on Meta Keywords Frequencies (shown in Figure 2).

This proposed architecture have many functional components are Form extractor, Meta keywords extractor, Frequency calculator and Field matcher. Functioning of main components is as follows:-

- **Form extractor:** First, it checks for form available on web page, if web page contains form, it extracts all searchable forms from web page and sends these forms to field matcher.
- **Field matcher:** It uses a high frequency selector formula to select meta keywords with high frequencies. All selected meta keywords matches with the form information. Then selects the relevant field of form that uses to extracts the form's information and stores relevant form's information into database. This information further utilized for searching hidden web.
- **Meta keywords extractor:** Extracts keywords of all Meta tags available on a web page containing forms. Extracted keywords are the combination of complete string separated by comma (,) and these keywords are further separated using blank space.
- **Thesaurus:** Contains the list of words according to similarity of meaning (containing synonyms and antonyms), which provides definitions for words. This helps the user to find the words having the similar meaning.
- **Frequency calculator:** It computes relative frequency of each extracted Meta keyword with all other contents of the web page. The resultant of this module is a Meta table that consist Meta keywords and their corresponding frequency.

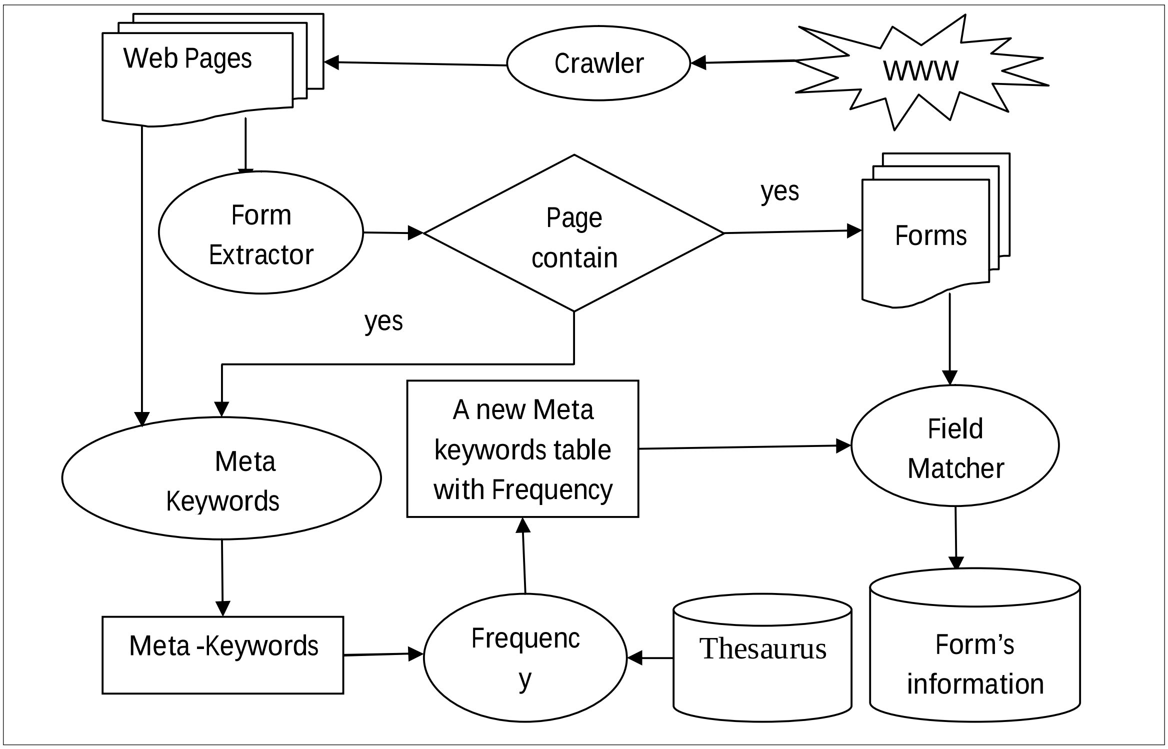

Figure 2: Proposed architecture of field selection based on Meta keywords frequencies

High Frequency Selector Formula (HFS) is used to select the Meta keywords with high frequencies.

$$HSF = \sum_{i}^{n} \frac{sum\ of\ all\ meta\ keyword's frequencies}{total\ number\ of\ meta\ keywords}$$

Or

$$HSF = \sum_{i}^{n} \frac{mf1 + mf2 + mf3 + mf4 + \cdots + mfn}{total\ number\ of\ meta\ keywords}$$

By applying this function we get the average frequency, which is compared with the each Meta keyword frequencies.Meta keywords with the same or higher frequency of average frequency are selected for matching with the form fields with the Field Matcher.

The algorithm used to extract hidden web contents is as follows:

```
Algorithm RelevantFieldInfo()
{
        1. Download web page from the World Wide Web.
        2. Form extractor checks for form the downloaded web page contains.
        3. If yes go to step 4 else go to step 9.
        4. Form extractor extract forms available on the web page and send all the forms to   the Field matcher.
        5. Meta keywords extractor creates a Meta table by extracting all the Meta keyword found on the web page.
        6. Frequency calculator calculates the frequency of all Meta keywords listed in      Meta table.
        7. Field matcher applies High Frequency Selector Formula (HFS) to select Meta keywords with high frequencies. Then matches form contents with selected Meta keywords and extract relative information.
        8. Add extracted information to the database for searching relevant hidden web content.
        9. Go to Step1.
}
```

```
Algorithm FrequencyCalculator()
{
        1.  Extract all keywords (lexicon) from the downloaded web pages.
        2.  Matches all keywords with meta keywords extracted by meta keyword extractor to calculate frequencies of
            meta keywords.
              For each keyword 1 to n
                  For each meta keyword 1 to n
                   If(keyword = meta keword or meta keyword synonym= keyword)
                    $FreqMetaKey_i = FreqMetaKey_i + 1;$
}
```

IV. Implementation and Results

Meta tags content on downloaded web page form http://www.naukri.com are:
<meta name =”robots" content="INDEX, FOLLOW">
<meta name="Description" content="Find the Best Jobs in Naukri.com , India’s No. 1 Job Site. Search for Job Vacancies across Top Companies in India. Post your Resume now to find your Dream Job!">
<meta name="Keywords" content="naukri.com, naukri, jobs, job, career openings, jobs in india, job site in india, it jobs india, software jobs india, it jobs in india, jobs india, india jobs, job search in india, online jobs in india, accounting jobs in india, part time jobs in india, banking jobs in india, finance jobs in india, jobs and careers in india, call center jobs in india, marketing jobs in india">
<meta name="application-name" content="Naukri.com – Indias No.1 job site ">

The extracted form from the web page is http://www.naukri.com (shown in Figure 3).

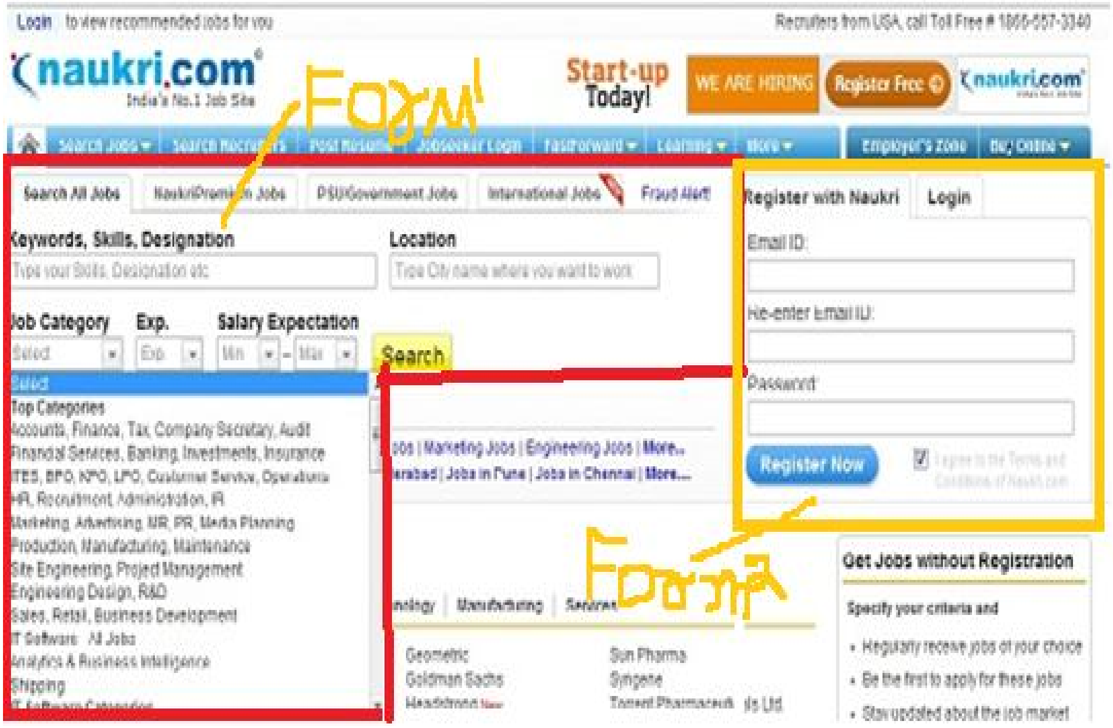
Login to view recommended jobs for you
Recruiters from USA, call Toll Free # 1866-557-3340
naukri.com
India's No.1 Job Site
Form1
Start-up Today!
WE ARE HIRING
Register Free
Search Jobs
Search Recruiters
Post Resume
Jobseeker Login
Fastforward
Learning
More
Employer's Zone
Buy Online
Search All Jobs
NaukriPremium Jobs
PSU/Government Jobs
International Jobs
Fraud Alert
Keywords, Skills, Designation
Type your Skills, Designation etc
Location
Type City name where you want to work
Job Category
Exp.
Salary Expectation
Select
Exp
Min
Max
Search
Select
Top Categories
Accounts, Finance, Tax, Company Secretary, Audit
Financial Services, Banking, Investments, Insurance
ITES, BPO, KPO, LPO, Customer Service, Operations
HR, Recruitment, Administration, IR
Marketing, Advertising, MR, PR, Media Planning
Production, Manufacturing, Maintenance
Site Engineering, Project Management
Engineering Design, R&D
Sales, Retail, Business Development
IT Software - All Jobs
Analytics & Business Intelligence
Shipping
IT Software Categories
obs | Marketing Jobs | Engineering Jobs | More...
erabad | Jobs in Pune | Jobs in Chennai | More...
nology
Manufacturing
Services
Form2
Geometric
Goldman Sachs
Headstrong New
Sun Pharma
Syngene
Torrent Pharmaceuticals Ltd
Register with Naukri
Login
Email ID:
Re-enter Email ID:
Password:
Register Now
☑ I agree to the Terms and Conditions of Naukri.com
Get Jobs without Registration
Specify your criteria and
• Regularly receive jobs of your choice
• Be the first to apply for these jobs
• Stay updated about the job market

Figure 3: The Web Page with Forms

Meta keywords table with frequencies generated by frequency calculator (shown in Table I)

TABLE I: Meta table with frequencies

Meta keywords	**Frequency**
Naukri	108
Jobs	126
Job	103
career openings	8
Career	34
Openings	23
jobs in india	14
India	55
job site in india	6
it jobs india	23
india jobs	16
job search in india	3
Search	9
online jobs in india	11
Online	17
accounting jobs in india	8
Accounting	11
part time jobs in india	14
Part	15
Time	23
banking jobs in india	17
Banking	23
finance jobs in india	3
Finance	16
jobs and careers in india	5
marketing jobs in india	3
Marketing	21

After apply high frequency formula selected meta keywords are shown in Table II .

$$HFS = (108 + 126 + 103 + 8 + 34 + 23 + 14 + 55 + 6 + 23 + 16 + 3 + 16 + 3 + 9 + 11 + 8 + 11 + 14 + 15 + 23 + 17 + 23 + 17 + 23 + 3 + 16 + 5 + 3 + 21)/27$$

$$HFS = 26.482$$

Table II: Selected Meta keyword with Frequency

Meta keywords	**Frequency**
Naukri	108
Jobs	126
Job	103
Career	34
India	55

The below Snapshot shows the selected field of forms after applying the High Frequency Formula(HSF) by field matcher in the web page.

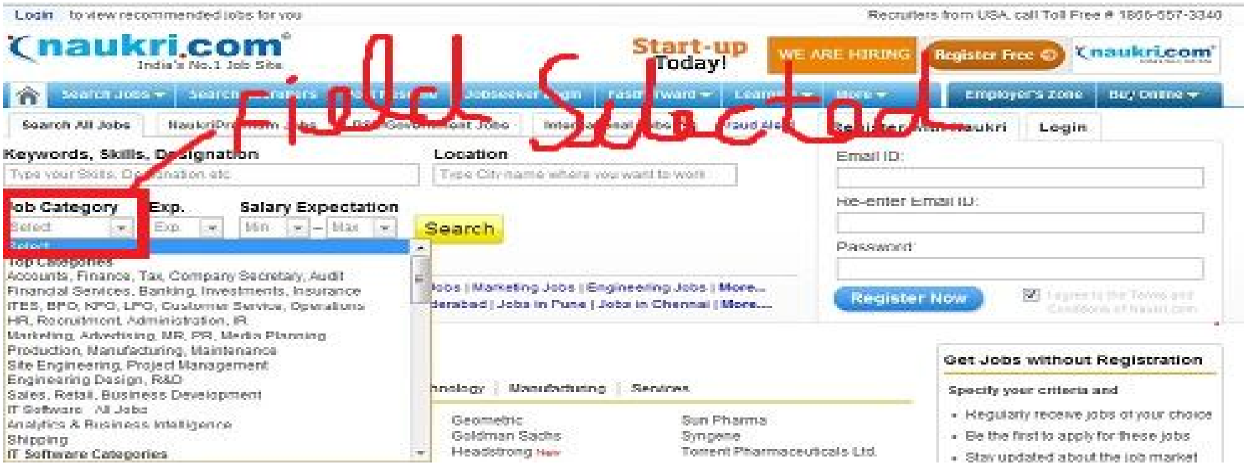
Field Selected
Login to view recommended jobs for you
Recruiters from USA call Toll Free # 1866-557-3340
naukri.com
India's No.1 Job Site
Start-up Today!
WE ARE HIRING
Register Free
Search Jobs
Jobseeker Login
Employer's Zone
Buy Online
Search All Jobs
Register with Naukri
Login
Keywords, Skills, Designation
Type your Skills, Designation etc
Location
Type City name where you want to work
Job Category
Select
Exp.
Salary Expectation
Min
Max
Search
Top Categories
Accounts, Finance, Tax, Company Secretary, Audit
Financial Services, Banking, Investments, Insurance
ITES, BPO, KPO, LPO, Customer Service, Operations
HR, Recruitment, Administration, IR
Marketing, Advertising, MR, PR, Media Planning
Production, Manufacturing, Maintenance
Site Engineering, Project Management
Engineering Design, R&D
Sales, Retail, Business Development
IT Software - All Jobs
Analytics & Business Intelligence
Shipping
IT Software Categories
Email ID:
Re-enter Email ID:
Password:
Register Now
I agree to the Terms and Conditions of Naukri.com
Get Jobs without Registration
Specify your criteria and
Regularly receive jobs of your choice
Be the first to apply for these jobs
Stay updated about the job market
Geometric
Goldman Sachs
Headstrong
Sun Pharma
Syngene
Torrent Pharmaceuticals Ltd.

Figure 4: Selected Field of Forms

V. CONCLUSION

Hidden web contains large amount of sensitive information. Traditional crawlers usually ignore this large amount of sensitive data behind the searchable forms. The proposed technique is proposed to extract only relevant information from hidden web content. This technique uses Meta keywords frequencies to select appropriate field of form to extract relevant hidden web content (Meta keyword with higher frequency are chosen for matching with the form fields). As we know almost eighty percent part of the web comes under the Hidden Web, so it is very difficult to extract the whole part of the Hidden web but we can extract the most relevant data (about twenty percent) of Hidden Web by this proposed technique. Thus this technique is proposed to improve searching capability to find relevant data from the hidden web.

REFERENCES

[1] S. Chakrabarti, K. Punera and M. Subramanyam, "Accelerated focused crawling through online relevance feedback", *Proceedings of 11th World Wide Web conference*, Hawaii, pp. 148–159, 2002.

[2] Vinit Kumar, Niraj Singhal and Ashutosh Dixit, "A Utility Function based Model for Extraction of Most Relevant Hidden Web Contents", *Proceedings of 4th National Conference on Recent Trends in Advanced Computing*, Electronics and Information Technology (CICON-2014**),** Shobhit University, Meerut, p. 48, March 2, 2014.

[3] S. Chakrabarti, M. Van Den Berg and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery", *Proceedings of 8th World Wide Web conference*, Toronto, pp. 1623–1640, 1999.

[4] S. Raghavan, Hector Garcia-Molina, "Crawling the Hidden Web", *Proceedings of 27th International Conference on Very Large Databases*, Rome, Italy, pp. 11-14, 2001.

[5] M. K. Bergman, "The Deep Web: Surfacing Hidden Value", *Proceedings of 66th IFLA Council and General Conference*, pp. 154-157, 2000.

[6] P. Lyman and H. R. Varian, "How Much Information ?", *Technical report, UC Berkeley*, 2003. (www.sims. berkeley.edu/research/projects/how-muchinfo-2003/internet.htm).

[7] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles and M. Gori, "Focused Crawling Using Context Graphs", *Proceedings of International Conference on Very Large Databases*, pp. 527– 534, 2000.

[8] Lage, "Collecting Hidden Web Pages for Data Extraction", *Proceedings of the 4th International Workshop on Web Information and Data Management*, pp. 69-75, 2002.

[9] Jufeng Yang, Guangshun Shi, Yan Zheng and Qingren Wang, "Data Extraction from Deep Web Pages", *Proceedings of International Conference on Computational Intelligence and Security*, pp. 237-241, 2007.

[10] S. W. Liddle, D. W. Embley, D. T. Scott and S. H. Yau, "Extracting Data behind Web Forms", *Proceedings of 28th VLDB Conference LNCS*, Vol. 2784, HongKong, China, pp. 402-413, 2002.

[11] Sonali Gupta and Komal Kumar Bhatia, "HiCrawl: A Hidden Web crawler for Medical Domain", *Proceedings of IEEE International Symposium on Computing and Business Intelligence*, Delhi, India, pp. 152-157, 2013.

[12] Brian E. Brewington and George Cybenko, "How dynamic is the web", *Procseedings of the 9th International World Wide Web Conference on Computer Networks*, pp. 257-276, June 2000.

[13] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. "Measuring index quality using random walks on the web". *Proceedings of the 8th Int. World Wide Web Conference (WWW8)*, pp 213–225, 1999.