ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

A Survey on Multiword Smart Crawler for Effective Web Searching

Gangwal Pallavi Kailas[1], Prof.D.B.Kshirsagar[2]

PG Student, SRES'COE, Kopargaon, SPPU, Maharashtra, India[1]

HOD, Department of Computer, SRES'COE, Kopargaon, SPPU, Maharashtra, India[2]

ABSTRACT: In today's world, the number of web pages available in the Internet is growing tremendously. Therefore, searching of relevant information from the Internet is hard task. Also ,there is problem of hidden information behind deep web. A lot of this information is hidden behind query forms that we are considered as hidden or deep web. Retrieving this hidden information by using Traditional search engines is really challenging task. Therefore, we propose a two-stage framework, namely Multiword Smart Crawler, for effectively and efficiently harvesting web pages. This System will try to achieve wide coverage for deep web interfaces and maintains highly efficient crawling. It is a focused crawler consisting of two stages: efficient site locating and balanced in-site exploring which implements to achieve higher harvest rates than other crawlers. Also it considers multi-word similarity of sites and pages. By ranking collected sites and by focusing the crawling by extracting words and matching similarity with query terms, it will try to achieve more accurate results

KEYWORDS: Deep web, HTML form, reverse searching, prioritizing, adaptive learning, ranking, term-similarity, multi-word crawling.

I. INTRODUCTION

Most of the data available on internet is in deep web form. It is very difficult work to locate deep web pages, because they are not recorded by any traditional search engines. These pages are usually rarely distributed and keep constantly changing. To solve this problem, previous work has proposed two types of crawlers which are generic crawlers and focused crawlers. Generic crawler fetches all the searchable forms and do not focus on a specific topic whereas Focused crawlers[3][5][10] are the crawler which focuses on a specific topic. Form-focused crawler (FFC) and Adaptive crawler for hidden web entries (ACHE) aims to efficiently and automatically detect other forms in the same domain. The accuracy of these focused crawlers is low in terms of retrieving relevant forms and time to access the relevant information.This problem is addressed by using Multiword Smart Crawler. This crawler is divided into two stages: site locating and in-site exploring. The site locating stage helps achieve wide coverage of sites for a focused crawler, and the in-site exploring stage can efficiently perform searches for web forms within a site. Site locating technique employs a reverse searching technique (e.g. using Google's "link:" facility to get pages pointing to a given link) and incremental two-level site prioritizing technique for prioritizing relevant sites,& achieving more data sources. During the in-site exploring stage, link tree is used for balanced link prioritizing, eliminating bias toward webpages in popular directories. To achieve more accurate result, system introduces the advantages of multi-word crawler, which gives more accurate and relevant pages by checking its word similarity. It results in retrieving more number of pages efficiently with higher harvest rate within less time than other crawlers.

II. LITERATURE SURVEY

1) Combining Classifiers to Identify Online Databases:
In 2007, Luciano Barbosa et.al. has addressed the problem of identification of the domain of online databases. Here, new strategy is developed that automatically and accurately classifies online databases based on the features of Web forms. By partitioning the space of form features, the simple classifier can be constructed using learning techniques that are better suited for the features of each partition. The use of different classifiers leads to high accuracy, precision and recall.[7]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

2) Search Interface On The Web: Querying & Characterizing.
In 2008, Denis Shestakov has studied the problem of discovering hidden portion Web behind web search interfaces. Three classes of problems around the deep Web has considered: characterization of hidden web, finding and classifying deep web resources, and querying web databases.
User-friendly and expressive form of query language is used which allows to retrieve information behind search interfaces and extract useful data from the result pages based on specified conditions.[9]

3) Web Crawling Foundation & Trends in Information Retrieval:
In 2010, Christopher Olston et.al. introduced the steps in crawling of deep web-
-Locating sources of web content.
-Selection of relevant sources.
-Extracting the underlying content of deep web pages.
Here is the problem of retrieving unwanted pages which needs more time to crawl relevant results. [6]

4) Crawling Deep Web Entity Pages:
In 2013, Yeye Hey et.al. has a built a system that specializes in crawling entity-oriented deep web sites. It handles important problems of deep web crawling such as query generation, empty page filtering and URL duplication in specific context of entity oriented deep web sites.[10]

5) Comparative Study of Hidden Web Crawlers:
In 2014, Sonali Gupta et.al. give review on working of the various Hidden Web crawlers. They mentioned the strengths and weaknesses of the techniques implemented in each crawlers. Crawlers are differentiated on the basis of their underlying techniques and behaviour towards different kind of search forms and domains. This study will useful in research perspective [3].

- **STRENGTHS AND WEAKNESS OF SUPPORTIVE REFERENCE PAPERS:**

Sr. No.	Author & Title	Publication & Year	Strengths	Weakness
1	S.Raghavan et.al. "Crawling the Hidden web"[5]	Proceedings of the 27th VLDB Conference, Roma, Italy, 2001	1) Effective form processing & label matching process. 2) LITE technique is introduced. 3) Automatic extraction of semantic information from search forms & response pages.	1) Task specific approach requires good quality input from human. 2) Unable to recognize & respond to simple dependencies between form fields. 3)Lack of support for partial information in form fields.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

2	P.G.Ipeirotis et.al. "Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection"[8]	Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002	1) Gives the facility of meta search tools. 2) Content-summarization technique is used.	1) Query does not consider flat Classification i.e. it requires hierarchical classification only.
3	Dr. Khanna et.al. "Concept of Search Engine Optimization in Web Search Engine"[4]	International Journal of Advanced Engineering Research and Studies , 2011	1) SEO technique is used to find & rank websites in response to user queries over millions of pages.	1) Requirement of choosing the right keyword for optimization.
4	Feng Zhao, et.al."SmartCrawler: A Two Stage Crawler for Efficiently Harvesting Deep-Web Interfaces",[1]	IEEE, 2015	1) Use of hierarchical tree generation to avoid bias nature of crawler. 2) Reverse searching & site prioritizing mechanisms are used. 3) Gives efficient results & harvest more number of forms.	1) Crawling large amount of data may lead more time consumption. 2) Need to classify deep web forms to improve accuracy of form classifier.
5	Nandhini.G et.al. "Multi-word Crawler Harvesting For DWI"[2]	International Journal of Advanced Research in Biology Engineering Science and Technology, March 2016	1) Word similarity is used. 2) Efficient clustering technique is used with English text. 3) It supports multi-keyword ranked search & synonym based search. 4) Bookmark concept is included which can be visible globally.	1) Need to provide prior input data for clustering and text matching.

III. PROPOSED SYSTEM OVERVIEW

In this project, following steps are performed-

1. Seed site collection:-
2. Reverse Searching:-
3. Multiword Extraction:-
4. Site Ranking
5. Link Ranking

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

IV. SYSTEM ARCHITECTURE

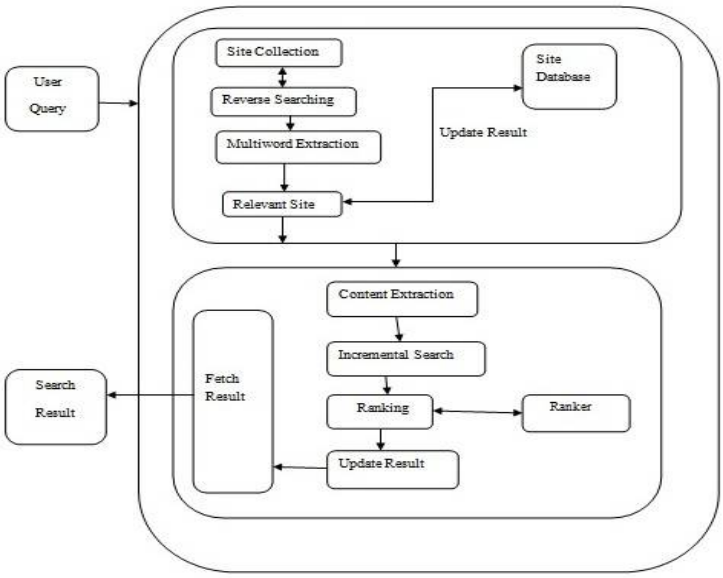

Fig No 01 System Architecture

Problem Solving Strategy:

1) Two Stage Crawling-

It is difficult to search the deep web databases, because they are not registered with any search engines, are usually sparsely distributed, and keep constantly changing. To address this problem, we use generic crawlers and focused crawlers. Generic crawlers fetch all searchable forms and cannot focus on a specific topic. The link classifiers in these crawlers play a pivotal role in achieving higher crawling efficiency than the best-first crawler However, these link classifiers are used to predict the distance to the page containing searchable forms, which is difficult to estimate, especially for the delayed benefit links (links eventually lead to pages with forms). As a result, the crawler can be inefficiently led to pages without targeted forms.

2) Ranking

When combined with above stop-early policy. We solve this problem by prioritizing highly relevant links with link ranking. However, link ranking may introduce bias for highly relevant links in certain directories. Our solution is to build a link tree for a balanced link prioritizing In this example, servlet directory is for dynamic request; For links that only differ in the query string part, we consider them as the same URL. Because links are often distributed unevenly in server directories, prioritizing links by the relevance can potentially bias toward some directories.

3) Adaptive learning

In the site locating stage, high relevant sites are prioritized and the crawling is focused on a topic using the contents of the root page of sites, achieving more accurate results. During the in site exploring stage, relevant links are prioritized for fast in-site searching.. The results also show the effectiveness of the reverse searching and adaptive learning.

V. IMPLEMENTATION PLAN

4. Seed site collection:-
   Collecting seed sites from internet resources for user requested query.
5. Reverse Searching:-
   We randomly select a known deep website or a seed site and use general search engine's facility to find center pages and other relevant sites,
6. Multiword Extraction:-

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

The system searches the user input data and also the word has similar meaning and provide the relevant data for search result. The input key term is preprocessed with term similarity generation using word net tool. The input terms are classified with the help of database. The input data sets finally go to the web server after processed. The web server provides the result.

4. Site Ranking

Site Ranker assigns a score for each unvisited site that corresponds to its relevance to the already discovered deep web sites and also consider the site's meta data.

6. Link Ranking

This component search text data from the body part of the HTML page. We internally parse the data to a parse tree using document parser. From this parse tree we extract all the paragraph tags that are in the body of the HTML page. We find root word from all the texts. This is because we consider that the paragraph tags in the body contain text that is unique.

VI. PERFORMANCE EVALUATION

The evaluation can be done based on following factors:
i) Performance matrices such as Precision, Recall.
ii) Impact of timing to find relevant documents.
iii) Harvesting more number of documents.

VII. CONCLUSION

In this project, System will try to achieve wide coverage for deep web interfaces and maintains highly efficient crawling. It is a focused crawler consisting of two stages: efficient site locating and balanced in-site exploring which implements to achieve higher harvest rates than other crawlers. Also it considers multi-word similarity of sites and pages. By ranking collected sites and by focusing the crawling by extracting words and matching similarity with query terms, it will try to achieve more accurate results. In future work, System can combine pre-query and post-query approaches for classifying deep-web forms to further improve the accuracy of the form classifier and green crawling concept for giving more quick result by applying search on log files.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my thanks to my guide Prof.D.B.Kshirsagar for his esteemed guidance and encouragement. His guidance always encourages me to succeed in this work. I am also very grateful for his guidance and comments while designing part of my research paper and learnt many things under his leadership.

REFERENCES

[1] Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin "SmartCrawler: A Two Stage Crawler for Efficiently Harvesting Deep-Web Interfaces",IEEE, VOL. 99, 2015.
[2] Nandhini.G, Murali.D, Kumaresan.A "Multi-Word Crawler Harvesting For DWI" International Journal of Advanced Research in Biology Engineering Science and echnology (IJARBEST) Vol. 2, March 2016.
[3] Sonali Gupta, Komal Kumar Bhatia "A Comparative Study of Hidden Web Crawlers", International Journal of Computer Trends and Technology (IJCTT) Vol. 12, Jun 2014.
[4] Dr. Khanna SamratVivekanand Omprakash , "Concept of Search Engine Optimization in Web Search Engine", International Journal of Advanced Engineering Research and Studies(IJAERS) Vol. I, October-December, 2011.
[5] Sriram Raghavan, Hector Garcia-Molina, "Crawling the Hidden Web", Proceedings of the 27th VLDB Conference, Italy, 2001.
[6] Olston and M. Najork, "Web Crawling, Foundations and Trends in Information Retrieval", vol. 4, No. 3, pp. 175–246, 2010.
[7] Luciano Barbosa, Juliana Freire, "Combining Classifiers to Identify Online Databases", International World Wide Web Conference Committee (IW3C2), May 2007.
[8] Panagiotis G. Ipeirotis, Luis Gravano , "Distributed Search over the Hidden Web:
Hierarchical Database Sampling and Selection", Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002.
[9] Denis Shestakov, Tapio Salakoski, "Search Interfaces on the Web: Querying and Characterizing", TUCS Dissertations, May 2008.
[10] Yeye He, Dong Xin, Venkatesh Ganti, Sriram Rajaraman, and Nirav Shah, "Crawling Deep Web Entity Pages", In Proceedings of the sixth ACM international conference on Web search and data mining, ACM, 2013.