Indian Journal of Science and Technology, *Vol 10(17), DOI: 10.17485/ijst/2017/v10i17/110304, May 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Survey on Uniform Resource Locator and Content Matching to Discover Deep- Web Pages

Sayali Shelke* and Parth Sagar

Department of Computer Engineering, RMDSSOE, Pune – 411502, Maharashtra, India; Sayalishelke52@gmail.com, Parthsagar.rmdssoe@sinhgad.edu

Abstract

Objectives: 1) The Objective is to harvest the deep web pages efficiently[1] 2) Personalize search according to user interest. 3) Combine pre-query and post-query approach. **Methods/Statistical Analysis**: Three methods are used 1. URL Matching: This method is used to match the query content in URL. For that, the system gets a link from online and site database. Links are extracted to match the user entered query content. 2. Content matching: This method is used extracting the links and getting form content and matching the user-entered query .If match calculates the occurrence frequency of that query on the form. For that, it use Jsoup library.3. Pre-query Algorithm: This method used to display pre-query result after entering focus in the search box. For that user login to the system that time system select his profile and according to that links will display to the user. **Findings:** As in the existing system, most of the search engines display the results according to the most visited sites or recently added sites. To find the deep web pages from the databases is a challenging, because they are not enrolled with any web indexes and keep constantly changing. In this system, Smart Crawler performs URL matching and content matching to discover the deep web pages. Proposed crawler proficiently gets deep-web network from wide destination and accomplishes the higher outcome from different crawlers. Page ranking is performed and it displays high ranked results on the result page .Here it provided personalized search, results display according to the user professions. Maintaining log file and the pre-query result will reduce time. First time this crawler perform personalize search this means that this crawler is unique. During the evaluation, notice that proposed approach is more efficient than the existing crawler. **Application/Improvements:** The application gathers real-time user profile information from user accounts. Therefore, it must be reliable and keep those data in safe. This crawler is used as the search engine for e-learning application, E-shopping site. Links can be a bookmark for future use. As in improvement, it can rank the pages according to user-entered review for each link. Also, the opened link will display page content in task extraction form. That is code, concept, URL on the page.

Keyword: Deep Web, IP, Positioning, Smart Crawler

1. Introduction

A Web Crawler is a framework for expand downloading of pages from the web. This crawlers are valuable as scope of goals. Web crawlers are the main components of web search engines, it assembles huge of web pages, and create entries for a search engine index, and enable clients to issue questions against the list and discover the web pages which is like the queries.[1]On web, deep web is expanding there has been expanded enthusiasm for strategies that assistance to find deep-web interfaces. Because of the enormous use of web assets and furthermore the changing way of the deep web, it is challenging to fetch relevant data from the search engines as per user's requirement. To solve this problem "Smart Crawler". Smart-crawler gets seed from seed database. To avoid going by an expansive number of pages, the first step is, it performs Link based searching for the middle pages with the assistance of any web search tools. In the second stage it used to match frame content, then it grouping pertinent and insignificant site. Here the designer customized scan for proficient outcomes and keeping up the log for productive time

**Author for correspondence*

administration. Techniques used in various approach as seen in Table 1.

Table 1. Techniques used in various approach

No.	Name	Techniques
1	Comparative Study of Hidden Web Crawlers	1) Text similarity to match fields and domain attributes. 2) Partial page layout and visual adjacency for identifying form elements 3) Hash of visually important parts of the page to detect errors
2	Web Crawling	1) Batch crawling. 2) Incremental crawling.
3	An active crawler for discovering geospatial Web services and their distribution pattern - A case study of OGC Web Map Service	1)Proposing an accumulated term frequency–used for prioritized crawling, 2) concurrent multi-threading technique 3) To update the metadata of identified services we adopting an automatic mechanism.
4	Focused crawler	Web resource discovery; Classification; Categorization; Topic distillation
5)	Supporting Privacy Protection in Personalized Web Search	1)GreedyDP 2)GreedyIL , for runtime generalization.
6)	Improving the Efficiency of Web Crawler by Integrating Pre-Query Approach	Dempster–Shafer theory
7	Preprocessing Techniques for Text Mining - An Overview	Natural Language Processing
8	Deep Web Entity Monitoring	1)MetaQuerier 2)IntegraWeb
9	Crawling the Hidden Web	1)Matching Function 2)Label Matching.
10	A Hierarchical Approach to Model Web Query Interfaces for Web Source Integration,	1) Query Interface Matching 2) Building Unified Query Interfaces 3) Deep Web Crawling

2. Research Method

[1] Comparative Study of Hidden Web Crawlers[2]-
Audit on working of the different Hidden Web crawlers. They said the qualities and shortcomings of the systems actualized in every crawler. Crawlers are separated on the premise of their fundamental techniques and furthermore the conduct towards different sorts of areas. This review is helpful in research perspective.[2]

[2] Web Crawling Foundation & Trends in Information Retrieval -Presented the means in crawling of profound website pages

- ✓ Locates wellsprings of web substance.
- ✓ Selection of significant sources.
- ✓ Extracting the hidden substance of profound website pages. Here is the issue of recovering undesirable pages which require more opportunity to creep important outcomes.

[3] Search Engines going beyond Keyword Search[3]-
This survey is basically based on to solve the problem of information, search engines also known as current information tool need to be improved. It has to arrange the search and filter the processes completely and give relevant information for this it should be embedded in the search tool. This paper shots to group the significant difficulties for watchword web search tools to plan for the fast development of web and bolster the client require in brisk time.[3]

[4] Personalization on E-Content Retrieval Based on Semantic Web Services[4]-
Learning object repositories have been increased drastically along with the current educational system. LOR is mainly available in large databases available over the internet. An LOR is storing content as well as also their metadata records. Different labeling standards such as LOM, Dublin Core. Over the large database available across the internet there is a need to develop a solution which will provide an efficient technique to search for distributed and heterogeneous situations.[4]

[5] Focused crawling: new approach to topic-specific web resource discovery[5]-
Two new techniques have been proposed to overcome the problem in existing crawling. The one is a classifier which searches the focus topic with the relevance of hypertext document. Another technique is a distiller which identifies hypertext nodes that access focused points within

few links with relevant pages. Focus crawling discover a largely overlapping set of resources in spite of disruption and it is robust. Proposed crawling searches for required pages steadily which overcome the problem in standard crawling which quickly disperses from its aim.[5]

[6] Improving the Efficiency of Web Crawler by Integrating Pre Query Approach[6]-
The measure of information devoured by crawler while looking is immense. The crawler looks a lot of information that may contain heaps of immaterial data. A great deal of time is squandered for seeking significant information among the colossal measure of immaterial outcomes returned by crawler and client needs to squander a period while creeping on the web while filtering insignificant connections too. Pre/Post inquiry preparing methodologies and site-based seeking methodology can be incorporated keeping in mind the end goal to prehandle the client question. By coordination of various preparing methodologies and connection, positioning methodologies a great deal of profitable client time is spared. Post question approach may likewise sift through all unessential data which is a bit much as indicated by the inquiry which is been terminated and gives the normal outcomes.[6]

[7] Pre-processing Techniques for Text Mining[7]-
Data mining is utilized for finding the valuable data from a lot of data. Data mining strategy used to execute and take care of various sorts of research issues. The exploration related regions in data mining are content mining, web mining, text mining, picture mining. This paper examined the content mining and its pre-handling systems. Content mining is the way toward mining the helpful data from the content archives. It is called knowledge discovery in the text (KDT). Text mining is a procedure which removes data from both organized and unstructured data and discovering designs. Text mining procedures are utilized as a part of different sorts of research areas like common dialect handling, data recovery, content arrangement and content bunching .[7]

[8] Deep Web Entity Monitoring[8]-
When the information has been searched on the web using the different search engine are available. It fetches, few data among all available information on the web. Apart from the search results fetched by search engines, there is still a huge measure of information accessible on the web which did not look via web indexes such information is characterized as concealed web or profound web which is not open through the web crawlers. Different interfaces have been developed to find this deep or hidden data. Querying for the deep or hidden data through the information sources might be helpful but could also be a troublesome, tedious and tiring undertaking. Surface web is the hidden data available over the internet which can't be accessed through search engines and this is much bigger than as compare to the data fetched by search engines. Considering the enormous measure of data accessible on the web in the form of hidden data, the user might interested in accessing all this data along with the data fetched by search engines. Therefore there is huge demand to develop a search engine which categories the users into different categories and will narrow the search domain and fetch as the large quantity of information as possible.[8]

[9] A Hierarchical Approach to Model Web Query Interfaces for Web Source Integration[9]-
Basically, this paper describes that the enough information on the web is hidden back of WQI. Web database coordination and Profound Web slithering which require a programmed utilization of these interfaces. So a fundamental issue to be tended to is the programmed extraction of inquiry interfaces into a proper model. The presence of an arrangement of area free "CDR" that aides the development of WQI. Every one of these standards change over inquiry interfaces into pattern trees. WQI extraction calculation, which joins the geometric format of these tokens and HTML tokens in a Page. Tokens are partitioned into various classes out of which the most essential ones are content tokens and field tokens. The tree structure is inferred for these tokens utilizing their geometric design. The progressive portrayal is accomplished by joining these two trees. Toward the end changed from the extraction issue into a coordination issue.[9]

3. Existing System

In the existing strategies, it creates a solitary profile for each client, when client's advantage differs for a similar inquiry but most of the time conflict occurs. Example, when a user wants to know about the banking exam and enter a query "Bank" or interest, may vary from banking account to accounts of cash bank however not under any condition intrigued in the list of the blood bank in all states. This time conflict may occur so with adverse choice for obtaining the fine grain between the interested outcomes and not interested outcomes.[1,10]

Consider the accompanying two conditions:

1) In view of document technique:
In this technique delaying at capturing user's clicking and browsing behaviours of the customer. It manages click through information from the customer it implies the records customer has clicked on. Click through information in search motors can be considered as triplets {q, r, c} Where,

q is for query
r is for ranking
c is for the arrangement of connections clicked

2) Based on concept methods:
In this techniques, the purpose is capturing user's conceptual needs. User's browsing all the record and research past. Customer profiles help to show customer significance and furthermore to infer their aims for searching recent queries.

DRAWBACKS -

1) Profound web interfaces.
2) It achieve large analysis and challenging issue is a high efficiency

3.1 Comparison

Comparison between existing and proposed crawler

1) Two-stage Smart crawler: Fundamentally in light of the estimation of the weight of the page with the thought of the active links, approaching links and title tag of the page at the time of seeking.

Limitation: It is constructed just with respect to the request of the different website page

Proposed Smart crawler: Proposed crawler is based graph based algorithm which is based on connection structure of pages. Consider the back connections in the rank computations. Rank is figured on the premise of the significance of pages.

2) Two-stage Smart crawler: Uses Incremental site prioritizing: that calculates out of site link of pages and according to that it classifies pages as relevant and displays it to the user. That takes more time to execute process.[11]

Proposed Smart crawler: Uses Incremental site prioritizing that directly check query word on page content. So it takes minimum time to execute or calculate result.

3) Generic crawlers: This crawlers are created for describing deep web and index development of deep web assets that don't restrict seek on a particular topic.[12]

Proposed Smart crawler: Focus on specific topic and perform domain classification for links. This uses naïve bays algorithm for domain classification

4) Form-Focused crawler: Form-focused crawling, which sift through non-searchable and insignificant structures. Doesn't allow users to perform personalized web search and doesn't produce pre-query result.[13]

Proposed smart crawler: Allow the user to personalized search according to user profession. Gives relevant result to user on searched query.

4. Conclusion

Because of the gigantic utilization of web assets and furthermore the changing way of the deep web, it is challenging to fetch relevant data from the search engines according to client's necessity and which expends additional time than normal. Smart crawler gives productive outcome than another crawler. Smart Crawler works in two aspects: URL matching as well as content matching. The ranking used to get pertinent outcomes that are the post query result. It personalizes the looking as indicated by client profile so that it is simple to find a precise outcome to the user.Maintaining log file will decrease time to list items. In pre-query results are displayed according to user personalized result after placing focus on the search box.So it will give the solution in the better way.Therefore, smart crawler gives efficient result than another crawler.

5. References

1. Zhao F, Zhou J, Nie C, Huang H, Jin H. Smart Crawler: A Two-Stage Crawler For Efficiently Harvesting Deep-Web Interfaces. IEEE transactions on services computing. 2016 Jul/Aug; 9(4):608–20. Crossref
2. Gupta S, Bhatia KK. A Comparative Study of Hidden Web Crawlers. International Journal of Computer Trends and Technology (IJCTT). 2014 Jun; 12(3):111–8. Crossref
3. Rahman M. Search Engines going beyond Keyword Search: A Survey. International Journal of Computer Applications by IJCA Journal. 2013 Aug; 75(17):1–8. Crossref
4. Gil1 AB, Rodrguez S, de la Prieta F, De Paz JF. Personalization on E-Content Retrieval based on Semantic Web Services. International Journal of Computer Information Systems and Industrial Management Applications. 2012; 5. ISSN 2150-7988.
5. Chakrabarti S, Den Berg MV, and Dom B. Focused crawling: A new approach to topic-specific web resource discovery. Comput Netw. 1999; 31(11):1623–40. Crossref

6. Shukla V. Improving the Efficiency of Web Crawler by Integrating Pre-Query Approach. International Journal of Innovative Research in Computer and Communication Engineering. 4(1):172–5.
7. Vijayarani S, Ilamathi J, Nithya. Preprocessing Techniques for Text Mining - An Overview. International Journal of Computer Science and Communication Networks. 5(1):7–16.
8. Kabisch T, Dragut EC, Yu C, and Leser U. Deep web integration with visqi. Proc VLDB Endowment. 2010; 3(1/2):1613–6. Crossref
9. Dragut EC, Kabisch T, Yu C, Leser U. A hierarchical approach to model web query interfaces for web source integration. Proc VLDB Endowment [Online]. 2(1):325–36. Crossref
10. Li W, Yang C, Yang C. An active crawler for discovering geospatial Web services and their Distribution pattern - A case study of OGC Web Map Service. International Journal of Geographical Information Science. 24(8):1127–47. Crossref
11. Sheng C, Zhang N, Tao Y, Jin X. Optimal algorithms for crawling a hidden database in the web. Proc VLDB Endowment. 2012; 5(11):1112–23. Crossref
12. Chakrabarti S, Den Berg MV, and Dom B. Focused crawling: A new approach to topic-specific web resource discovery. Comput Netw. 1999; 31(11):1623–40. Crossref
13. Wu W, Yu C, Doan A, Meng W. An interactive clustering based approach to integrating source query interfaces on the deep Webroot. ACM SIGMOD Int Conf Manage Data. 2004. p. 95–106.