Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 9; April-June, 2015 pp. 48-51

http://www.krishisanskriti.org/acsit.html

Design of A Novel Ranking Technique for Hidden Web Pages

Jyoti Yadav[1] and Manvi[2]

[1]M.tech Student Dept of Computer Science Engineering YMCA University of Science & Technology Faridabad, India
[2]Dept of Computer Science Engineering YMCA University of Science & Technology Faridabad, India
E-mail: [1]jyadav200@gmail.com

Abstract—*Web is a wide term which mainly consists of surface web and hidden web. The hidden or deep web refers to content that is hidden behind HTML forms. This contains a large collection of data which is unreachable by link-based search engines. A study conducted at University of California, Berkeley estimated that the deep web consists of around 91,000 terabytes of data, whereas the surface web is only about 167 terabytes. Several researchers have explored various methods for crawling deep web content. To access this content one must submit valid input values to the HTML forms. These hidden web crawlers return huge result set for the user query. But users commonly look at top ten or twenty results that can be seen without scrolling. Users rarely look at results coming after first response page so ranking of the results is needed. Till now ranking of hidden web pages is a big challenge, enough work has not been done in this area. In this paper, a novel technique for the ranking of hidden web pages is designed.*

Keywords*: Hidden Web , Deep Web Ranking Algorithms Ranking Techniques.*

1. INTRODUCTION

The World Wide Web (WWW) consists of two types of web pages: surface web (or visible web) and deep web (or the hidden web or the invisible web).The Surface Web[3] refers to the part of the Web that can be crawled and indexed by general purpose search engines and the hidden Web[1] refers to the abundant information that is "hidden" behind the query interfaces and not directly accessible to the traditional search engines. Examples of hidden web resources includes online banking websites, shopping websites , online book data stores etc. As the size of the web is increasing day-by-day, similarly the size of hidden web is also increasing[3][4]. A typical link based search engine cannot access the deep web pages because the content is hidden behind these web forms. The huge amount of valuable information stored on the hidden web in back end database of websites is accessible only after the user enters a query through a search interface. So specific hidden web crawlers[6] are needed to extract information from websites having hidden web data. In general Query Engine may return several hundreds or thousands of URL that match the keywords for a given query. But often users look at top ten results that can be seen without scrolling. Users seldom look at results coming after first search result page, which means that results which are not among top ten are nearly invisible for general user. Therefore to provide better search result, page ranking mechanisms are used by most search engines for putting the important pages on top leaving the less important pages in the bottom of result list. There are various page ranking algorithms for surface web. Some of the common page ranking algorithms of surface web are PageRank Algorithm [7], Weighted PageRank Algorithm [8] and Hyperlinked Induced Topic search Algorithm [9]. But there is a scarcity of ranking algorithms for hidden web pages.

The aim of the paper is to design architecture for a novel efficient ranking technique for Hidden Web data. Section 2 is Literature Review in which existing techniques for ranking of hidden web data are analyzed, section 3 explains the proposed architecture for ranking technique of hidden web pages. Section 4 is conclusion and future work.

2. LITERATURE REVIEW

Many researchers are trying to develop novel ideas to improve ranking technique for hidden web pages in order to provide better experience for users. A brief overview at few of them is given in the following section:

Balakrishnan et al[13] considered the emerging problem of ranking the deep web data considering trustworthiness and relevance. In this paper end-to-end deep web ranking by focusing on: (i) ranking and selection of the deep web databases (ii) topic sensitive ranking of the sources (iii) ranking the result tuples from the selected databases has been discussed.

Neha Batra et al[10] proposed ranking algorithm consists of four different attributes. These are: PageRank , Term Weighting Technique [TWT] ,User's Feedback ,Visitor. In this paper user feedback and no. of user visiting the websites is considered important for ranking of web page. Term weight is calculated on basis of document length and term frequency. The limitation of this paper is that it is ranking of the pages on basis of PageRank Algorithm which checks link structure of

page in account to calculate rank of hidden web page whereas hidden web pages lacks link structure.

1. Brian Wong in et al[12] gave ranking algorithm which utilizes best-fit scoring functions using ten quality factors and a dynamic weighting algorithm that changes the factor weighting based on user behavior. This algorithm is scalable and requires minimal pre-processing to generate the factor weightings.
2. Saravanan, Nan Zhang and Gautam Das et al[14] introduced problem of rank discovery over hidden web databases. This paper define a comprehensive spectrum of ranking functions according to various dimensions such as query-dependent vs. static, observable vs. proprietary, and whether the scoring attribute can be queried or not. This paper discuss the feasibility of rank discovery for each type of ranking function, and show that different types of ranking functions require fundamentally different approaches for rank discovery. For proprietary and observable ranking functions, they developed RANK-EST(algorithm) which interleaves two separate procedures for handling high and low ranked tuples, respectively. This paper also present theoretical analysis of ranking of hidden web.

3. PROPOSED ARCHITECTURE

To present the results to the user in an ordered manner, Page Ranking methods are applied, which can arrange the results in order of their relevance, importance and content score. These ranking methods can be query dependent or query independent method. Query-dependent are all ranking methods that are specific to a given query, while query-independent factors are attached to the results, regardless of a given query. Query-dependent factors used by search engines are measures such as word documents frequency, the position of the query terms within the result page or the inverted document frequency, which are all measures that are used in traditional Information Retrieval System. Some of the query independent factors are Link popularity, Page-content popularity and upto-dateness etc.

The design and algorithms for a Ranking technique is proposed which uses factors for both query dependent and query independent ranking methods. Factors of query dependent such as page frequency , query – page content matching are used and factors of query independent such as page content popularity , page source rank, user feedback are also used to design a novel and efficient ranking technique.

Major components of proposed Architecture are as:-

I. Weight assignment
II. Page Source Rank Determiner
III. Frequency Calculation
IV. Final Rank Assignment.

Architecture of proposed Ranking technique is shown below in fig.1.

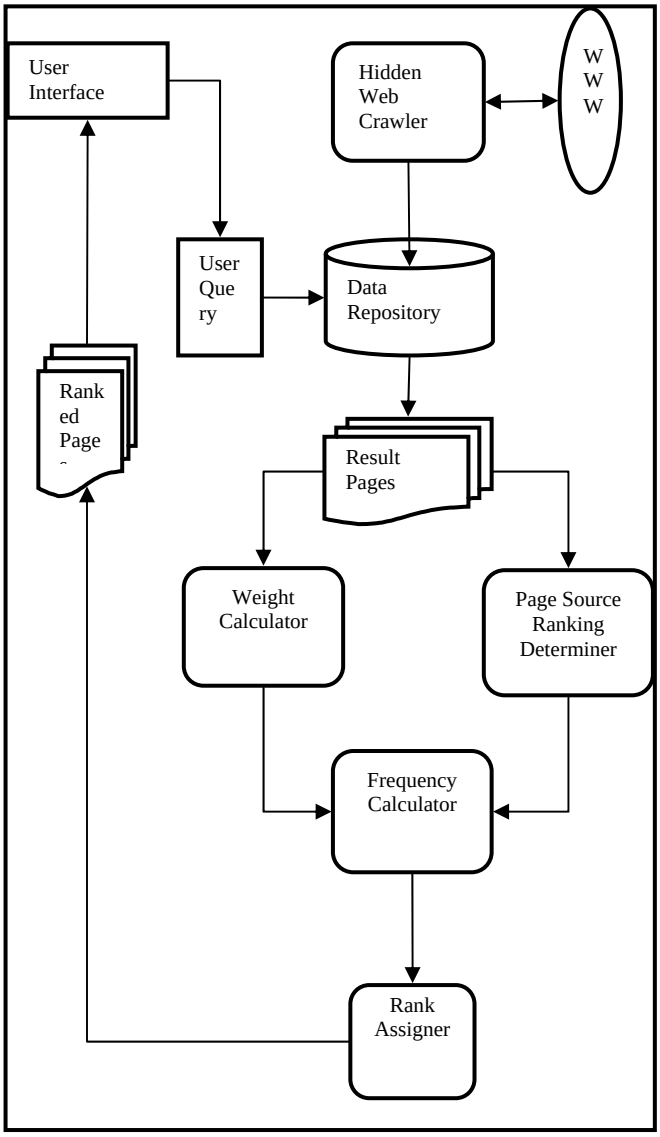

Fig. 1: Architecture of Ranking Technique

Each module of the architecture of ranking technique for hidden web pages in detail along with the algorithms used to implement them is explained below.

3.1 .Main Modules Of Architecture:-

3.1.1 .Weight Assignment Module

In this module weight , W is assigned to each url/web page on basis of three factors. One is rating on the web page(w1), second Users Feedback present on the web page (w2) on basis of previous visits and third is Query-Page Content Matching(w3), as shown in fig. 2 below. This Weight , w will be used as one of factors to calculate rank of the web page.

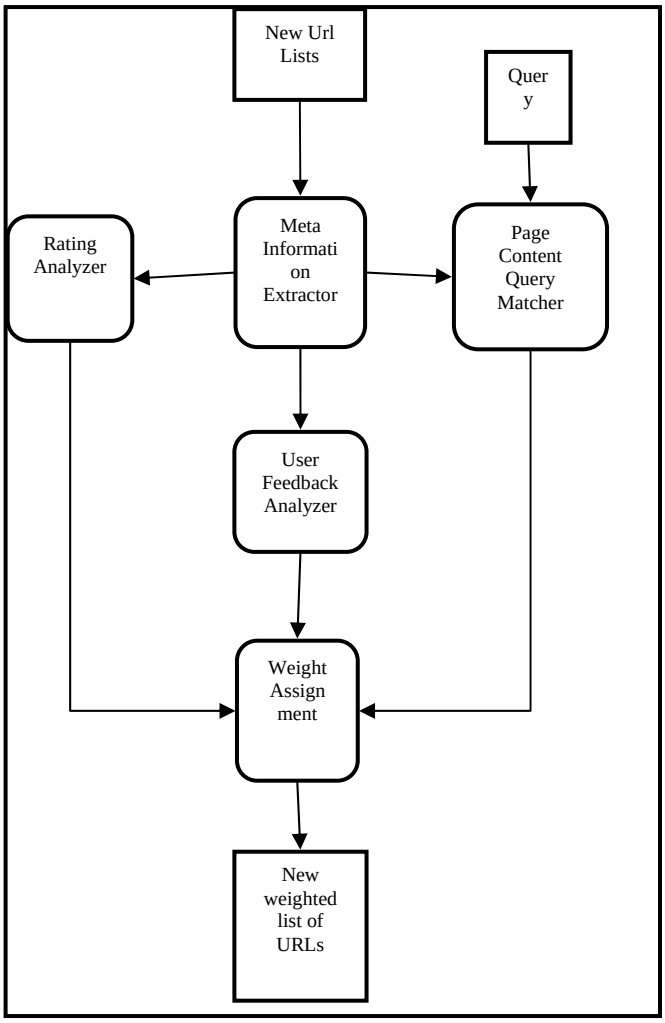

Fig. 2: Weight Assignment Module

Sub-Modules of Weight Assignment Module:-

a) **Rating Analyzer:** Rating on any web page shows the popularity of the website owner or service provider on web. Such as. For a book domain any rating on book on any web page are basically bookseller ratings which are based on seller's completion rate. Completion rate represents a seller's percentage of successfully completed orders that is the number of orders a seller receives versus the number of orders cancelled or returned. Booksellers with a higher Bookseller Rating have cancelled fewer orders and received fewer returns. Ratings on various website normally shown in form of stars. Meaning of these star rating is as follows.

5 full stars :- 96-100% completion rate.

4 full stars :- 90-95% completion rate.

3 full stars :- 85-89% completion rate.

2 full stars :- 70-84% completion rate.

1 full stars :- 0-69% completion rate.

Thus Rating analyzer will extract this rating present on the web page and assign this value to w1 sub factor of weight ,W.

b) **Users Feedback Analyzer:** On web pages there are some sort of users feedback is there which user provide in form of likes/ dislikes, user reviews etc. These feedback shows the usability of the page content means whether the information/product available on the webpage is up to the requirement or not, whether user liked it or not.

User Feedback analyzer will extract these reviews from the web page and try to analyze them and set w2 a numerical value as shown in algorithm.

c) **Query-Page Content Matching :** In this sub module the query entered by user is matched with the content of the web page. The similarity percentage between query and web page will be calculated. Such as web page having a book shows its title as 'computer programming' and author as 'Y. S. Singh' And user entered query having book title 'computer' and author 'Singh' then the matching % is calculated between user query and title and author on the webpage.

This matching is done by using the Levenshtein distance method. The Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. It is named after Vladimir Levenshtein, who considered this distance in 1965. Levenshtein distance may also be referred to as edit distance, although that may also denote a larger family of distance metrics. It is closely related to pairwise string alignments.

d) **Weight Assignment :** Weight w can be calculated as a combination of w1,w2 and w3 as-

W=w1+w2+w3 where

W is the final weight assigned to the web page

W1 is one sub-factor of the weight on basis of rating provided on web page.

W2 is sub-factor of the weight on basis of users feedback provided on the web page.

W3 is sub-factor of the weight on basis of Query-Web Page Content matching.

3.1.2 . Page Source Rank Determiner Module

In this module rank of the source website from where the hidden pages are generated by filling form is determined. This page rank of source is determined with the help of Google ranking of the websites. The websites which Google ranks on the 1st page of of its search results for any given search term

are the ones that they consider to be the most relevant and useful. Google determine which websites are the most useful and relevant by using a complex algorithm (mathematical process) which takes into account 200+ different factors. Google doesn't let people know what those factors are, however, through a combination of research, testing and experience, some most important factors are identified as:-

- Keyword usage
- Site structure
- Site speed
- Time spent on site
- Number of inbound links
- Quality of inbound links

3.1.3. Frequency Calculation Module

As the hidden web crawler will crawl hidden web pages by filling HTML form on the various websites. It can be possible that hidden web crawler may fill same set of values in HTML form on more than one website. It may result in generation of web pages having same content from two different websites. Such as two websites of book domain may generate two web pages with different url having description of same book. Frequency calculator will calculate such number of pages and set this as frequency of web page.

3.1.4. Final Rank Assignment Module

In this module final rank value is assigned to each url . Rank value is used to arrange all urls in a ranked list. This rank value, Rv is calculated on basis of weight assigned to each url(W) , rank of source of web page(Sr) and frequency of the web page(f) as:

$$Rv = (W + Sr)*f$$

After calculating rank value of each retrieved url , store the rank value and display the results to the user after arranging all the urls in descending order on basis of rank value(Rv).

4. Conclusion And Future Work

Hidden Web data is now becoming highly important so extraction of hidden data is quite relevant. For getting accurate results, ranking of extracted Hidden web pages is needed. This paper entitled 'Design of a Novel Ranking Technique Of Hidden Web pages" provides a simple and efficient technique for ranking hidden web pages using some query dependent and independent factors. It will provide more accurate and efficient results to users.

This paper provide for ranking of hidden web pages, further for more better results we also need to explore better indexing technique and also need to explore various other query independent or dependent factors for ranking of hidden web pages.

REFERENCES

[1] http://en.wikipedia.org/wiki/Deep_Web

[2] The Deep Web: Surfacing Hidden Value, *September 2001, http://www.brightplanet.com/deepcontent /tutorials/DeepWeb/deep webwhitepaper.pdf*

[3] http://en.wikipedia.org/wiki/Surface_Web

[4] Bin He, Mitesh Patel, Zhen Zhang, Kevin Chen : "Accessing the Deep Web: A Survey" *Computer Science Department University of Illinois at Urbana-Champaign.*

[5] Chris Sherman and Garyprice : "The invisible web: uncovering sources search engines can't see ".

[6] Anuradha and A.K.Sharma : "A Novel Technique for Data Extraction from Hidden Web Databases" *in International Journal of Computer Applications, 2011.*

[7] Larry Page, and Sergey Brin, Rajeev Motwani, Terry Winograd, 'The PageRank Citation Ranking: Bring Order to the ' , *Technical report in Stanford University, 1998.*

[8] Wenpu Xing and Ghorbani Ali, 'Weighted PageRank Algorithm', *Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.*

[9] G.Kumar; N. Duhan; A.K. Sharma, 'Page Ranking Based on Number of Visits of Links of Web Page ', *International Conference on Computer & Communication Technology (ICCCT), 2011.*

[10] Neha Batra et al Content Based Hidden Web Ranking Algorithm (CHWRA*), IEEE International Advance Computing Conference (IACC) 2014.*

[11] Babita Ahuja , Dr. Anuradha "SCUM: A Hidden Web Page Ranking Technique", *International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 10 (November 2014).*

[12] Brian Wai Fung Wong's, "Deep-Web Search Engine Ranking Algorithm" ,MIT.

[13] Raju Balakrishnan's "Trust and Profit Sensitive Ranking for the Deep Web and On-line Advertisements ", *Arizona State University.*

[14] Saravanan Thirumuruganathan, Nan Zhang, Gautam Das :- "Rank Discovery From Web Databases" by *University of Texas at Arlington; George Washington University.*