Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org
Volume 5, Special Issue 2, December (2015)

Annotating Ranking Techniques for Hidden Web: A Review

Manvi, Komal Kumar Bhatia,Ashutosh Dixit
Department of Computer Engineering, YMCA University of Science and Technology, Faridabad

Abstract – The hidden or deep web refers to content that is hidden behind HTML forms. This contains a large collection of data that is unreachable by normal hyperlink-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. To access this content, one must submit valid input values to the HTML form. There are many methods for extraction data from hidden web. Various crawlers to access deep web content have already been developed by many researchers. These hidden web crawlers return huge result set for the user query so ranking of these results is needed. Ranking of hidden web may be classified into content based and structure based. Till now ranking of hidden web data is a big challenge, enough work has not been done in this area. In this paper, various ranking methods for the hidden web data will be explored.

Index Terms – Hidden Web, Deep Web, Ranking, Structure, Content Ranking

This paper is presented at International Conference on Recent Trends in Computer and information Technology Research on 25th& 26th September (2015) conducted by B. S. Anangpuria Instituteof Technology & Management, Village-Alampur, Ballabgarh-Sohna Road,Faridabad.

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 [1] refers to the part of the Web that can be crawled and indexed by general purpose search engines and the hidden Web [2] 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 include online flight and railway reservation systems, shopping websites, online book data stores, library catalogues etc. As the size of the web is increasing day-by-day, similarly the size of hidden web is also increasing [3]. A typical link based search engine cannot access the deep web pages because the content is hidden behind the web forms. This high quality 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 [4] are needed to extract information from websites having hidden web data. After extracting various hidden web pages it is needed to rank web pages to provide efficient results to the users.

The aim of the paper is to comparatively analyze the existing ranking algorithms or techniques for hidden web pages. Section 2 is Hidden web Search engine which shows the importance of ranking, section 3 explains various proposed ranking techniques for hidden web pages. Section 4 shows the comparison of ranking techniques and section 5 is conclusion and future work.

2. HIDDEN WEB SEARCH ENGINE

The search engine is a computer program that searches for the particular keywords entered by the user and returns a list of documents in which they were found. The search engines crawls the web and returns hundreds of web pages as result of user query [5]. The resultant pages generated by the search engine are first searched in its own local database and if the desired web pages are not found there then it fetches from web. The major components of search engine are Crawler, Indexer and Query processor as shown in figure 1.

Hidden Web crawler traverses the hidden web, fill various search forms and extract information hidden behind the html forms. It downloads the web pages and store them in a large database. It starts with seed URL and collects documents by recursively identifying and filling forms and storing the extracted URL’s into a local repository. The Indexer processes and indexes the pages collected by the crawler. It extracts keywords from each page and records the URL where each word has occurred.

The query processor is responsible for receiving and filling search requests from user. When a user fires a query, query engine receives it and after matching the query keywords with the index, returns the URL’s of the pages to the user. In general Query Engine may return several hundreds or

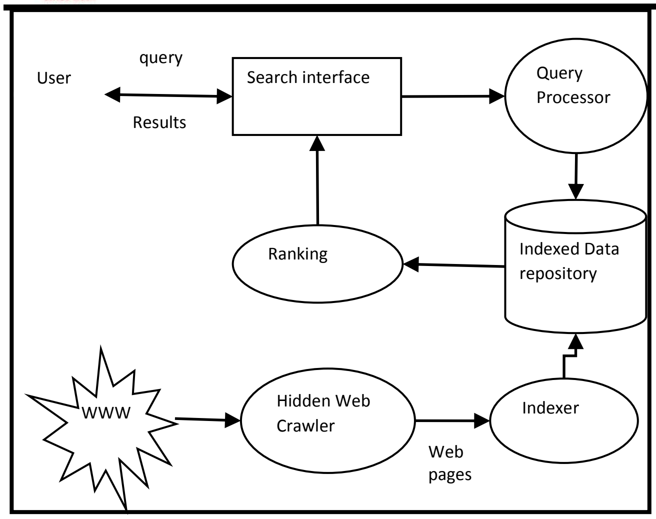

Figure 1. Architecture of a search engine

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 and hidden web. Some of the common page ranking algorithms of surface web are described in Page Rank Algorithm [2, 3], Weighted Page Rank Algorithm [4] and Hyperlinked Induced Topic search Algorithm [5].

3. RANKING TECHNIQUES

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. Search engines use two different kinds of ranking factors: **Query-dependent factors and Query Independent Factors** to calculate the rank of a web page. Query-dependent factors are all ranking factors 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 measures word documents frequency, the position of the query terms within the result page or the inverted document frequency, that are used commonly in any basic search engine. Some of the query independent factors are Link popularity, Click popularity and up to-datedness of the page etc. Few of the above stated important ranking techniques are discussed below.

3.1 Content Based Hidden Web Ranking Algorithm (CHWRA)[6]

In this paper N. Batra et al proposed a ranking algorithm which consists of four different attributes. These are: a) Page Rank b) Term Weighting Technique [TWT] c) User's Feedback and d) Visitor Count.

Following is the brief description of each term:

a) Page Rank: The Page Rank component in this paper checks the entire link structure of the website and calculates the PR value of web pages and distributes it to the links within the web page.

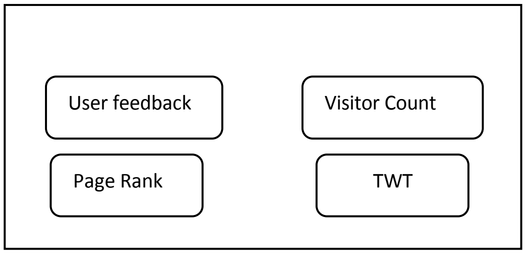

Figure 2: Ranking Attributes in CHWRA

The formula used in calculation of page rank value is given as:

$$PR(A)=(1-d)+d[PR(T1)/C(T1)+.....+PR(Tn)/C(Tn)] \quad (1)$$

b) Term Weighting Technique: The term weighting technique is based on probabilistic and vector space model. There are three main parameters used in calculating TWT. The parameters are document length, document frequency and term frequency.

c) User's Feedback: This method takes user's feedback into account in the form like and dislikes count. Like and dislike count are taken as the positive or negative response respectively to the web page and affects the popularity of the web page, thus affecting the rank value of the web page.

d) Visitor Count: In this method hits on the web page are considered as the visitor count. It is assumed that more the number of hits on the web page and higher the popularity of the web page.

Limitations of CHWRA:

i. The Page Rank algorithm is commonly used by the conventional search engines. It is not effective for Hidden Web pages.

ii. Some fraud websites knowingly add a lot of popular keywords which are not related to the content in the title or the content of the page to cheat search engines.

iii. This technique uses user feedback and visitor count for ranking but it didn't explain well how these are taken into consideration.

3.2 *SCUM: A Hidden Web Page Ranking Technique [7]:*

By Anuradha et.al. says that the pages from deep web are ranked by analysing the structure, the contents of the web page, measuring the human interest and attention devoted by them to the pages. In order to extract the characteristics of the web pages such as structure and content of web pages the web mining is required. Web mining is the use of data mining techniques to automatically discover and extract information from web.

Ranking of Hidden Web pages here uses three steps:

a) Structure Page Rank Calculation
b) Content Page Rank Calculation
c) Usage Page Rank Calculation

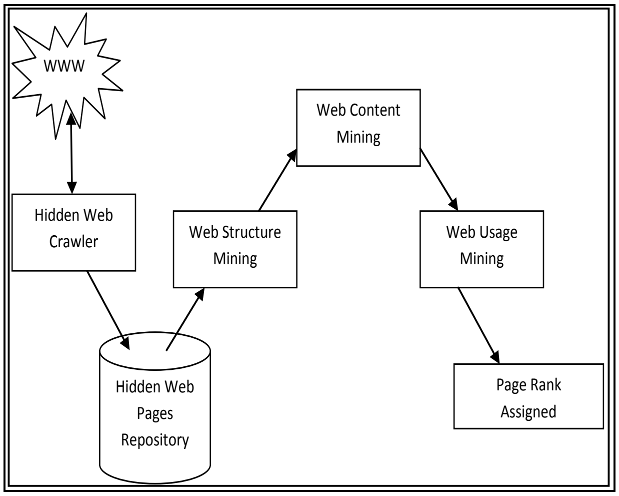

Figure 3: Architecture for rank calculation

i) ***Content Page Rank Calculation**:* The content of extracted web pages will be analyzed and on the basis of the content the pages will be ranked. The relevance of the page will be analyzed on the basis of the domain, the quality of content, spam detection.

$$\text{Rank}_{(c)}A = \text{Relevance} + \text{Quality}$$

ii) ***Structure Page Rank Calculation***: Graph databases are used, the nodes represent the entities (web pages) and edges represents the relationships (here inlinks and outlinks). The pattern recognition is done very easily in the case of graph database. For ex: it is very easy to extract the pattern that the node no. 699 of domain car receives all the inlinks from the web pages of different domain i.e. property.

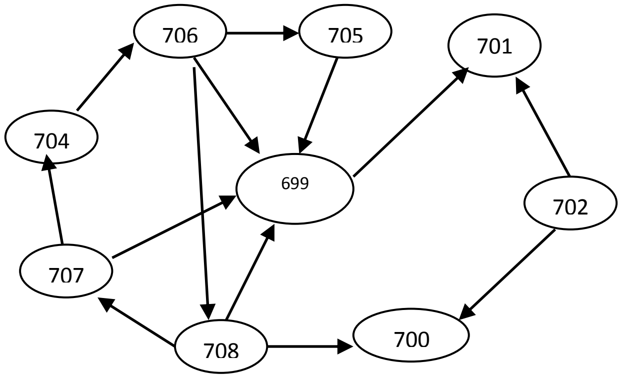

Figure.4. Interconnection of web page.

***iii) Usage Page Rank Calculation**:* In this the users access pattern and the time spend by the user on the web pages will be analyzed. When user will revisit and issue the query his pre-processed access pattern will be fetched and rank of web pages will be updated dynamically.

$$\text{Rank}_{(u)}A = \text{Access}(\text{User Can} + \text{User } A_{tn}) \qquad (2)$$

Here Rank(u) A is the usage page rank of page A, User $A_{c\ n}$ refers to the no of clicks(c) of User n on page A and User A_{tn} refers to the time(t) spend by User n on page A.

iv) ***Overall page Rank Calculation**:* Once the rank has been calculated by using all the mining techniques the final rank will be calculated. The formula for the final page rank is given below.

$$\text{Rank}_{(f)}A = \text{Rank}_{©}\ A + \text{Rank}_{(s)}\ A + \text{Rank}_{(u)}\ A. \qquad (3)$$

Where $\text{Rank}_{(f)}A$ refers to the final page rank of page A.

Advantages of the SCUM algorithm

i) It not only considers the links but also the contents of the web pages for the rank calculation.
ii) The interest of the users is also considered. So every user is able to see the pages of its own interest at the top.
iii) The proposed algorithm will rank the pages from the hidden web and surface both.

Limitations of SCUM algorithm

i) This algorithm in structure rank calculation focus on inlinks and outlinks on the page. No. of inlinks are not relevant for calculating hidden web page ranking.
ii) In usage rank calculation emphasis is on particular user. Thus this algorithm is more user based but user priority may change with time.

iii) Proper design and implementation of algorithm is not given.

3.3 *Deep-Web Search Engine Ranking Algorithm by Wong et al [8] :*

This algorithm introduce the ten-factor model for ranking deep-web search results. The ten-factor linear regression model has the following general form:

$$R=\alpha_0 \sum_0^{10} \alpha i\ .F_i + \text{€} \qquad (4)$$

The entity R, which ranges from 0 to 10, is defined as the quality score of a particular search result. These ten factors can be classified among best-fit scoring functions using quality factors and a dynamic weighting algorithm that changes the factor weighting based on user behavior.

Definition of 10 factors

F1=Binary indicator for presence of business supplied image
F2 = Binary indicator for use of Google images Visually related
F3 Continuous variable C (0... 1) provided by user for indicating the quality of "More Photos" tab
F4 Distance metric for indicating proximity to desired search location Distance/Time related
F5 Time metric for indicating closeness to search date
F6 Binary indicator for presence of "Description"
F7 Binary indicator for including a "Detailed location"
F8 Binary indicator for including a "Phone number" Data related
F9 Binary indicator if price is available
F10 Continuous variable C (0...1) provided by user for indicating the

Search engine utilizes two factor scoring function to rank results – a combination of distance score d and referral score r. The distance score is inversely proportional to the physical distance between search result and location of interest. The referral score represent the popularity of the result amongst the searched websites.

***Advantages*:**

i. This algorithm is scalable- and requires minimal pre-processing to generate the factor weightings.

ii. It ranks results considering user behavior and requirements.

iii. This technique efficiently ranks the hidden web pages.

3.4 *Rank Discovery From Web Databases by Saravanan et.al.[9]:*

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.

3.5 *Trust and Profit Sensitive Ranking for the Deep Web by Raju Balakrishnan et al[10]:*

This work considered the emerging problem of ranking the deep web data considering trustworthiness and relevance. In this paper end-to-end deep web ranking is discussed 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.*

A method namely SourceRank is developed to assess the trustworthiness and relevance of the sources based on the inter-source agreement. Second the Source Rank is used to consider the topic of the agreeing sources in multi-topic environments. Further, a ranking sensitive to trustworthiness and relevance for the individua results returned by the selected sources is formulated.

4. COMPARISON OF EACH WORK

Below is the comparison table made in accordance with the fact that which work has used query dependent and which has used query independent factors. Also there is a column made to specify whether the work can be used in Hidden Web or not. One can make other factors also for example whether ontology is used or not. Table 1 delineates the comparison.

Ranking Algo	**Use of Query Dependent factors**	**Use of Query Independent Factor**	**Technique Used**	**Relevancy to Hidden Web**	**Remarks**
Content based hidden web ranking [6]	NO	YES	Pagerank and Term Weighting Technique	Less	Do not emphasis on ranking of hidden web pages.
SCUM[7]	NO	YES	Web structure, content and usage mining	More	
Brian Wong et al[8]	YES	YES	Best-fit scoring functions using ten quality factors and a dynamic weighting algorithm	More	Scalable and requires minimal pre-processing to generate the factor weighting
Rank Discovery by Saravanan et al[10]	YES	NO	Developed RANK-EST(algorithm) which interleaves two separate procedures for handling high and low ranked tuples,	Partial	It discuss the feasibility of rank discovery for each type of ranking function.
Raju Balakrishnan et al[9]	YES	YES	Various score functions are used.	More	

5. CONCLUSION

In this paper we present the review on various techniques for ranking of Hidden web pages. Algorithms to the corresponding techniques are also discussed with each technique. From the existing techniques it can be concluded that an efficient algorithm for ranking hidden web contents should be based on source of the page, structure of the page, content and popularity of the web page. Ranking the web pages in a particular domain should also be based on user requirement or user search history.

REFERENCES

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

[2]. [2].The Deep Web: Surfacing Hidden Value, September2001,http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/deepwebwhitepaper.pdf

[3]. [3].Komal Kumar Bhatia, A.K.Sharma, "A Framework for an Extensible Domain- specific Hidden Web Crawler (DSHWC)", communicated to IEEE TKDE Journal Dec 2008.

[4]. Bin He, Mitesh Patel, Zhen Zhang, Kevin Chen : "Accessing the Deep Web: A Survey" Computer Science Department University of Illinois at Urbana-Champaign,2006.

[5].

[6]. [5]. N. Batra, A. Kumar, Dr. D. Singh and Dr. R.N. Rajotia "Content Based Hidden Web Ranking Algorithm (CHWRA)" Advance Computing Conference (IACC), 2014 IEEE International, 2014 IEEE, DOI 10.1109/IAdCC.2014.6779390 Page(s): 586 – 589.

[7]. [6]. 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).

[8]. [7]. Brian Wai Fung Wong's, "Deep-Web Search Engine Ranking Algorithm", MIT.

[9]. [8]. Saravanan Thirumuruganathan, Nan Zhang, Gautam Das :- "Rank Discovery From Web Databases" by University of Texas at Arlington; George Washington University published in Proceedings of the VLDB Endowment, Vol. 6, No. 13 Copyright 2013 VLDB Endowment 21508097/13/13.

[10]. [9]. Raju Balakrishnan's "Trust and Profit Sensitive Ranking for the Deep Web and On-line Advertisements ", Arizona State University ,August 2012.

[11]. [10]. Jianguo Lu "Ranking Bias in Deep Web Size Estimation Using Capture Recapture Method", School of Computer Science, University of Windsor, Canada, March 12,2010.