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. 3, Issue 5, May 2015

A Comparative Study of Ranking Techniques for Hidden Web and Surface Web

Jyoti Yadav, Anil Kumar, Seema Rani

Student M.Tech, Dept. of CE, YMCA University of Science & Technology, Faridabad, India

ABSTRACT: The web consist of Surface web and hidden web. Surface web is also known as publically indexable web. It can be accessed by search engines using hyperlinks present on the pages and using simple keyword matching schemes. Hidden web refers to content that is hidden behind HTML forms. This contains a large collection of data that are 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. The hidden and surface 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 the web data is a big challenge. Various scholars tried to propose better and efficient techniques for ranking. In this paper, various ranking methods for the hidden web as well as surface web will be explored.

KEYWORDS: Surface Web, Hidden Web, Deep Web, Ranking Techniques.

I. INTRODUCTION

The World Wide Web (WWW) byname The Web, the leading information retrieval service of the Internet (the worldwide computer network) 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 that encompasses normal crawler[7] which work as basic keyword matching scheme while hidden Web[1] refers to the abundant information that is "hidden" behind the query interfaces and not directly accessible to the traditional search engines, so to access hidden web search engines must be enabled with a special Hidden Web crawler. Example of surface web resources include various websites through which one can move simply by clicking on hyperlinks present on the websites while 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 [4]. 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 along with general crawlers [7], specific hidden web crawlers [6] are also needed to extract information from web having websites with hidden web data. After extracting various 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 both surface web as well as hidden web. Section 2 is Web Search engine, section 3 explains various existing and proposed ranking techniques for surface and hidden web pages. Section 4 shows the comparison of ranking techniques and section 5 is conclusion and future work.

II. 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 perform following basic tasks:

- They search the Internet based on important words.
- They keep an index of the words they find, and where they find them.
- They allow users to look for words or combinations of words found in that index.
- Search engine technology has had to scale dramatically to keep up with the growth of the Web.

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. 3, Issue 5, May 2015

The search engines crawls the web and returns millions of web pages as result of user query. The resulted url list generated by the search engine is either first searched in its own local database and if the desired web pages is not found there then it fetches from web. The major components of search engine as shown in fig. 1 are as follows: Crawler, Indexer, Query processor, Ranking

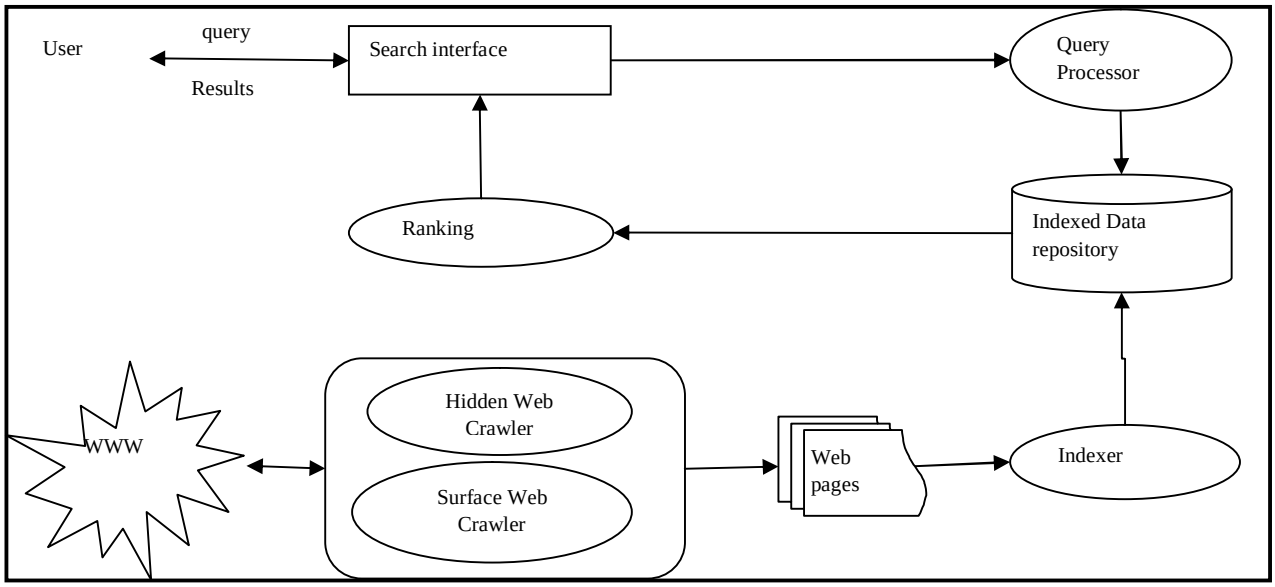

Fig 1. Architecture of a Web search engine

❖ **Detailed explanation of each Module :**

2.1 Crawler Module: This module has two crawler i.e. one for surface web and other for hidden web .Both crawler download the various pages available on the web and store them in data repository for indexing.

i) **Surface web crawler:** It search the seed URL on the web and linked page with that single URL. Crawler behavior is dependent on following policies [21]:-

- Selection Policy: - This policy decides that which pages will be downloaded or which will be discarded.
- Revisiting policy: This policy decides that which page will be revisited for the changes.
- Parallelization policy that states how to coordinate distributed web crawlers.

ii) **Hidden Web Crawler:** Strategic approaches are needed to target deep Web content and for this specific hidden web crawlers are used. They need to fill the search form present on the website and for filling these form they need specific database and mapping functions [22].Data extracted from the results of one Web form submission can be taken and applied as input to another Web form thus establishing continuity across the Deep Web in a way not possible with surface web crawlers.

2.2 Indexer Module: Search engine indexing collects, parses, and stores data to facilitate fast and accurate information retrieval. Popular engines focus on the full-text indexing of online, natural language documents. Media types such as video and audio and graphics are also searchable. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power. For example, while an index of 10,000 documents can be queried within milliseconds, a sequential scan of every word in 10,000 large documents could take hours.

2.3 Query Processor: Query Processor of a search engine receives search requests from users in the form of keywords or content or phrases. It then tokenize the query, remove stop words and apply stemming on it. And search the processed query in the index of the data repository.

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. 3, Issue 5, May 2015

2.4 Ranking module: In general Query Processor 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. This ranking of the web pages can be done on basis of some query dependent and query independent factors. Some general ranking factors are as:

a. Location of terms in web page**. b.** Term frequency **c**. Link analysis **d.** Web page Popularity **e.** Query – Page content matching **f.** Date of Publication **g.** Length or size of web page. **h**. User feedback **i**. Proper nouns

III. RANKING TECHNIQUES

Many researchers are trying to develop novel ideas to develop better ranking technique for Surface web as well as hidden web pages in order to improve the quality of user results. A brief overview at few of them is given in the following subsections:

3.1 PageRank Algorithm [8]: Page Rank algorithm uses link structure to determine the importance of web page. This algorithm is based on random surfer model. The random surfer model assumes that a user randomly keeps on clicking the links on a page and if she/he get bored of a page then switches to another page randomly. Thus, a user under this model shows no bias towards any page or link. PageRank (PR) is the probability of a page being visited by such user under this model. Page Rank algorithm assumes that if a page has a link to another page then it votes for that page. Therefore, each inlink to a page raises its importance. PageRank is a recursive algorithm in which the PageRank of a page depends upon the PageRank of the pages linking to it. Thus, not only the number of inlinks of a page influences its ranking but also the page ranks of the pages linking to it. A page confers importance to the pages it references to by evenly distributing its PageRank value among all its outlinks.

The PageRank of page P is given as, follows:

$$PR(P) = 1 - d + d \sum^{n}_{t=0} (PR(N_i) / O(N_i) \quad \text{Where}$$

N0...Nn are the pages that point to page P.

O(Ni) is defined as the number of links going out of page P.

The parameter d is a Damping factor, the probability of user's following the direct links which can be set between 0 and 1.

3.2 Weighted PageRank Algorithm[9] : Weighted PageRank (WPR) algorithm is a modification to the PageRank algorithm. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance. The importance is assigned in terms of weight values to the incoming and outgoing links. Quality of the pages returned by this algorithm is high as compared to PageRank algorithm.

3.3 Hyperlink-Induced Topic Search Method[10]: HITS algorithm, also known as Hubs and Authorities, is a link analysis algorithm for the web. It is executed at query time and is used to modify the ranking of the results of a search by analyzing the link structure of the pages that will appear in the result of the search. HITS algorithm assigns two different values to each web page: its authority value, and its hub value. The authority value of a page represents the value of the content in the page; meanwhile the hub value estimates the value of its links to other pages. The first step in the HITS algorithm is to retrieve the set of pages in the result of the search, as the HITS algorithm only analyzes the structure of the pages in the output of the search, instead of all the web pages.

3.4 Page Ranking based on Link-Visits (PRLV) [11]: This techniques an efficient page ranking mechanism that roams around two major web mining techniques namely, the Web Structure Mining and Web Usage Mining. It takes the user visits of pages/links into account that directs the calculation of the grandness and

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. 3, Issue 5, May 2015

relevance score of the web pages. The system constitutes many subtasks, till final rank gets determined, which are outlined below:

- Storage of user's access information (hits) on an outgoing link of a page in related server log files.
- Fetching of pages and their access information by the targeted web crawler.
- For each page link, computation of weights based on the probabilities of their being visited by the users.
- Final rank computation of pages based on the weights of their incoming links.
- Retrieval of ranked pages corresponding to user queries.

In this, ranking is assigned to pages according to clicks a page has got from users with respect to some query, and this rank, instead of being distributed evenly among pages, is assigned according to the weight their incoming links possess.

3.5 SCUM : A Hidden Web Page Ranking[12]: In this Babita Ahuja et al[12]proposed ranking of Hidden Web pages having three steps :

a) Structure Page Rank Calculation: The web pages from the WWW are highly connected. Graph databases are used, the nodes represent the entities (web pages) and edges represents the relationships (here inlinks and outlinks). The graph is created by using NEO4J and cypher query language.

b) Content Page Rank Calculation: The content mining is extraction of knowledge from text in the web pages. In this the content of extracted web pages will be analyzed and on the basis of the contents 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.

c) 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.

3.6 Deep-Web Search Engine Ranking Algorithm [13]: In this paper Brian Wong et al proposed ranking algorithm which utilizes best-fit scoring functions using quality factors and a dynamic weighting algorithm that changes the factor weighting based on user behavior. 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.

3.7 Content Based Hidden Web Ranking Algorithm (CHWRA)[14] : In this paper N. Batra et al proposed a ranking algorithm which consists of four different attributes. These are:- Page Rank , Term Weighting Technique [TWT], User's Feedback and Visitor Count.

The PageRank component checks the entire link structure of the network .The term weighting technique is based on probabilistic and vector space model. There are three main parameters used in calculating TWT i.e document length, document frequency and term frequency. This technique takes user's feedback into account in the form like and dislikes count while the 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.

3.8 Rank Discovery From Web Databases[15] : In this paper Saravanan, Nan Zhang and Gautam Das et al[29]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.

3.9 Trust and Profit Sensitive Ranking for the Deep Web and On-line Advertisements [16] : In this Raju Balakrishnan et al[16] 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

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. 3, Issue 5, May 2015

of the deep web databases (ii) topic sensitive ranking of the sources (iii) ranking the result tuples from the selected databases has been discussed.

3.10 Ranking of Web Documents using Semantic Similarity[17] : In this paper Poonam Chahal et al have proposed a ranking scheme for the semantic web documents by finding the semantic similarity between the documents and the query which is specified by the user. The novel approach proposed in this paper not only relies on the syntactic structure of the document but also considers the semantic structure of the document and the query. The approach used here includes the lexical as well as the conceptual matching. The combined use of conceptual, linguistic and ontology based matching has significantly improved the performance of the proposed ranking scheme. Poonam Chahal et al explored all relevant relations between the keywords exploring the user's intention and then calculate the fraction of these relations on each web page to determine their relevance with respect to the query provided by the user.

3.11 SR Rank algorithm [18] : In this paper K. P. Shyam et al used the reinforcement concepts and the link structure of the web pages to rank the web pages. In SR Rank algorithm the Agent is the surfer of the wed and the State is the each web page. The surfer (agent) clicks on the any available link in each page (state) and traverse the web pages in an uniform probability and goes to the next state. Therefore, the agent clicks randomly on the available links in order to traverse the web pages with a uniform probability. In other words, when an agent selects a link by clicking randomly on one of the available link in the current state, then the policy π is equal to 1/O , where O denoting the current state is the out degree of the current state. When a transition occurs from j to i , where j is the current state.

3.12 An Offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters [20] : This paper describes the new algorithm for calculating web page rank according to different parameters. The proposed algorithm called M-HITS (Modified HITS) is a new version of HITS algorithm. It is developed by extending the properties of HITS algorithm. Author present new algorithm in which six parameters are used to evaluate rank for web page. Future work can be done by using some AI techniques in addition to these proposed techniques to improve the rank of web pages.

IV. COMPARISONOF VARIOUS RANKING TECHNIQUES

In this section above explained all the ranking techniques are compared on basis of some important parameters such as query dependency, technique used, advantages and limitations of the various techniques.

Table 1. Comparison of All Ranking Techniques

S.No	Ranking Algorithm	Query Dependency	Technique / Model Used	Specific For	Remarks
1	Page Rank Algorithm[8] ,1998	Query Independent	Random surfer model	Surface Web	**Advantage:** -It returns important pages as Rank is calculated on the basis of the popularity of a page. -Less time consuming and more feasible. **Disadvantages:** -it favors older pages.
2	Weighted Page Rank Algorithm[9] ,2004	Query Independent	Web structure mining	Surface Web	**Advantage:** -It assigns larger rank value to more important pages. - more efficient than pageRank **Disadvantages:** -it considers only link structure not the content of the page, it returns less relevant pages to the user query.
3	Hyperlink-Induced Topic Search Method[10]	Query Dependent	Web Structure and Web Content analysis	Surface Web	**Advantage:** -It returns More Relevant pages to user. **Disadvantages:** -it is time consuming.

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. 3, Issue 5, May 2015

4	Page Ranking based on Link-Visits (PRLV) 2011[11]	Query Dependent	Web Structure Mining and WebUsage Mining	Surface Web	**Advantage:** -It returns relevant pages as dependent on user query and on basis of user history. **Disadvantages:** -Need additional database to store user history
5	Scum 2014 [7]	Query Independent	Web structure, content and usage mining	Hidden Web	**Advantage:** -It makes use of all aspects of web mining to calculate the page rank. **Disadvantages:** -It is more user based but user priority may change with time.
6	Deep-Web Search Engine Ranking Algorithm[13]	Uses both Query Independent and dependent factors.	Best-fit scoring functions using ten quality factors	Hidden Web	**Advantage:** -This algorithm is scalable and requires minimal pre-processing to generate the factor weightings.
7	Content based hidden web ranking, 2014 [6]	Query Independent	Pagerank and Term Weighting Technique	Hidden Web	**Advantage:** -It returns more important pages as it uses Page rank. **Disadvantages:** -Less relevant Results.
8	Rank Discovery From Web Databases,2013[15]	Query Dependent	RANK-EST(algorithm) which interleaves two separate procedures for handling high and low ranked tuples,	Surface Web	**Advantage:** It 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
9	Trust and Profit Sensitive Ranking for the Deep Web,2012 [16]	Uses both Query Independent and dependent factors.	Various trust score functions are used.	Hidden Web	**Advantage:** -High relevancy in results. -Trustworthy results. **Disadvantages:** -it is time consuming technique
10	Ranking of Web Documents using Semantic Similarity, 2013[17]	Query Dependent	Syntactic and Semantic Structure Mining.	Hidden Web	**Advantage:** - It uses conceptual, linguistic and ontology based matching functions so give better results. **Disadvantages:** -Some Fraud websites may use same title and name but without any useful information.
11	SR Rank algorithm , 2013[18	Query Independent	Link Structure and Reinforcement concepts	Surface Web	**Advantage:** -This algorithms provide the users with the results determining the links availability in a particular page. **Disadvantages:** -Less Efficiency.
12	An Offline SEO Based Algorithm to Calculate Rank , 2013[20]	Query Dependent	Web Structure and Web Content analysis (Extended HITS Algorithm with more parameters).	Surface Web	It is developed by extending the properties of HITS algorithm using some more parameters.

V. CONCLUSION AND FUTURE WORK

In this paper we explored the various proposed techniques for ranking of web pages. Some of these techniques are query dependent, some are independent and few uses combination of both. From the existing techniques it can be concluded that ranking algorithm should consider source, content, results and popularity of the web page for 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. 3, Issue 5, May 2015

the various web pages. Ranking of the web pages in a particular domain specially in hidden web should also be based on user requirement or user search history. Hence a better efficient ranking technique which can rank both hidden web pages and surface web pages effectively is required.

REFRENCES

[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,2006.

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

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

[7]. Shalini Sharma, "*Web Crawler*", published in International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 4, April 2014.

[8]. Sergey Brin and Lawrence Page. The anatomy of a large-scale hyper textual web search engine. In Computer Networks and ISDN Systems, pages 107-117, 1998.

[9]. Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm" Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04) 0-7695- 2096-0/04 $20.00 © 2004 IEEE.

[10]. J. Kleinberg. Authoritative sources in a hyperlinked environment. In 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[11]. A. K. Sharma, Neelam Duhan, Gyanendra Kumar, "A Novel Page Ranking Method based on Link- Visits of Web Pages". Int. J. of Recent Trends in Engineering and Technology, Vol. 4, No. 1, Nov 2010, pp: 58-63.

[12]. 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).

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

[14]. 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.

[15]. 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. 13Copyright 2013 VLDB Endowment 21508097/13/13.

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

[17].Poonam Chahal, Manjeet Singh, Suresh Kumar, "Rankingof Web Documents using Semantic Similarity", Information Systems and Computer Networks (ISCON), 2013 International Conference on 2013 IEEE, DOI 10.1109/ICISCON.2013.6524191 Page(s): 145 – 150.

[18]. K. P. Shyam, Sharath Jagannathan, Maheswari Rajavel ,"A New Technique for Ranking Web Pages and Adwords",International Journal of Computer Applications (0975 – 8887) Volume 82 – No 12, November 2013.

[19]. Ashish Jain, Rajeev Sharma, Gireesh Dixit and Varsha Tomar ,"Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages", Communication Systems And Network Technologies (CSNT), 2013 International Conference On, 2013 IEEE, DOI 10.1109/CSNT.2013.137 Page(S): 640 – 645.

[20]. Parveen Rani and Er. Sukhpreet Singh, "An Offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters", International Journal Of Computers & Technology Vol. 9, No 1 J U L Y 1 5, 2 0 1 3.

[21].Joeran Beel, Bela Gipp, and Erik Wilde, "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co." , in Journal of Scholarly Publishing, 41 (2): 176–190, January 2010.

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