International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 10 (November 2014) www.ijirae.com

SCUM: A Hidden Web Page Ranking Technique

Babita Ahuja*
Information Technology, M.D.U University

Dr. Anuradha
Computer Science, YMCAUST, University

Abstract— Hidden Crawler enables to index the pages from World Wide Web, which the traditional crawlers fail to do. The hidden web crawlers are not able to adequately rank the pages from deep web. The hidden web data is very much high in quality in comparison with the surface web data. So a very large and high quality data of WWW is left unranked. These days it has become a need and trend among the website administrators to make the web pages of their websites dynamic. Because of this need the data sank into the web deeply. The traditional search engines fail to index and rank these pages. The end users have to do extract the important and relevant pages from deep web manually. So there is a need of novel page ranking techniques for ranking pages from deep web. So here we have proposed a novel technique SCUM (structure, content and usage mining) for ranking the pages from the deep web. The pages will be ranked and ordered by new ranking techniques. While ranking the deep web pages the behaviour of end user will also be considered. These ordered set of pages will be then displayed to the end users and effort of end users in extracting their desired pages will be reduced drastically.

Keywords— WWW, page rank, surface web, hidden web, Web Mining, RDF, Graph databases, SPARQL, NEO4J, Cypher Query.

I. INTRODUCTION

In this digital world every one seeks for the information on the internet. Most Website developers support to place the data in the relational databases. The high quality information is placed in databases by universities, libraries and government agencies and is being made available online. The 99% part of WWW is made up of the hidden web pages as shown in Fig. 1 [1]. The Web Query Interfaces acts as the front door to access the data from these databases [2]. So in order to access the relational databases residing behind query interfaces the URL will be created automatically by using the Dynamic Query Processing for Hidden Web Extraction [3]. After dynamic query processing the web pages are fetched. So in this paper we propose a novel page ranking technique SCUM to rank the web pages fetched from the Hidden web. 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.

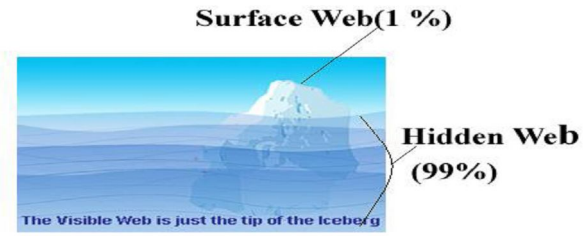

Fig. 1 Hidden Web and Surface Web

II. WEB MINING

Web mining is a kind of data mining technique to extract the knowledge from the web pages. The knowledge can be structure or content of web pages and services. Web mining is categorized into three types [4], [5] as shown in Fig. 2.

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

In web content mining [6] the text from web pages is extracted and processed. The traditional search engines use the web content mining to index the web pages on the basis of the text and media present on web pages. This indexing later helps the search engines to fetch the results according to user query. Few tools are also developed to automate the processing of text and other graphic content present on web pages. Some of the web content mining tools are Screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda, and WebContent Extractor.

Web structure mining [7] focuses on the inter-connected property of WWW. The WWW is a highly connected. The web pages are connected by the anchor tags or the hyperlinks. Web structure mining is very much used for calculating the rank of web pages. It is also used in find the duplicate websites.

Web usage mining [8] is the technique of understanding the behaviour, need and problems faced by users visiting the websites. When a user visits a website then the blueprint of his complete access path is recorded in the server log files.

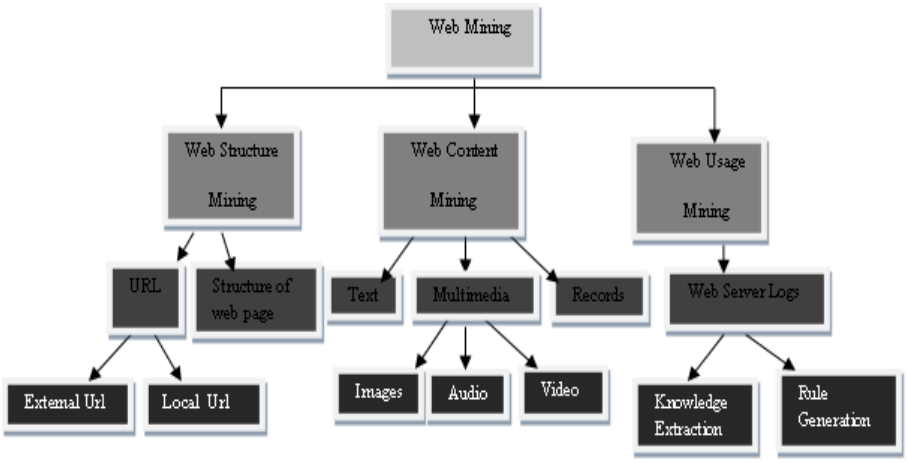

Fig. 2 Taxonomy of Web Mining

The server log files records all the browsing activities of every user. Web usage mining processes the contents of server log files and extracts the useful information. This information is very helpful for website management purpose and at the same time enhances the sales of ecommerce websites drastically.

III. RELATED WORK

The WWW has a limitless collection of web pages. The end users have to fetch the web pages of their need from this huge ocean of data. It is very difficult for users to find the desired data from WWW. So many researchers have developed the algorithms to rank the web pages from WWW. Mostly the page ranking algorithms uses structure web mining to rank the pages. Three important page ranking algorithms are:

- Pagerank algorithm
- Weighted page rank algorithm
- Hyper-link induced topic search algorithm

The pagerank algorithm was developed by Brin and Page at Stanford University [9]. This algorithm is based on the web structure mining. Pagerank algorithm counts the number of incomings links or inlinks to a web page. If a page receives a inlink from a important web page then the page is assigned a higher rank. So in this algorithm the importance of a web page linking to a page is also considered. Brin and Page gave formulae to count the page rank. The formulae to calculate the PageRank of a page A is given below

$$PR(A)=(1-d) + d(PR(T_1)/C(T_1) +..+ PR(T_n/C(T_n))$$

Here $PR(T_i)$ is the Pagerank of the Pages Ti which links to page A, $C(T_i)$ is number of outlinks on page T_i and d is damping factor. It is used to stop other pages having too much influence. The total vote is "damped down" by multiplying it to 0.85. It is an iterative algorithm which follows the principle of normalized link matrix of web. Pagerank of a page depends on the number of pages pointing to a page.

Weighted page rank algorithm was proposed by Wenpu Xing and AliGhorbani [10]. This algorithm is based on the web structure mining. This algorithm is an extension of Pagerank algorithm. The pagerank algorithm assigns a rank to a page depending on the importance of the pages which are linking to it. It does not check the importance of the page whose rank is being calculated. Weighted page rank algorithm does not assign rank by dividing its importance equally among all web pages it connect but also on the importance of those web pages itself. In this algorithm the importance is represented in terms of weight values to incoming and outgoing links. They are denoted as $W_{in}(m, n)$ and $W_{Out}(m, n)$ respectively. $W_{in}(m, n)$ is the weight of link(m, n) . It is calculated on the basis of number of incoming links to page n and the number of incoming links to all reference pages of page m.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \in R(m)} I_p}$$

In is number of incoming links of page n, I_p is number of incoming links of page p, R (m) is the reference page list of page m. $W_{Out}(m ,n)$ is the weight of link(m, n). It is calculated on the basis of the number of outgoing links of page n and the number of outgoing links of all the reference pages of page m.

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p \in R(m)} O_p}$$

O_n is number of outgoing links of page n, O_p is number of outgoing links of page p. So the weighted page rank is given by formula shown below.

$$WPR(n) = (1-d) + d \sum_{m \in B(n)} WPR(m) W^{in}_{(m,n)} W^{out}_{(m,n)}$$

Klienberg [11] introduced the new algorithm Hyper-link Induced Topic Search (HITS). This algorithm is based on the web content mining. In this algorithm the web pages are of two types hubs and authorities. Hubs are the pages that act as resource lists while authorities are the pages having important content. A hub page of a domain is good if it is linked with many authoritative pages of the same domain. An authoritative page of a domain is good if many good hub pages of that domain points to it. The formulae to calculate the weight of Hub (H_p) and the weight of Authority (A_p) are given below.

$$H_p = \sum_{q \in I(p)} A_q$$

$$A_p = \sum_{q \in B(p)} H_q$$

Here H_q is Hub Score of a page, A_q is authority score of a page, I ($_p$) is set of reference pages of page p and B (p) is set of referrer pages of page p, the authority weight of a page is proportional to the sum of hub weights of pages that link to it. Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

The algorithms discussed above do not utilize the capabilities of web mining techniques to the fullest. These algorithms calculate the rank of the web pages that are stored statically on the server of websites. The ranks calculated by these algorithms do not change on the basis of the behaviour of the end users also. So I proposed a new page ranking algorithm that will not only use the entire web mining techniques but will also calculates the page rank of dynamic web pages which are actually deep web pages and will also keep track of end user interests.

IV. PROPOSED WORK

All of the page rank algorithm works on the surface web. Most of the algorithms do not focus on the hidden web. The hidden web is 99% of the WWW. So it means major part of the web pages in WWW are left unranked. The page ranks are assigned to the web pages only by analysing the links available on the web pages. Most of the algorithms do not examine the content of the web pages and the usage of the web pages while calculating the importance of the web page. The proposed algorithm will analyse the web page on the basis of not only the links in the page but also the content and the usage statistics of the page as shown in Fig. 3. The rank of the pages will not be constant. It will vary from user to user by learning the interest of the user in the page. The interest of the user will be learned by examining and processing the website server log. The major part of the deep web resides behind the query interfaces. So these query interfaces act as a gateway to access the data behind them. The traditional crawlers are not able to pass this gateway. By using the dynamic query string processing technique [3] the data behind the query interfaces will be fetched. Once the data behind these query interfaces is fetched, then they will be ranked by the SCUM algorithm.

A. Steps of the proposed Algorithm

Thousands of web pages from hidden web are fetched dynamically by processing the query which user has entered in a single search textbox like Google. Now the primary goal of the SCUM technique is to provide the most relevant information and the best results at the top to the users to cater their needs. In order to accomplish this task the web pages downloaded from the hidden web are ranked. The ranking of the pages will be done in three steps.

- Structure Page Rank Calculation
- Content Page Rank Calculation
- Usage Page Rank Calculation

1) *Structure Page Rank Calculation*: The content mining is extraction of knowledge from text in the web pages. When users submit the query, thousands of web pages behind query interfaces are fetched. So the content of these pages will be analysed and on the basis of the contents the pages will be ranked. The relevance of the page will be analysed on the basis of the domain, the quality of content, spam detection. In the beginning the spam websites will be removed. The spam websites do keyword stuffing for disguise. They insert the keyword inside the text and keep the font color of keyword as background color. So they easily bluff the search engines. So the foreground color of text and background color will also be checked. After spam website elimination the relevance and content quality check of web pages will be done. The web pages are created on the machines for the end users i.e. human beings. But the irony is that the machines are not able to understand the content on the web pages. In order to make the content present on the web page machine understandable we will create a Resource Description file (RDF) for every web page [12]. The RDF document will contain all the meta data of the web page and the complete description of all resources present on the page. We will use SPARQL a query language for RDF. By using the SPARQL [13] we will make query on the RDF and find out the relevance and quality of data on the web page. Using SPARQL we will fetch the title, headings and meta tags of page.

The system will check the degree of relevance of the pages. The SPARQL query will be used to fetch the data record from RDF. The quality of the data will be checked.

The following steps are done for content mining.

• Spam website Removal

• RDF document creation of web pages

• SPARQL query to check the relevance and quality

Rank(c)A = Relevance + Quality

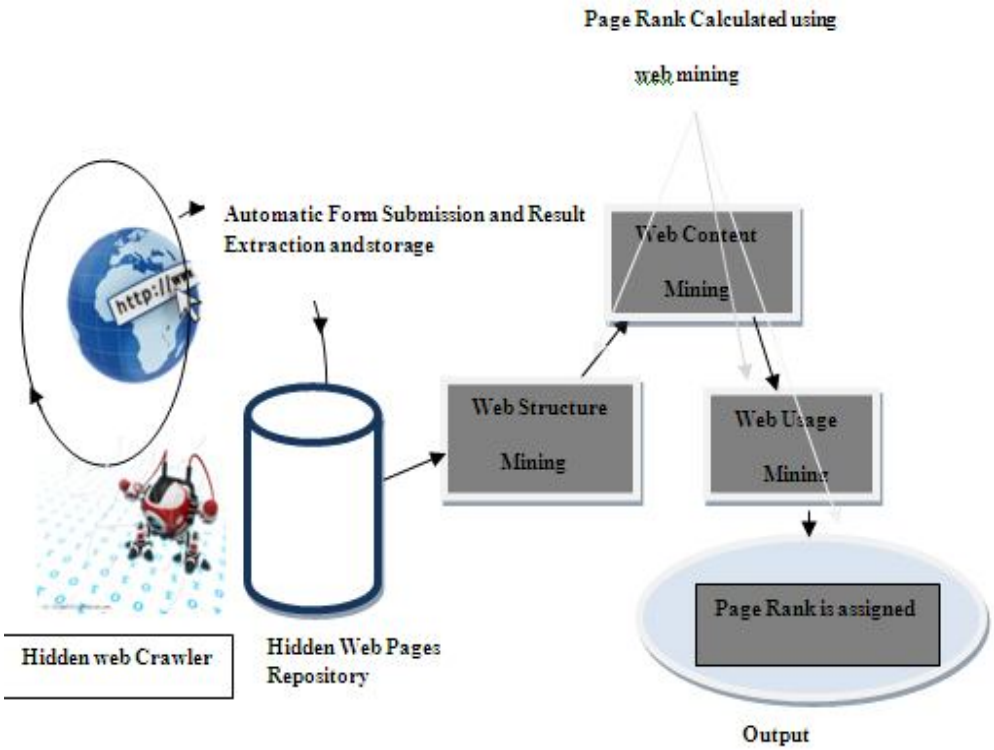

Fig. 3 Proposed Architecture for Rank Calculation

2) Content Page Rank Calculation: The web pages from the WWW are highly connected. If the data about incoming and the outgoing links are stored in relational databases then it becomes difficult and time consuming to extract the knowledge from the relational databases. So in order to remove the problems faced by relational databases in highly connected data we are using the graph database [14]. For experimental purposes NEO4J [15] which is a property graph database is being used. In property graph databases the nodes represent the entities (web pages) and edges represents the relationships (here inlinks and outlinks). The web pages from two domains car and property are taken as shown in Fig. 4. The nodes in the graph are the query interfaces or result pages and the edges are the hyperlinks between web pages as shown in Fig. 5. The graph is created by using NEO4J and cypher query language. 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. Most of the newbie websites gave their advertisements on the high rank popular websites. Mostly the domains of these websites are far apart from each other. The popular websites get financial benefit from it. So they promote the newbie websites. These new websites then are able to share the page rank of the popular websites. So in order to avoid such type of page rank sharing the new technique should be developed which will avoid this incorrect page rank sharing policy. So we here propose new formulae for calculating the page rank. For calculating the rank of page "A" which assume that it receive "n" inlinks from same domain and "m" inlinks from different domains. The pagerank will be shared using the formulae given below.

$$Rank(s)\ A = (Rank(s)_{P1} / (OS_{P1}) + Rank\ (structure)_{Pn} / (OS_{Pn})) \quad \text{if } OS_{P1}>0$$

Here Rank(s) A is structure rank of page A, OSP1 is the no of outlinks of web page P1 to web pages of same domain. So the page that has inlinks from different domains will not be able to share the page rank. The pages which are being linked by pages from same domain will get share of pagerank. So node 700 will share pagerank from 699 and 702. It will not get the share from node 708.The node 699 of domain car will not be able to get the vote from nodes 705,706,707 and 708 which belong to the property domain. Rank(s) 699=0 because OS=0 here.

3) *Usage Page Rank Calculation:* In the beginning the result hidden pages are displayed to user according to the rank calculated above. When user browses the pages, his access pattern will be recorded in the server's log file. The server logs will be processed, users access pattern and the time spend by the user on the web pages will be stored. 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. By the help of the proposed algorithm the rank of the web pages will be calculated on the fly. So the user will be able to view pages of his interest at the top.

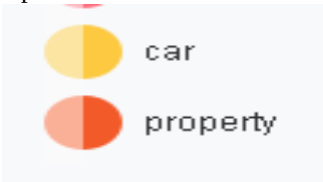

Fig. 4 Nodes of domain car and property

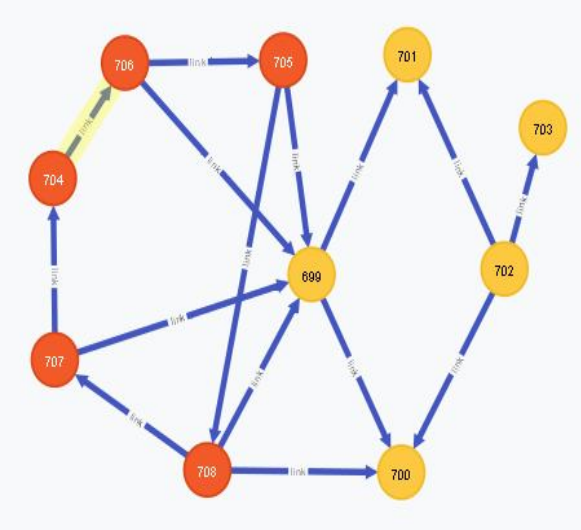

Fig. 5 Interconnection among web pages

$$\text{Rank}_{(u)A} = \text{Access}(\text{User}_{Acn} + \text{User}_{Atn})$$

Here $\text{Rank}_{(u)A}$ is the usage pagerank of page A, User_{Acn} refers to the no of clicks (c) of User_{n} on page A and User_{Atn} refers to the time(t) spend by User_{n} on page A.

4) *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.

$$\text{Rank (f)}_{A} = \text{Rank}_{(c)A} + \text{Rank}_{(s)A} + \text{Rank}_{(u)A}$$

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

B. Advantages of the SCUM algorithm

- The proposed algorithm also considers the domain of web pages casting the vote in the rank calculation.
- It not only considers the links but also the contents of the web pages for the rank calculation.
- The interest of the users is also considered. So every user is able to see the pages of its own interest at the top.
- The final page rank is calculated dynamically.
- The algorithm makes use of all aspects of web mining to calculate the page rank.
- The proposed algorithm will rank the pages from the hidden web and surface both.

REFERENCES

[1] The deep web: Surfacing hidden value. Accessible at http://brightplanet.com, July 2000.
[2] B.He, Z. Zhang, K. C.-C. Chang, Data integration: Knocking the door to the deep Web: integrating Web query interfaces, Proc. SIGMOD'04, Paris, France, June 2004, 913-914.
[3] Babita Ahuja , Dr. Anuradha , Ashish Ahuja,Hidden Web Extraction using Artificial Intelligence (Communicated)
[4] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.

[5] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pag-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACM SIGKDD Explorations Newsletter, January 2000, Volume 1 Issue
[6] Raymond Kosala, Hendrik Blockeel, " Web Mining Research: A Survey ",ACM SIGKDD,Vol 2, Issue 1, pp 1-15
[7] Wang Jicheng, Huang Yuan, Wu Gangshan, Zhang Fuyan. Web mining: knowledge discovery on the Web. Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference- on Volume 2, Page(s):137 - 141 vol.2 - 12-15 Oct. 1999
[8] Alex G. Biichner, Maurice D. Mulveena" Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining", (1998).
[9] S. Brin, and L. Page, The Anatomy of a Large Scale Hypertextual Web Search Engine,, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117,1998.
[10] Wenpu Xing and Ali Ghorbani, Weighted PageRank Algorithm, Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.
[11] J. Kleinberg, Authoritative Sources in a Hyper-Linked Environment, Journal of the ACM 46(5), pp. 604-632, 1999.
[12] Hayes, P.: RDF Semantics. http://www.w3.org/TR/2004/REC-rdf-mt-20040210/ (2004)
[13] SPARQL Query Language for RDF, W3C Working Draft, 12 October 2004, http://www.w3.org/TR/2004/WD-rdf-sparql-query-20041012/.
[14] Darshana Shimpi ,Sangita Chaudhari "An overview of Graph Databases", International Conference in Recent Trends in Information Technology and Computer Science (ICRTITCS - 2012)
[15] Graph Databases NEO4J by Ian Robinson, Jim Webber O'Reilly Publication

