Web Search Result using the Rank Improvement

Rashmi Rani [*], **Vinod Jain** [**]

[*] B.S.Anangpuria. Institute of Technology and Management, Faridabad, India
[**] Department of Computer Science and Engineering B.S.Anangpuria. Institute of Technology and Management, Faridabad, India

Abstract- The problem of finding relevant documents has become much more difficult due to the return a large number of Web pages generally in the form of ranked list data on the WWW. This result increases the users' searching time to find the desired information within the search results, while in general most users just want to result pages to find new/different results. Thus a work is done which reduced a search space and high priority pages are to move upwards in the result list. The Web mining tools are used to classify, cluster and order the documents so that users can easily navigate through the search result and find the desired Information content .The method first performs query clustering in query logs and then capture the weight of clicked web pages in each cluster and also capture the rank of that pages then we find out the new rank by adding the weight and existing rank, Now apply the Insertion Sort and set the level according the high priority

In this paper, architecture is being proposed that introduces methods that order the results according to both the relevancy and the importance of documents. This proposed work results in reduced search space as user intended pages tend to move up words in result list.

Index Terms- WWW; Query log; Cluster; Search Engine; Ranking Algorithm, Insertion Sort

I. Introduction

WWW is one of the popular resource for text, image, audio, video, and metadata .In order to analyze such data, some technique called web mining technique are used by various web application and tools. Web mining describes the use of data mining technique to automatically discover Web documents and services, to extract information from the web resource and to uncover the general pattern on the Web. However, with the overwhelming volume of information on the Web, the task of finding relevant information related to a specific query /topic is becoming increasingly difficult. Many advanced Web searching technique have been recently developed to taken this problem and are being used in the commercial Web search engines such as Google and Yahoo. Google [3] has been found out that more than 50% of the search engine users consult no more than first two screens of results [4]. To get the required information, the user may have to sift through a very large list of documents displayed by search engines, posing the problem of information overkill thus necessitating the need to look for alternative techniques for documents presentation.

R.Cooley et al [4] and Dr. M.H.Dunham [5] divide web mining into three Categories namely web content mining, web structure mining and web usage mining. Web Content Mining emphasis on the content of Web page instead of its embedded links. Web Structure Mining is used to generate structural summary about the Web sites and Web pages. Web Usage Mining tries to discover user navigation patterns from web data and useful information from the interactions of the users while surfing on the web. Nowadays, providing a set of web pages based on query words is not a big problem

In search engines [2] .Instead, the problem is that a search engine returns a large number of web pages in response to user queries and users have to spend much time in finding their desired information from the long list resulting in information overload problem [2] .Almost all search engines store their user activities in the form of query logs . Query logs provide an excellent opportunity for gaining insight into how a search engine is used and what the user's interest are since they form a complete record of what use searched for in a given time frame.

In this paper the work proposed is to optimize the result of a search engine by returning the more relevant and user desired pages on the top of search result list .the paper has the following 4 steps: Section II describe the related work and terminologies used in the work; section III illustrates the proposed optimization technique in detail along with proposed page rank with Insertion sort algorithms. The practical evaluation of the work is given in section IV; section V concludes the paper with some discussion on future research.

II. RELATED WORK

The notation of Web log Mining has been a subject of interest since many years. Most of the search engines use Page ranking algorithms, which can arrange the documents in order of their relevance, importance and content Page ranking algorithms [9, 10] have been proposed in the literature such as Page Rank, Weighted Page Rank. The typical logs [3] of search engine includes the following entries (1) User IDs , (2) Query q issued by the user, (3) URL u selected by the user, (4) Rank r of the URL u clicked for the query q and (5) Time t at which the query has been submitted for search.Table1 for this format.

A number of researchers analyzing the problems of query logs [6, 7, 8, 9].The information contained in query logs has been used in many different ways, for examine to provide context doing search to clustering values. In values studies [9, 10], researcher and search engine operators have used information from query logs to learn about the search process and thus improve search engines.

Thorsten joachnns [7] proposed an approach to automatically optimizing the retrieval quality of search engines using data stored in query logs and the logs of links the users clicked on in the presented ranking.Hao.Maetal. [8] Suggested an accommodation system by extracting the relations between user submitted queries stored in query logs. They developed an effective and entered two-level query suggestion model by mining click though data in the form of two bipartite graphs.

2.1 Page Rank Algorithm

Page Rank [10, 11, 12] was developed at Stanford University by Larry page (cofounder of Google search engine) and Sergey Brin. Google uses this algorithm to order its search results in such a way that important documents move up in the results of a search while moving the less important pages down in its list. This algorithm states that if a page has some important incoming links to it, then its outgoing links to other pages also become important, thus it takes back links into account and propagates the ranking through links. When some query is given, Google combines precompiled Page Rank scores with text matching scores to obtain an overall ranking score for each resulted web page in response to the query. Although many factors determine the ranking of Google search results but Page Rank continues to provide the basis for all of Google's web search tools.

Random Surfer Model [11] which states that not all users follow the direct links on WWW. The modified version is given in (1).

$$PR(u) = (1-d) + d \sum_{V€B(u)} \frac{Pr(v)}{Vn} \qquad (1)$$

TABLE 1 DATA FOR QUERY

UserID	Query	Clicked URL	Time
12345	Marutiswift	www.mauutiswift.com	2012-03-14 00:01:9
12345	Gaadi	www.gaddi.com	2012-03-14 00:04:18
12345	marutidzire	http://www.marutidzire.com	2012-03-14 00:04:18
……	…………	………………..	………….

Where u represents a web page, B (u) is the set of pages that point to u .PR (v) and PR (u) are rank scores of page u and v respectively. Nv denotes the number of outgoing links of page v, d is a damping factor [6] that is usually set to 0.85. d can be thought of as the probability of users‟ following the direct links and (1 − d) as the page rank distribution from non-directly linked pages.

2.1.1 Example illustrating working of PR

To explain the working of Page Rank, let us take an example hyperlinked structure shown in Figure 2, where A, B and C are three web pages. The Page Ranks for pages A, B, C can be calculated using (2) as shown below.

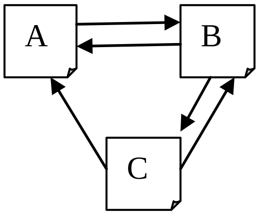
$$PR(A) = (1-d) + d ((PR(B)/2+PR(C)/2) \qquad (a)$$

$$PR(B) = (1-d) + d (PR(A)/1+PR(C)/2) \qquad (b)$$

$$PR(C) = (1-d) + d (PR(B)/2) \qquad (c)$$

By calculating the above equations with d=0.5 (say), the page ranks of pages A, B and C become:

PR (A) =1.2, PR (B) =1.2, PR(C) = 0.8



Example Hyperlinked Structure

2.2 Weighted Page Rank Algorithm

Wenpu Xing and Ali Ghorbani [8] proposed an extension to standard Page Rank called Weighted Page Rank (WPR). It assumes that more popular the web pages are more linkages other web pages tend to have to them or are linked to by them. This algorithm assigns larger rank values to more important pages instead of dividing the rank value of a page evenly among its outgoing linked pages.

The weight of the URL is estimated to be its popularity and order of access corresponding to the user query. Suppose a sequential pattern < {A}, {B}, {C}> belongs to a query cluster matched with the query .This pattern can be graphically Presented as shown in figure

$$Weight(X) = \frac{In (lenpat(x))}{level(x)}$$

Where lenpat(x) is the efficient length/depth of the sequential pattern in which X occur and level (X) is the depth of X in the pattern .Consider the example pattern of figure the weight of pages comes out to be:

Weight (A) = 0.65
Weight (B) = 0.93
Weight(C) = 0.60

III. THE PROPOSED OPTIMIZATION SYSTEM

A novel optimization method based on learning from historical query logs is proposed to predict user’s information in a better way.

The search engine architecture is modified so as to add the component for calculating importance and relevancy of pages. The modified architecture is displayed in figure:

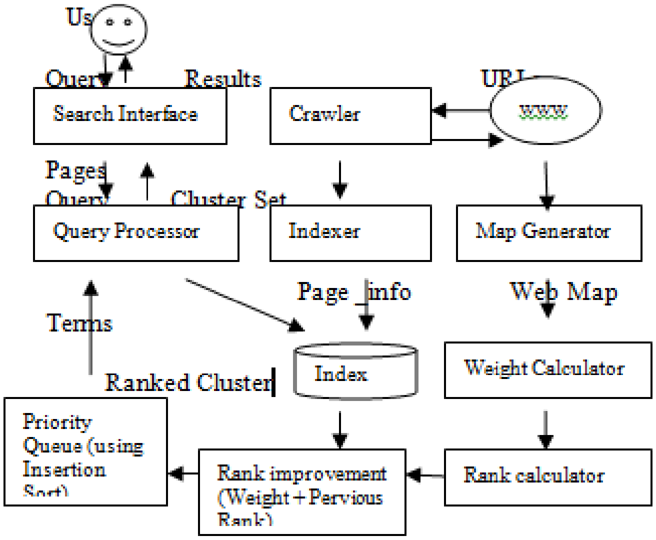


Figure: Priority Based Search Engine Architecture

When a user submit a query on the search engine interface, the query processor component match the query terms with the index repository of the search engine and return a list of matched documents in response .The user browsing behavior including the submitted queries and clicked URLs get stored in the logs .The Rank Updater component works online and takes input the matched documents received by query processor. It improves the ranks of page based on the weights assigned to each according to a sequential pattern which were discovered offline. And a heap sort which gave the first priority according to the new Page Rank.

The working and algorithm for different functional modules are explained below.

a) WWW: The World Wide Web is a system of interlinked hypertext documents accessed via the Internet With a web browser, one can view web pages that may contain text, images, videos, and other multimedia, and navigate between them via hyperlinks

b*) Crawler*: A Web crawler is one type of boot, or software agent, or computer program, it starts with a list of URLs to visit, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier.Webcrawling providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code.

c) *Search Interface*: It is a Graphical interface of a search engine on which the user can enter his query e.g. the Google interface. There are two parts to a search engine's user experience: the user interface (design of the search forms and results pages) and the functionality (how well it matches and sorts pages). When you install a search engine, you should consider both aspects, design the interface carefully and test all aspects of the usability.

d) *Query Processor*: It is the component used to taking the user query from the search interface and processing it word by word.

e) *Indexer*: It collects and stores data to facilitate fast and accurate information retrieval. The index is usually built in an alphabetical order of term and contains extra information regarding the page such as its URL, frequency and position of term. It provides a more useful vocabulary for the internet or onsite search engine. Index design incorporates interdisciplinary concepts from linguistics, cognitive psychology, mathematics, informatics, physics, and computer science. An alternate name for the process in the context of search engines designed to find web pages on the Internet is web indexing.

f). *Map Generator*: This module generates a map/graphical structure of the WWW. The map is used further find out the inlinks and outlinks of the web pages

g) Rank improvement: The rank of a page can be improved with the help of its assigned weight .The new rank now becomes:

New Rank(X) =Rank(X) +Weight(X) (5)

Where rank(X) is the exiting rank value (Page Rank) of page X and weight(X) is the popularity given to X.

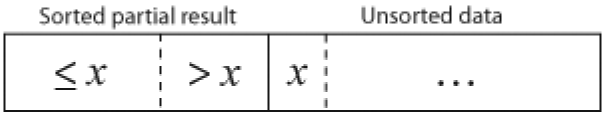
h) *Priority queue*: A priority queue is a collection of elements that each element has been assigned a priority and such that order in which elements are deleted and processed comes from the following riles:

1) An element of higher priority is processed before any element of lower priority.

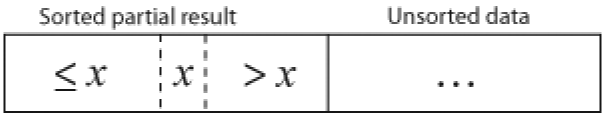
2) Two element with the same priority are processed according to the order in which they were added to the queue.

***Insertion Sort*:** Every repetition of insertion sort removes an element from the input data, inserting it into the correct position in the already-sorted list, until no input elements remain. The choice of which element to remove from the input is arbitrary, and can be made using almost any choice algorithm.

Sorting is typically done in-place. The resulting array after *k* iterations has the property where the first *k* + 1 entries are sorted. In each iteration the first remaining entry of the input is removed, inserted into the result at the correct position, thus extending the result:



becomes



A novel optimization method based on learning from historical query logs is proposed to predict user's information needs in a better way. The final approach is to re_rank the search result list by modifying the already assigned rank score of the web page using the discovered sequential pattern. The rank updating improves the relevancy of the web page which are already visited by some users in the past .By this way , the time user spends for seeking out the required information from search result list can be reduced and the more relevant Web pages can be presented.

The algorithm is based on the simple perspective; initially all queries are considered to be unassigned to any cluster. Each unclassified query is examined against all other queries.

Algorithm: rank improve (Q,n)

Given: A set of n queries and corresponding clicked URLs stored array Q [qi, URL1……., URLm], 1≤i≤n
Output: A set C= {C1, C2…, Ck} of k query.
// Start of algorithm

```
{
K=0;
For (each query P in Q)
Set Clusterid (P) = NULL;
          For (each P€ Q with clustered (P) = NULL)
          {
        I = n, page = Q (n);
        Clusterid (p) =ck;
        $\text{Weight}(X) = \frac{\text{In (lenpar}(X))}{\text{level}(X)}$
          $\text{Page\_ rank}(X) = (1-d) + d \sum_{V€B(X)} \frac{PR(v)}{Nv}$
        NewPage _rank(X) = Page_ rank + Weight(X)

        For (j=2; j≤page; j++)
         {
        NewPage_rank= Q[j];
        i=j-1;
       While ((i≥1) && (item<Q[i]))
        {
       Q [i+1] = Q[i];
        i=i-1;
        }
      Q [i+1] =Newpage_rank;
      }
    }
K=k+1;
}
```

It may be noted from the literature that page ranking algorithms have become more and more efficient in order to achieve higher precision, they have been made to satisfy the needs of all users in a concise manner. The returned documents should be arranged in a user friendly manner. The paper focuses on other techniques like clustering also to represent the results according to the user needs

Rank improvement: This module takes the input from the query processor and matched documents of a user query and an improvement is applied to improve the rank score of the returned pages. The module operates online at the query time and applied the improvement on the current documents.

Step 1: Given an input user query q and matched document D collected from the query processor, the cluster ck is found to which the query q belongs.

Step 2: Sequential pattern of the concerned cluster the retrieved from the local repository maintained by the sequential pattern generator.

Step 3: The level weight are calculated for every page X present in the sequential pattern.

Step 4: The rank are calculated for every page X present in the sequential pattern. The improved is calculated as the summation of pervious rank and assigned weight value.

By improving the ranks using a priority queue, the result of a search engine can be optimization so as to better serve the user need. The user can now find the popular and relevant pages upwards in the result list.

IV. EXPERMENTAL RESULTS

To show the practical evaluation of the proposed architecture, a fragment of sample query log is consider .The following function is tested:

1. Pattern generation
2. Weight Calculation and Rank updation
3. Priority Rank based indexing

1. *Pattern generation*

For each cluster c1 and c2, corresponding URLs form the Query Cluster Database are extracted .Two columns (ClusterID and URL) need to be retrieved. URLs are assigned to different variable

For example,
A= www.marutiswift.com,
B=www.gaadi.com,
C=www.swiftdezire.com

2. *Weight calculation and Rank Improvement*

Weight of each URL can be determined in the following ways:
Weight (A) =1.099, Weight (B) =0.549
Weight(C) =0.549, Weight (A=0.366)
The ranking of search results has been modified to a great extent and the more relevant Web pages can be implemented in this ways.

Table RANK IMPROVEMENT WITH WEIGHT OF PAGES

Page/URL	Rank	Weight	New Rank
A	1.2	0.65	1.85
B	1.2	0.93	2.13
C	0.8	0.60	1.40
…..	………..	….	……

3. Priority Rank Based Indexing

High priority takes the first place in this ways

Page/URL	Priority queue
B	2.13
A	1.85
C	1.40
….	….

V. CONCLUSION AND FUTURE SCOPE

Web mining is used to extract useful information fromUsers'past behavior. In this paper the Page Rank and Weighted Page Rank algorithms are used by many search engine but the users may not get the required relevant documents easily on the top few pages. To solve this problem we use the Weighted Page Content Rank has been proposed which employ Web structure mining as well as Web Content mining technique. This algo is improving the order of the page in the result list so that the user gets the relevant and important pages in the list.

A query log analysis the proposed for implementing effective web search. The most important feature is that the result optimization method is based on users' feedback, which determines the relevance between Web pages and user query words. The returned pages with improved page ranks are directly mapped to the user feedbacks and dictate higher relevance than pages that exist in the result list. Bipartite graph technique can be employed on query logs to retrieve a better clustering of user queries and thus returning more efficient results.

REFERENCES

[1] A.K Sharma, Neha Aggarwal, Neelam Dhuan and Ranjna Gupta,"Web Search Result Optimization by Mining the Search Engine Query logs,". Proceeding IEEE International Conference on methods and models in Computer Science (ICM2CS-2010).

[2] A. Arasu, J.Cho, H. Garcia -Molina, A.Paepcke, and S. Raghavan, "Searching the Web," ACM Transactions on Internet Technology, Vol. 1, No. 1, pp. 97_101, 2001.

[3] A. Spink, D.Wolfram, B.J.Jansen, T.Saracevis,"Searching the Web: The public and their queries".journal of the American Society for information Science and technology 52(3), 2001, 226-234.

[4] R.Cooley, B.Mobasher and J.Srivastava, "Web mining: Information and pattern discovery on the World Wide Web," In 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 97) 1997.

[5] M.H.Dunham, Companion slides for the text," Data mining Introductory and advanced topics".Prentice Hall,2002.

[6] J.Wen, J.Mie, and H.zhang," Clustering user queries of a search engine".In Proc of 10th International WWW Conference .W3C, 2001.

[7] Thorsten joachims,"optimizing search engine using click though data" Proceeding of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, 2002, pp: 133-142, New York.

[8] H.Ma, H.Yang,I.King,and M.R.Lyu,"learning latent semantic relations from click though data from query suggestion ".InCIKM'08:Proceeding oh the 17th ACM conference on information and knowledge management ,pages 709-708,New York,ny,USA,2008,ACM.

[9] Isak Taka, Sarah Zelikovitz, Amanda Spink, "Web Search log to Identify Query Classification terms "Proceeding of IEEE International Conference on Information Technology (ITNG'07), pp: 50-57, 2008.

[10] L. Page, S.Brin, R.Motwani, T.Winograd,"The page rank citration ranking: Bringing order to the Web".Technical report, Stanford Digital Libraries SIDL-WP-1990-0120, 1999.

[11] C. RIDINGS AND M.Shishigin,"Pagerank uncovered".Technical report, 2002.

[12]]http://pr.efactory.de/e-pagerank-algorithm.shtml.

[13] Neelam Duhan, A.K Sharma, "A Novel Approach for Organing Web Search Results using Ranking and Clustering," International Journal of Computer Applications (0975-8887) Volume 5-No.10, August 2010.

AUTHORS

First Author – Rashmi Rani, B.S.Anangpuria. Institute of Technology and Management, Faridabad, India
Second Author – Vinod Jain, Department of Computer Science and Engineering B.S.Anangpuria. Institute of Technology and Management, Faridabad, India