World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:8, No:8, 2014

Comparative Analysis of Different Page Ranking Algorithms

S. Prabha, K. Duraiswamy, J. Indhumathi

Open Science Index, Computer and Information Engineering Vol:8, No:8, 2014 publications.waset.org/10000128/pdf

***Abstract*—**Search engine plays an important role in internet, to retrieve the relevant documents among the huge number of web pages. However, it retrieves more number of documents, which are all relevant to your search topics. To retrieve the most meaningful documents related to search topics, ranking algorithm is used in information retrieval technique. One of the issues in data miming is ranking the retrieved document. In information retrieval the ranking is one of the practical problems. This paper includes various Page Ranking algorithms, page segmentation algorithms and compares those algorithms used for Information Retrieval. Diverse Page Rank based algorithms like Page Rank (PR), Weighted Page Rank (WPR), Weight Page Content Rank (WPCR), Hyperlink Induced Topic Selection (HITS), Distance Rank, Eigen Rumor, Distance Rank Time Rank, Tag Rank, Relational Based Page Rank and Query Dependent Ranking algorithms are discussed and compared.

***Keywords*—**Information Retrieval, Web Page Ranking, search engine, web mining, page segmentations.

I. Introduction

DATA mining is to extract or mine knowledge from a lot of data called Knowledge Discovery in Databases (KDD), which is the result of information technology natural which is the result of information technology natural evolution. In recent years, the data mining technology produced great attention among the information industry, which is developing rapidly. Data mining is an inter-discipline subject, influenced by multiple disciplines, including database system, statistics, machine learning, data analysis, etc. At present, according to the different types of mining knowledge and mining the different objects, many data mining methods and special tools are available. Many research fields such as database, data analysis, machine learning, also benefited a lot from the data mining. Information Retrieval is a technique used in Data Mining for searching in huge databases to retrieve related documents. Information Retrieval (IR) is the science of searching for information within relational databases, documents, text, multimedia files, and the World Wide Web. Many users are affianced in the IR field especially reference librarians, governmental agents, professional researchers, political analysts, and market forecasters. The applications of IR are diverse but not limited to extraction of information from huge documents, spam filtering, probing in digital libraries, information filtering, object extraction from images, automatic summarization, document classification and clustering, and web searching. Google's PageRank algorithm is the one of the best-known algorithms in web search. With the increasing number of Web pages [42] and users on the Web, the number of queries submitted to the search engines are also growing rapidly day by day. Therefore, the search engines needs to be more efficient in its processing way and its output. Web mining techniques are employed by the search engines to extract appropriate documents from the web database documents and provide the necessary and required information to the manipulators. The search engines become very successful and popular if they use efficient ranking mechanisms. Now these days it is very successful because of its PageRank algorithm. Page ranking algorithms [32] are used by the search engines to present the search results by considering the significance, reputation, and content score [30] and web mining techniques to order them according to the user interest. Some ranking algorithms [37] depend only on the link structure of the documents i.e. their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining), while some use a permutation of both i.e. they use content of the document as well as the link structure to assign a rank value for a certain document [5]. If the search results are not displayed according to the user interest then the search engine will mislay its fame. So the ranking algorithms become very important. The sample architecture of a search engine is shown in Fig. 1.

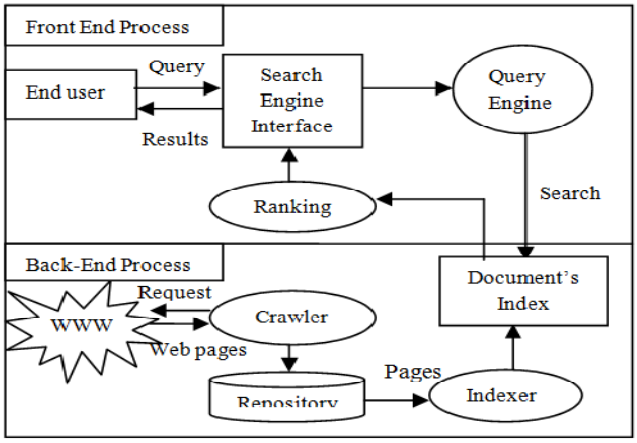


Fig. 1 Architecture of search engine

Ms.S. Prabha, Associate Professor, is with the Department of Information Technology, K.S.Rangasamy College of Technology, Tamil Nadu India (e-mail: prabha.dw@gmail.com).

Dr. K. Duraiswamy, Academic Dean and Professor, is with the Department of computer science, K. S. Rangasamy College of Technology, Tamil Nadu India.

J. Indhumathi, PG Scholar, is with the Department of Information Technology, K. S.Rangasamy College of Technology, Tamil Nadu India (e-mail: indhuksrct@gmail.com).

Open Science Index, Computer and Information Engineering Vol:8, No:8, 2014 publications.waset.org/10000128/pdf

There are three vital components in a search engine known as Crawler, Indexer and Ranking mechanism. The Crawler is also called as a robot or spider that navigates the web and downloads the web pages. The downloaded pages are being transferred to an indexing module that parses the web pages and erect the index based on the keywords in individual pages. An alphabetical index is normally sustaining using the keywords. When a query is being drifted by a user, it means the query transferred in terms of keywords on the interface of a search engine, the query mainframe section examine the query keywords with the index and precedes the URLs of the pages to the user. But before presenting the pages to the client, a ranking mechanism is completed by the search engines to present the most relevant pages at the top and less significant ones at the substructure. It makes the search outcomes routing easier for the user.

II. Ranking Technique Overview

Web mining [1] is the mechanism to classify the web pages and internet users by taking into consideration the contents of the page and behavior of internet user in the past. An application of data mining technique is a web mining [36] which is used spontaneously to find and retrieve information from the World Wide Web (WWW). According to analysis targets, web mining [51] is made of three basic branches i.e. web content mining (WCM), web structure mining (WSM) and web usage mining (WUM).

A. Web Content Mining (WCM)

Web Content Mining [18] is the progression of extracting useful information from the contents of web credentials. The web credentials may consists of text, audio, video, image or structured records like tables and lists. Mining can be purposeful on the web documents as well the results pages fashioned from a search engine. There are bi approaches in content mining called agent based approach and database based approach. The agent based method focus on searching appropriate information using the uniqueness of a particular domain to interpret and organize the collected information. The database approach is used for get back the semi-structure data from the web.

B. Web Usage Mining (WUM)

Web Usage Mining is the method of hauling out useful information from the secondary data consequent from the interactions of the user while surfing on the Web. It extracts data accumulated in server referrer logs, access logs, agent logs, user profile and Meta data client-side cookies.

C. Web Structure Mining (WSM)

The aim of the Web Structure Mining [17] is to generate the structural abstract about the Web site and Web page. It tries to determine the link structure of the hyperlinks at the bury document level. Basic underpinning on the topology of the hyperlinks, Web Structure mining [34] will classify the Web pages and spawn the information like similarity and relationship between different Web sites. This type of mining can be carried out at the document level (intra-page) or at the hyperlink level (inter-page). It is important to appreciate the Web data structure for Information Retrieval. The three categories of web mining [47] described and its own appliance areas including site improvement, business intelligence, web personalization, site modification, usage characterization and ranking of pages ,classification etc.

III. Different Page Ranking Algorithms

The page ranking algorithms [38] are generally used by search engines to find more important pages. Different Page Rank based algorithms [43] like Page Rank (PR), Weighted Page Rank (WPR), Weight Page Content Rank (WPCR), Hyperlink Induced Topic Selection (HITS), Distance Rank, Eigen Rumor, Distance Rank Time Rank, Tag Rank, Relational Based Page Rank and Query Dependent Ranking algorithms.

A. Page Ranking Algorithms

In the ranking algorithms [4] the usage of web is drastically increases day by day. The search engine is very useful to retrieve the relevant documents from web easily. The ranking algorithms [12] are very important because the search results are not according to their user needs then the search engine loss their popularity. Google is the famous search engine tool in page ranking algorithm. Some ranking algorithms depend only on the popularity score i.e. web structure mining and web content mining. The PageRank values are calculated based on the number of pages that point to a page. Adaptive Methods for the Computation of PageRank algorithm [2], [10] is used to speed up the computation of PageRank is nearly 30%. Filter-Based Adaptive PageRank and Modified Adaptive PageRank algorithm [50] is used to reducing the redundant computation. The page rank consider only back link to decide the page score. In the below equation the variable d is a damping factor [20], [31] values can be set between 0 and 1. PR(A) is the PageRank of page A,T1….Tn is all pages that link to page A, PR(Ti) is the PageRank of page Ti, Q(Ti) is the number of pages to which Ti links to PR(Ti)/Q(Ti) is PageRank of Ti distributing to all pages that Ti links to,(1-d) is to make up for some pages that do not have any out-links to avoid losing some page ranks [40].PR(A)is the incoming link to page A and C(T1) is the outgoing link from page PR(T1). The PageRank of a page A is given below

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

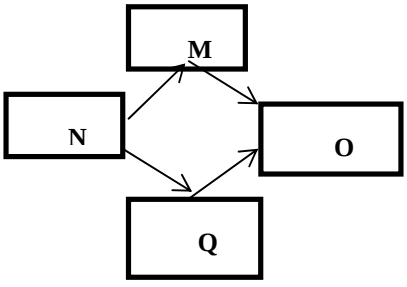


Fig. 2 Back links Example

Open Science Index, Computer and Information Engineering Vol:8, No:8, 2014 publications.waset.org/10000128/pdf

An example of back link is shown in Fig. 2, N is the back link of M & Q and M & Q are the back links of O.

Implementation of Page Rank Algorithm

The following steps explain the method for implementing Page Rank Algorithm [33]:

Step 1. Initialize the rank value of each page by 1/n. where n is total no. of pages to be ranked. Suppose we represent these n pages by an Array of n elements. Then A[i] = 1/n where 0≤ i< n

Step 2. Take some value of damping factor such that 0<d<1. e.g. 0.15, 0.85 etc.

Step 3. Repeat for each node i such that 0≤ i< n. Let PR be an Array of n element which represent PageRank for each web page.

PR[i]← 1-d

For all pages Q such that Q Links to PR[i] do

PR[i] ← PR[i] + d * A[Q]/Qn

where Qn = no. of outgoing edges of Q

Step 4. Update the values of A

A[i]= PR[i] for 0≤ i< n

Repeat from step 3 until the rank value converges i.e. values of two consecutive iterations match.

The advantages of page rank are less query time, Less susceptibility to localized links, more efficiency and feasibility.

B. Weighted Page Rank

Weighted Page Rank algorithm (WPR) [14] is the extension of Page Rank algorithm. In the WPR contains both in link and out link, the link is assigned based upon the page rank priority. The terms [48] of weight values to the incoming and outgoing links and are denoted as $w^{in}(m,n)$ and $w^{out}(m,n)$ respectively, $w^{in}(m,n)$ is the weight of link (m,n) calculated based on the number of in links of page n and the number of in links of all reference pages of page m.

$$W^{in}(m,n) = \frac{I_n}{\sum P \varepsilon R(m) I_P} \quad (2)$$

where In and Ip represent the number of in links of pages n and page p, respectively. R(m) denotes the reference page list of page m. $w^{out}(m,n)$ is the weight of link(m,n) calculated based on the number of out links of page n and the number of out links of all reference page of page m.

$$W^{out}(m,n) = \frac{O_n}{\sum P \varepsilon R(m) O_n} \quad (3)$$

where O_n and O_p represent the number of outlinks of the page n and page p, respectively. R(m) denotes the reference page list of page m. Modification of page rank formula is given:

$$WPR(n) = (1-d) + d \sum_{m \epsilon B} WPR(m)\, w^{in}(m,n) w^{out}(m,n) \quad (4)$$

The values of *WPR*(*A*), *WPR*(*B*), *WPR*(*C*) and *WPR*(*D*) are shown in equations respectively. The relation between these are *WPR*(*A*)>*WPR*(*B*)>*WPR*(*D*)>*WPR*(*C*).

This results shows that the Weighted PageRank order is different from *PageRank*.

C. Weighted Page Content Rank

Weighted page content algorithm (WPCA) [46] is the modification of original page rank algorithm. It is used to give the sorted order to the web page. WPCR [35] is assigned numerical value based on which the web pages are given an order. This algorithm employs both web structure mining and web content mining techniques. Web structure mining is used to analyse the popularity of the page and web content mining is used to find the page relevancy. The calculation is based on the in links and out links of the page [47]. For example Google Web search receive 34,000 queries per second (2 million per minute; 121 million per hour; 3 billion per day; 88 billion per month) for most queries, there exist thousands of documents containing some or all of the terms in the query.

Algorithm WPCR

Input:
Query text Q
Set of pages {Pi} →→Google (Q)
Output:
New (Pi)
Relevance calculation
Find f(Pi) = {number of frequency of logical combination of Q}
Find content weight factor CWF(Pi)= GPA(f(Pi))
Reorder and return the new {Pi}

The proposed algorithm SWPCR design new methods to calculate the relevance of a page based on two factors:

1) Find f(Pi) = the frequency of logical combination of query text, the number of times that term appears in page Pi.
2) Find content weight factor CWF(Pi)) = GPA (f(Pi)) that is consider the core of SWPCR proposed algorithm based on: Given a matrix with m × n; n = number of words in a given query, each column contains the frequency of n words f(n) in each of the given pages; m = number of pages

TABLE I
SORTING ROWS IN ARRAY

P1	f(n)	f(n-1)	f(n-2)	……..	f(1)
P2	f(n)	f(n-1)	f(n-2)	……..	f(1)
P3	f(n)	f(n-1)	f(n-2)	……..	f(1)
P4	f(n)	f(n-1)	f(n-2)	……..	f(1)
…	f(n)	f(n-1)	f(n-2)	……..	f(1)
Pm	f(n)	f(n-1)	f(n-2)	……..	f(1)

D. HITS Algorithm

HITS (Hyper-link Induced Topic Search) algorithm [6] is used to ranks the web page by processing in links and out links. In this algorithm [9] a web page is named as authority and hub, if the web page is pointed by many hyperlinks it is named as authority, and if the page is pointed to various hyperlinks and a web page is named as HUB. Hubs are the

Open Science Index, Computer and Information Engineering Vol:8, No:8, 2014 publications.waset.org/10000128/pdf

pages that act as resource lists. Authorities are the pages having main contents. A decent hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is pointed by many good hub pages on the same content. A page may be a good hub and a good authority at the same time. It uses an iterative algorithm for computing the hub and authority weights. The HITS algorithm [49] gives WWW as directed graph G(V,E), where V is a set of vertices representing pages and E is set of edges corresponds to link. It has two steps first one is sampling step and second one is iterative step. Sampling Step a set of relevant pages for the given query are collected, iterative step Hubs and Authorities are found using the output of sampling step. Following expressions are used to calculate the weight of Hub (Hp) and the weight of Authority (Ap).

$$HUB(H_P) = \sum_{q \varepsilon I_P} A_q \quad (5)$$

$$AUTHORITY(A_P) = \sum_{q \epsilon B_P} Hq \quad (6)$$

Here Hub Score of a page is (Hq) and authority score of page is (Aq). I(p) is set of reference pages of page p and B(p) is set of referrer pages of page p. The weight of authority pages is proportional to the weights of hub pages that link to the authority page. Another one is, hub weight of the page is proportional to the weights of authority pages that hub links.

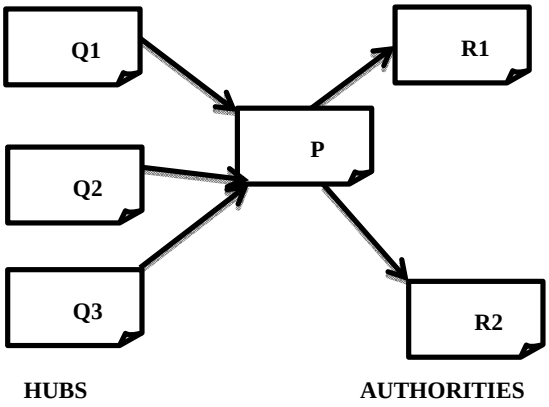


Fig. 3 Hubs and authorities

Hubs and authorities are calculated by way of:

$$A_{P=}H_{Q1} + H_{Q2} + H_{Q3} \quad (7)$$

$$H_P = A_{R1} + A_{R2} \quad (8)$$

Advantage of HITS:

HITS [11] scores due to its ability to rank pages according to the query string, resulting in relevant authority and hub pages. The ranking may also be combined with other information retrieval based rankings. HITS [39] is sensitive to user query (as compared to PageRank). Important pages are obtained on basis of calculated authority and hubs value. HITS is a general algorithm [45] for calculating authority and hubs in order to rank the retrieved data. HITS induces Web graph by finding set of pages with a search on a given query string. Results demonstrate that HITS calculates authority nodes and hubness correctly.

E. Distance Rank Algorithm

A distance rank algorithm [22] is proposed by Ali Mohammad Zareh Bidoki and Nasser Yazdani. This algorithm is also called as intelligent ranking algorithm based on support learning algorithm. In this algorithm [25] the page is considered as a distance factor and rank the distance between web pages in search engine. The main goal of the ranking algorithm [8] is computed on the basis of shortest logarithmic distance between two pages and ranked according to them. The higher rank is assigned to which page has a smaller distance. The Advantage of this algorithm is being fewer sensitive, it can find pages faster with high quality and more quickly with the use of distance based solution as compared to other algorithms. Distance Rank algorithm implements the PageRank properties. Then, a page has a high rank value if it has more incoming links on a page. Formulas for distance rank algorithms

$$Distance\ n[j] = (1-a) * Dist_{n-1}[j] + a * \min i\,(a * Distance_{n-1}[i] + \log(0[i])), i \epsilon B(j) \quad (9)$$

B(j) shows list of pages that link to j and O(i) is the number of out links in page i Sort url_queue by distance vector in ascending order.

F. EigenRumor Algorithm

The EigenRumor algorithm [19] is proposed by Ko Fujimura that ranks each blog entry on basis of weighting the hub and authority scores of the bloggers it is based on eigenvector calculations. So this algorithm enables a high score to be assigned to a blog entry entered by a good blogger but not linked to by any other blogs .In the recent scenario nowadays number of blogging sites is increasing, there is a challenge for web service provider to provide good blogs to the users. Page rank and HITS are providing the rank value to the blogs but some issues arise, if these two algorithms are applied directly to the blogs. The EigenRumor algorithm calculates three vectors, i.e., authority vector a , hub vector h, and reputation vector r, from information provisioning matrix P and information evaluation matrix E.

These issues are:

1. The number of links to a blog entry is normally very small. As the result, the scores of blog entries are calculated by PageRank.
2. The rank scores of blog entries as decided by the page rank algorithm is often very low, so it cannot allow blog entries to be provided by rank score according to their importance. So resolve these issues, an Eigen Rumor algorithm is proposed for ranking the blogs. The Eigen Rumor algorithm has connections to PageRank and HITS in that all are based on eigenvector calculation of the adjacency matrix of the links. One important thing is an agent is used to represent an aspect of human being such as a blogger, and an object is used to represent any object

Open Science Index, Computer and Information Engineering Vol:8, No:8, 2014 publications.waset.org/10000128/pdf

such as a blog entity. Using the Eigen Rumor algorithm, the hub and authority scores are calculated as bloggers and the encouragement of a blog entity that does not yet have any in-link entered by the blogger can be computed.

G. Time Rank Algorithm

Time Rank algorithm is used to improving the rank score by using the visit time of the web page is proposed by H Jiang et al. [23], [41] measured the visit time of the page after applying original and improved methods of web page rank algorithm [15] to know about the degree of importance to the users. This algorithm consumes the time factor to increase the accuracy of the web page ranking. Due to the methodology used in this algorithm, it can be assumed to be a combination of content and link structure [3]. The results of this algorithm are very satisfactory and in agreement with the applied theory for developing the algorithm.

$$P_r(T(i)|q) = P_r\big(T(i)\big) * P_r(q|T(i)) \tag{10}$$

Ti $\rightarrow$ is the topic i of each page.

P_r (T(i)) $\rightarrow\rightarrow$ means the proportion of pages related to topic i in the whole pages set.

P_r (Ti | q) $\rightarrow\rightarrow$ means the probability of the query q belonging to topic i.

The topic sensitive page rank used in the Time Rank is given by:

$$TSPR\ t(i) = a\sum_{i\epsilon B}\frac{TSPR(i)}{|Fi|} + (1-a).Et(i) \tag{11}$$

Single jump probability 1/n is replaced by E_t = {E(1), E(2).....E(n)}, n is the no. of topics.

$$E(i) = \begin{cases} 1/nt \\ 0 \end{cases}$$

$n_t \rightarrow$ number of pages related to topic.

There are n TSPR scores corresponding to topics. It is calculated statically offline. After some running time of search engines the time vector related to topics for every page can be calculated and hence every page is assigned as page rank [21] based on time visited.

$$TIME\ PRt(j) = TSPRt(j) * T(t) \tag{12}$$

where time vector Tv = {T(1), T(2), ...T(n)}

T(i) - user's total visiting time of a page related to topic i. Time rank means that irrespective of similarity of similar link structure of two web pages, the page having longer visited time gets the high score.

H. TagRank Algorithm

Tag rank algorithm [24] is also known as novel algorithm it is used for ranking the web page based on social annotations is proposed by Shen Jie, Chen, Zhang Hui, Sun Rong-Shuang, Zhu Yan and He Kun. This algorithm calculates the tags by using time factor of the new data source tag and the annotations behavior of the web users. This algorithm [7] provides a better authentication method for ranking the web pages. This algorithm provides very accurate results and this algorithm indexes new information resources in a better way. Future work in this direction can be to utilize co-occurrence factor of the tag to determine weight of the tag and this algorithm can also be improved by using semantic relationship among the co-occurrence tags.

I. Relation Based Algorithm

Fabrizio Lamberti, Andrea Sanna and Claudio Demartini proposed a relation based algorithm [27] for ranking the web page for semantic web search engine. Various search engines are presented for better information extraction by using relations of the semantic web. This algorithm [16] proposes a relation based page rank algorithm for semantic web search engine that depends on information extracted from the queries of the users and annotated resources. [26] Results are very encouraging on the parameter of time complexity and accuracy. Further improvement in this algorithm can be increased the use of scalability into future semantic web repositories.

J. Query Dependent Ranking Algorithm

Lian-Wang Lee, Jung- Yi Jiang, ChunDer Wu and Shie-Jue Lee [28] have presented a query dependent raking algorithm for search engine. In this approach a simple similarity measure algorithm is used to measure the similarities between the queries. A single model for ranking is made for every training query with corresponding document. Whenever a query arises, then documents are extracted and ranked depending on the rank scores calculated by the ranking model. The ranking model in this algorithm [29] is the combination of various models of the similar training queries. Experimental results show that query dependent ranking algorithm is better than other algorithms.

IV. Comparison of Page Ranking Methods

On the basis of analysis [44], a comparison of various page ranking algorithms is done on the basis of some vaults such as main technique use, methodology, key in parameter, complexity, relevancy, quality of results, and limitations. On the basis of parameters we can find the powers and limitations of each algorithm.

Open Science Index, Computer and Information Engineering Vol:8, No:8, 2014 publications.waset.org/10000128/pdf

TABLE II
COMPARISON OF VARIOUS PAGE RANKING ALGORITHMS

Algorithms/ Criteria	Main Technique	Methodology	Input parameters	Complexity	Relavancy	Quality of results	Limitations
Page Rank	Web Structure Mining	This algorithm computes the score for pages at the time of indexing of the pages.	Back Links	O(log n)	Less (this algo. Rank the pages on the indexing time)	Medium	Results come at the time of indexing and not at the query time.
Weighted Page Rank	Web Structure Mining	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided.	Back links and Forward links.	<O(log n)	Less as ranking is based on the calculation of weight of the web page at the time of indexing.	Higher than PR	Relevancy is ignored.
Weighted Page Content Rank	Web Structure Mining, Web Content Mining	Gives sorted order to the web page returned by search engine as a numerical value in response for a new query.	Content, Back links and Forward links	<O(log n)	more (it consider the relative position of the pages)	medium	WPCR is a numerical value based on which the web pages are given an order.
HITS	Web Structure Mining, Web Content Mining	It computes the hubs and authority of the relevant pages. It relevant as well as important page as the result.	Content, Back links and Forward links	-	More (this algo. Uses the hyperlinks so according the Hen zinger, 2001 it will give good results and also consider the Content of the page)	Less than PR	Topic drift and Efficiency problem.
Distance Rank	Web Structure Mining	Based on reinforcement learning which consider the logarithmic distance between the pages.	Forward links	<O(log n)	Moderate due to the use of the hyperlinks.	High	If new page inserted between two pages then the crawler should perform a large calculation to calculate the distance vector.
Eigen Rumor	Web Structure Mining	Eigen rumor use the adjacency matrix, which is constructed from agent to object link not page to page link.	Agent/Object	O(log n)	High for Blog so it is mainly used for blog ranking.	Higher than PR and HITS	It is most specifically used for blog ranking not for web page ranking as other ranking like page rank, HITS.
Time Rank	Web Usages Mining	In this algorithm the visiting time is added to the computational score of the original page rank of that page.	Original Page Rank and Sever Log	O(log n)	High due to the updation of the original rank according to the visitor time.	Moderate	Important pages are ignored because it increases the rank of those web pages which are opened for long time.
Tag Rank	Web Content Mining	Visitor time is used for ranking. Use of sequential clicking for sequence vector calculation with the uses of random surfing model.	Popular tags and related bookmarks	-	Less as it uses the keyword entered by the user and match with the page title.	less	It is comparison based approach so it requires more site as input.
Relational Based Page Rank	Web Structure Mining	A semantic search engine would take into account keywords and would return page only if both keywords are present within the page and they are related to the associated concept as described in to the relational note associated with each page.	Keywords	-	High as it is keyword based algorithm so it only returns the result if the keyword entered by the user match with the page.	high	In this ranking algorithm every page is to be annotated with respect to some ontology, which is the very tough task.
Query Dependent Ranking	Web content Mining	This paper proposed the construction of the rank model by combining the results of similar type queries.	Training query	O(n log n)	High (because the model is constructed from the training quires).	high	Limited number of characteristics is used to calculate the similarity.

V. PAGE SEGMENTATION ALGORITHMS

Page segmentation algorithm is to partition web pages into blocks. Because of the special characteristics of web pages comparable several topics and varying length of the web page, dissimilar page segmentation methods have impact on the web search performance. There are four main types of methods of page segmentation algorithms. They are

Open Science Index, Computer and Information Engineering Vol:8, No:8, 2014 publications.waset.org/10000128/pdf

- Fixed-length Page segmentation
- DOM-based page segmentation
- Vision-based page segmentation
- Combined approach segmentation

A. Fixed –Length Page Segmentation (Fixed PS):

A fixed length PS passage contains stable number of words. For web documents, fixed-length page segmentation is identical to traditional window approach, except that all the attributes and HTML tags are removed. The only parameter is the length of window and from the previous experience; it can be of 200 or 250.

B. DOM-Based Page Segmentation (DOMPS):

DOM provides each web page with a fine-grained structure, which explains not only the content but also the presentation of the page. In general, similar to discourse passages, the blocks produced by DOM-based methods tend to partition pages based on their pre-defined syntactic structure, i.e., the HTML tags.

C. Combined Approach (Comb PS)

It takes the advantage of both the visual layout and length normalization. In this, a web page will be first passed to VIPS for segmentation, and then to a normalization procedure. Passage Retrieval: It helps to apply retrieval algorithms to portions of a document, especially when documents have varying length. In passage retrieval, passages can be of three classes. Discourse passages – rely on the logical structure of the documents marked by punctuation. Window passages – It contains fixed number of words. Semantic passages – It is obtained by partitioning a document into topics or sub topics according to its semantic structure. Vision-Based Page Segmentation (VIPS)

People view a web page through a web browser and get a 2-D presentation which provides various visual signs to help distinguish different parts of the page, such as images, lines, etc.. For the sake of easy browsing; a block within the web page is much likely about a single semantic. VIPS [13] is proposed to achieve a more accurate content structure on the semantic level.

D. X-Y Cut Segmentation Algorithm:

The x-y cut segmentation algorithm, it also referred to as recursive x-y cuts (RXYC) algorithm, and is a tree-based top-down algorithm. The root of the tree denotes the total document page. All the leaf nodes together represent the final segmentation. The RXYC algorithm recursively splits the document into two or more smaller rectangular zones which denote the nodes of the tree. At each step of recursion, the vertical and horizontal projection profiles of each node are computed.

E. Voronoi – Diagram Based Algorithm:

The Voronoi-diagram based segmentation algorithm by Kise et al. is also a bottom-up algorithm. The first step is it extracts sample points from the boundaries of the connected components using sampling rate rs. Then the noise removal is done using a maximum noise zone size threshold tn; in addition to height, width and aspect ratio thresholds. After that a Voronoi diagram is generated using sample points obtained from the borders of the connected components. The Voronoi edges that pass through a connected component are deleted to obtain an area Voronoi diagram. Finally, extra Voronoi edges are deleted to obtain boundaries of document components.

VI. Comparison of Various Page Segmentation Algorithms

On the basis of analysis, a comparison of various page segmentation algorithms is done on the basis of some vaults such as basic criteria, main techniques, merits, demerits and retrieval performance.

TABLE III
COMPARISON OF VARIOUS PAGE SEGMENTATION ALGORITHMS

Algorithms	Basic Criteria	Main Technique	Merits	Demerits	Retrieval Performance
Fixed-Length Page Segmentation Algorithm	Fixed number of words or fixed length passages	Web Content Mining	Simplicity, Very robust, Effective For improving performance	No semantic information is taken into account in the segmentation process	Medium
DOM-Based Page Segmentation Algorithm	Tags or tag types and also Content and link	Web Content Mining	Provides a hierarchical structure of every web page	Difficult to evaluate and compare	Less
Vision Based Page Segmentation Algorithm	Visual cues	Web Content Mining	Achieve more accurate content structure on the semantic level. Greatly improve the performance of pseudo relevance feedback	Suffer from lack of normalization. It remains unclear	Medium
Combined Approach Segmentation Algorithm	Visual layout and fixed length	Web Content Mining	Advantage of both Visual layout and length normalization	Little time consuming	More
X-Y Cut	Top-down approach	Web Content Mining	It is Fast and easy to implement	Presence of noise Under segmentation errors	Medium
Voronoi Diagram Based Algorithm	Bottom up approach	Web Content Mining	Voronoi diagram based algorithm is good in Layouts having different variations	Spacing variations can occur. Over segmentation errors is introduced	Medium

Open Science Index, Computer and Information Engineering Vol:8, No:8, 2014 publications.waset.org/10000128/pdf

VII. Conclusion

The algorithms that are described above are effective in retrieving the web pages from the search engines. The link analysis algorithms are based on link structure of the documents. The page which has many links has many references can improves retrieval efficiency. In the integrated ranking approach comes under personalized web search. In integrated approach both the content and the link are integrated to improve the retrieval efficiency. Page Segmentation algorithms are used to segment the page as blocks and by separating as blocks the retrieval performance in the web context could be improved. Each and every algorithm has got its own merits and demerits. As per the requirements of a search engine we can utilize the above said algorithms. It helps to enhance the current page rank algorithm used by the Google and these web page ranking algorithms could be used by several other search engines to improve the retrieval efficiency of the web pages as per the user's query.

References

[1] Cooley, R, Mobasher, B., Srivastava, J."Web Mining: Information and pattern discovery on the World Wide Web". In proceedings of the 9th IEEE International Conference on tools with Artificial Intelligence (ICTAI' 97).Newposrt Beach,CA 1997.

[2] Serge Abiteboul and Victor Vianu, Queries and Computation on the Web. Proceedings of the International Conference on Database Theory. Delphi, Greece 1997.

[3] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, R. Kumar, P.Raghavan, S. Rajagopalan, A. Tomkins, "Mining the Link Structure of the World Wide Web", IEEE Computer Society Press, Vol 32, Issue 8 pp. 60 – 67, 1999.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing order to the Web". Technical Report, Stanford Digital Libraries SIDL-WP 1999-0120,1999.

[5] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson,and J. Kleinberg, "Mining the Web's Link Structure", Computer, 32(8), PP.60–67, 1999.

[6] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "Link Analysis: Hubs and Authorities on the World". Technical Report: 47847, 2001.

[7] Yang, Y. and Zhang, H., "HTML Page Analysis Based On Visual Cues", In 6th International Conference on Document Analysis and Recognition (ICDAR 2001), Seattle, Washington, USA, 2001.

[8] Sung Jin Kim and Sang Ho Lee, "An Improved Computation of the PageRank Algorithm", In proceedings of the European Conference on Information Retrieval (ECIR), 2002.

[9] C.. H. Q. Ding, X. He, P. Husbands, H. Zha and H. D. Simon, "PageRank: HITS and a Unified Framework for Link Analysis". 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002.

[10] C. Ridings and M. Shishigin, "PageRank Convered". Technical Report, 2002.

[11] Longzhuang Li, Yi Shang, and Wei Zhang, "Improvement of HITS-based Algorithms on Web Documents", WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA. ACM 1-58113-449-5/02/0005.

[12] C.P.Lee, G.H.Golub, S.A.Zenios, A fast two-stage algorithm for computing PageRank, Technical report of Stanford University, 2003.

[13] D.Cai, S.Yu, J.-R.Wen, and W.-Y.Ma, "VIPS: a vision-Based page segmentation algorithm", Microsoft Technical Report, MSR-TR-2003-79, 2003.

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

[15] Amy N. Langville and Carl D. Meyer, Deeper Inside PageRank, October 20, 2004.

[16] Ricardo Baeza-Yates and Emilio Davis ,"Web page ranking using link attributes" , In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, PP.328-329,2004.

[17] M. G. da Gomes Jr. and Z.Gong, "Web Structure Mining: An Introduction", Proceedings of the IEEE International Conference on Information Acquisition, 2005.

[18] Lihui Chen and Wai Lian Chue, "Using Web structure and summarisation techniques for Web content mining", Information Processing and Management, Vol. 41 , pp. 1225–1242, 2005.

[19] Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki,, "The EigenRumor Algorithm for Ranking Blogs", In WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem, 2005.

[20] P.Boldi, M.Santini, S.Vigna, "PageRank as a Function of the Damping Factor", Proceedings of the 14th World Wide Web Conference, 2005.

[21] Abou-Assaleh T., Das T., Weizheng G., Yingbo M., O'Brien P., Zhen Z., "A Link –Based Ranking Scheme For Focused Search".In:WWW2003, ACM Press.2007.

[22] Ali Mohammad Zareh Bidoki and Nasser Yazdani, "DistanceRank: An Iintelligent Ranking Algorithm for Web Pages", Information Processing and Management, 2007.

[23] H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008..

[24] Shen Jie,Chen Chen,Zhang Hui,Sun Rong-Shuang,Zhu Yan and He Kun, "TagRank: A New Rank Algorithm for Webpage Based on Social Web" In proceedings of the International Conference on Computer Science and Information Technology,2008.

[25] A. M. Zareh Bidoki and N. Yazdani, "DistanceRank:An intelligent ranking algorithm for web pages" information Processing and Management, Vol 44, No. 2, pp. 877-892, 2008.

[26] X. Zhang and J. Chomicki, "On the semantics and evaluation of top-k queries in probabilistic databases," in *DBRank*, 2008.

[27] Fabrizio Lamberti, Andrea Sanna and Claudio Demartini , "A Relation-Based Page Rank Algorithm for. Semantic Web Search Engines", In IEEE Transaction of KDE, Vol. 21, No. 1, Jan 2009.

[28] Lian-Wang Lee, Jung-Yi Jiang, ChunDer Wu, Shie-Jue Lee, "A Query-Dependent Ranking Approach for Search Engines", Second International Workshop on Computer Science and Engineering, Vol. 1, PP. 259-263, 2009.

[29] Milan Vojnovic et al., "Ranking and Suggesting Popular Items", In IEEE Transaction of KDE, Vol. 21, No. 8, Aug 2009.

[30] NL Bhamidipati et al., "Comparing Scores Intended for Ranking", In IEEE Transactions on Knowledge and Data Engineering, 2009.

[31] Su Cheng,Pan YunTao,Yuan JunPeng,Guo Hong,Yu ZhengLu and Hu ZhiYu "PageRank, "HITS and Impact Factor for Journal Ranking", Inproceedings of the 2009 WRI World Congress on Computer Science and Information Engineering – Vol. 06, PP. 285-290, 2009 .

[32] Neelam Duhan ,A.K.Sharma and Komal Kumar Bhatia , Page Ranking Algorithms : In proceedings of the IEEE International Advanced Computing Conference (IACC),2009.

[33] Xiang Lian and Lei Chen , "Ranked Query Processing in Uncertain databases", In IEEE KDE, Vol. 22, No. 3, March 2010.

[34] P Ravi Kumar, and Singh Ashutosh kumar, "Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval", American Journal of applied sciences, 7 (6) 840-845 2010.

[35] Pooja Sharma, Pawan Bhadana, "Weighted Page Content Rank For Ordering Web Search Result", International Journal of Engineering Science and Technology, Vol 2, 2010.

[36] Kavita D. Satokar and Prof.S.Z.Gawali, "Web Search Result Personalization using Web Mining", International Journal of Computer Applications, Vol. 2, No.5, pp. 29-32, June 2010.

[37] Sharma, A.K., Duhan, N. and Kumar, G "A Novel Page Ranking Method based on Link- Visits of Web Pages". International Journal of Recent Trends in Engineering and Technology, Vol. 4, No. 1, pp 58-63. 2010

[38] Dilip Kumar Sharma, A.k. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2670-2676.

[39] Saeko Nomura, Tetsuo Hayamizu, "Analysis and Improvement of HITS Algorithm for DetectingWeb Communities". Volume 11-No 08,2011.

[40] J.Jayanthi.,K.S.Jayakumar., "An integrated Page Ranking Algorithm for Personalized Web Search". In International Journal of Computer Applications (0975-8887), Volume 12-No.11, January 2011.

[41] 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.

[42] Rekha Jain, Dr G.N.Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer application,Vol 13, Jan 2011.
[43] Tamanna Bhatia," Link Analysis Algorithms For Web Mining ", IJCST Vol. 2, Issue 2, June 2011.
[44] Dr. Paras Nath Gupta1, Pawan Singn, Punit Kr Singh and Amit Kumar"comparative analysis of page ranking algorithms"vol. 3,issue 10,2012.
[45] N. Senthil Kumar, P.M. Durai Raj Vincent " Web Mining An Integrated Approach" Vol 2, Issue 3,March 2013.
[46] Pooja Sharma, Deepak Tyagi, Pawan Bhadana, International journal of Engineering Science and Technology "Weighted Page Content Rank for ordering Web Search Result", Vol 2(12) 2010, 7301-7310.
[47] Parveen Rani, 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, July 15 ,2013.
[48] W.Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proc. Of the Second Annual Conference on Communication Networks and Services Research, IEEE,2013.
[49] Pooja Devi1, Ashlesha Gupta, Ashutosh Dixit"Comparative Study of HITS and PageRank Link based Ranking Algorithms"International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014.
[50] Punit Patel, "Research of Page ranking algorithm on Search engine using Damping factor" (IJAERD) Volume 1 Issue 1, February 2014, ISSN: 2348 – 4470.
[51] A.M. Sote, Dr. S. R. Pande" Application of Page Ranking Algorithm in Web Mining" International Conference on Advances in Engineering & Technology–2014.



Ms. S.Prabha has completed her B.E (Computer Science and Engineering) from Kongu Engineering College, Perundurai in 1998. She has worked as a lecturer in K.S.Rangasamy College of Engineering, Tiruchengode 2003.M.E(Computer Science and Engineering) from Anna University, Chennai in 2005.

She is doing her Ph.D. programme under the area Data Mining in Anna University, Chennai. She has a teaching experience of about 16 years. At present she is working as Associate professor in Information Technology department at K.S.Rangasamy College of technology, Tiruchengode. Her research interests include Database Systems, System Modeling, Compiler Design, Data Mining and Information Retrieval System. She is a life member of ISTE.



Dr. K.Duraiswamy received his B.E. degree in Electrical and Electronics Engineering from P.S.G. College of Technology, Coimbatore, Tamil Nadu in 1965 and M.Sc.(Engg) degree from P.S.G. College of Technology,Coimbatore,Tamil Nadu in 1968 and Ph.D. from Anna University, chennai in 1986. From 1965 to 1966 he was in Electricity Board. From 1968 to 1970 he was working in ACCET, Karaikudi, India. From 1970 to 1983, he was working in Government College of Engineering, Salem. From 1983 to 1995, he was with Government College of Technology, Coimbatore as Professor. From 1995 to 2005 he was working as Principal at K.S.Rangasamy College of Technology, Tiruchengode and presently he is serving as Dean in the same institution. He is interested in Digital Image Processing, Computer Architecture and Compiler Design. He received 7 years Long Service Gold Medal for NCC. He is a life member in ISTE, Senior member in IEEE and a member of CSI



Ms.J.Indhumathi holds a B.Tech degree in Information Technology from K.S.Rangasamy College of technology, Tiruchengode Tamil Nadu, India in 2013. Now she is an M.Tech student of Information Technology department in K.S.Rangasamy College of Technology. She has presented a paper in International conference. Her Research interests includes Data Mining and Web Mining.

Open Science Index, Computer and Information Engineering Vol:8, No:8, 2014 publications.waset.org/10000128/pdf