ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Comparative study of various Page Ranking Algorithms in Web Content Mining (WCM)

***Thirumala Sree Govada[1], and N Lakshmi Prasanna[2]**

1. Department of CSE, Vignan's LARA Institute of Technology & Science, JNTUK.

2. Department of CSE, Vignan's LARA Institute of Technology & Science, JNTUK.

Manuscript Info

Manuscript History:

Received: 15 May 2014
Final Accepted: 15 June 2014
Published Online: July 2014

Key words:
Correlation ranking, Eigen Rumor, HITS, QDR, TR, Web Content Mining, WLRA, WPCR

****Corresponding Author***

Thirumala Sree Govada

Abstract

World Wide Web (WWW) is a trendy and wealth of information.WWW is a huge, widely dispersed, global information service center so, it constitutes a rich source for data mining. Web mining is the make use of data mining technique to automatically extract and mine useful knowledge from the web. They are billions of HTML pages, images and other multimedia files available on web.web is facing plenty of problems i.e. 99% of information is not interested to 99% of users. Hundreds of irrelevant documents are returned in response to a user search query. It is a challenge for search engines to provide relevant information to the users according to their queries. Several ranking algorithms are defined to get the preferred result in response to user's search query. This paper refers the detailed explanation of web content mining (is a data mining technique), which is defined as "the process of eliciting useful information from the text, images and also from other forms of content that make up the pages."This paper also explores diverse web Page ranking algorithms for web content mining and compares those algorithms used for information retrieval. Different web Page Ranking algorithms like HITS (Hyperlink InducedTopicSearch), EigenRumor,WLRA(WeightedLinkRankalgorithm),TagRank,QueryDependentRanking(QDR)algorithms,weighted page content ranking,Tag rank,correlation rankingalgorithms are discussed and comparison of these algorithms in context of performance has been accomplished

Introduction

The World Wide Web (WWW) has been reached the peak of its success with respect to:

- Valuable resources of information
- Enormous number of users
- Multiform and multitude of data
- Efficient digital commerce

The profuse unstructured or semi-structured information on the web leads a great challenge for both the users, who are seeking for efficiently valuable information and for the business people, who needs to offer personalized service to the individual consumers, covered in the billions of web pages. To triumph over these problems, data mining techniques must be useful on the www. Nowadays, most of the people rely on web search engines to find and retrieve information [11]. The enormous growth, assorted, dynamic and unstructured nature of web makes internet awfully difficult in searching and retrieving relevant information and in presenting query results. There are tens and hundreds of search engines accessible but some are admired like Google, Bing, Yahoo etc., because of their brimming and ranking approaches. The search engines download, index and store up hundreds of millions of web pages. Every day search engines are giving response to millions of queries.The sample architecture [12] of a search engine is shown in Fig 1.

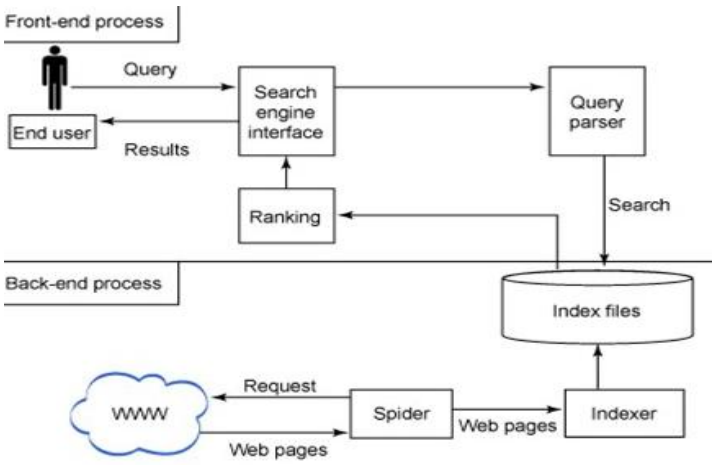

Fig 1: Web search engine

Components of web search engine are

- Search engine Interface
- Query Parser
- WWW Database
- Ranking Engine

1) Search engine Interface -It is the part of Web Search Engine interacting with the users and allowing them to query and view their query results according to their query.

2) Query Parser- It is the section providing term (keyword) extraction for both sides. The parsers found the keywords of the user query and all the terms of the Web documents.

Term extraction process includes the following sub procedures:

- Tokenization
- Normalization
- Stemming
- Stop word handling

3) WWW Database – This section contains all the text and metadata specifying the web documents.

4) Ranking Engine – This section is mainly contains the ranking algorithm operating on the current data, which is indexed, to be able to afford some order of relevance, for the web documents, with respect to the user query.

Web Mining

Web mining is the make use of data mining techniques to automatically discover and extract information from Web documents/services (Etzioni, 1996, CACM 39(11)) [13]. Naturally, deriving something helpful out of it is targeted use of web mining. It consists of following tasks,

1) *Resource finding*: It involves the task of retrieving intended web documents. It is the method by which we extract the data either from online or offline text resources accessible on web.
2) *Information selection and pre-processing*: It involves the automatic selection and pre processing of specific information from retrieved web resources.
3) *Transformation*: This process transforms the original retrieved data into information. The transformation could be rejuvenation of stop words, arising or it may be aimed for obtaining the desired representation such as finding phrases in training mass.
4) *Generalization:* It automatically determines general patterns at individual web sites as well as across multiple sites. Machine learning and data mining techniques are helpful in generalization.
5) *Analysis:* It involves the validation and interpretation of the mined patterns. It has an important role in pattern mining. A human has an important role in information on knowledge discovery process on web.

Web Mining Taxonomy

Web mining has three categories as shown in Fig 2[14].

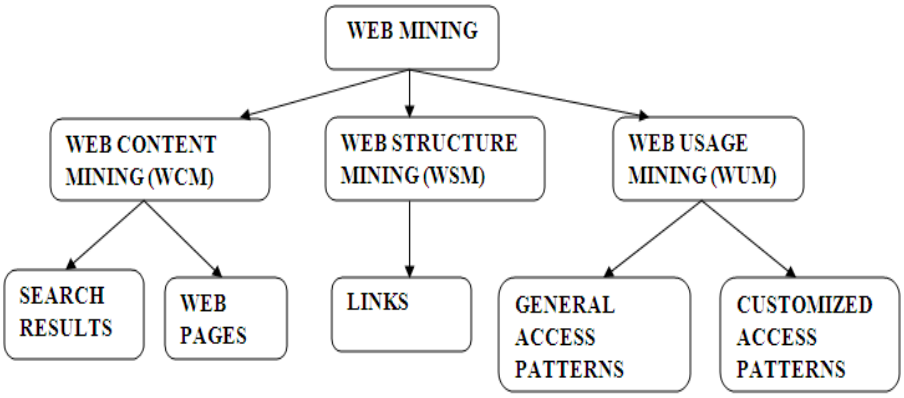

Fig 2: Web Mining Taxonomy

1. Web Usage Mining (WUM)

Web Usage Mining is the relevance of data mining techniques to realize interesting usage patterns from Web data in order to understand and to provide better needs of Web-based applications.

The web usage mining mines the secondary data i.e., the data from the web server, access logs, browser logs, proxy server logs, user profiles, registration data, user queries, user sessions or transactions and so on. Usage data grabs the identity or origin of Web users along with their browsing activities at a Web site.

2. Web Structure Mining (WSM)

Objective is to discover the structural synopsis about the web site and web pages. In essence, Web Structure Mining tries to discover the link structure of hyper links that is, focus on inter – document structure (within the web). This model can be used to classify the web pages and make it possible to compare or integrate different web pages. Web structure mining can be divided into two kinds:

- First extract the patterns from hyperlinks in the web: A hyperlink is a structural component that connects the web page to a diverse location.
- Second mine the document structure: analysis of the tree-like structure of page structures to clarify HTML or XML tag usage.

3. Web Content Mining (WCM)

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data correlates to the collection of information; a Web page was designed to pass on to the users. Web page may consist of text, images, audio, video, or structured records like lists and tables. Mining can be vital on the web documents as well the results pages produced from a search engine. Basically there are two approaches in content mining called agent based approach and database based approach. The agent based approach relies on searching proper 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.

Web Content Mining is mining the data from the content of web pages (Xu et al., 2011). Web Content Mining uses the ideas and ethics of data mining and knowledge discovery process. Using the Web for providing information is more complex than when working with static databases, due to Web dynamics and the large number of documents. Many researchers have been made to cover web content mining problems to advance the way that pages are presented to end users, improving the quality of search results and extract interesting content pages.

Ranking Algorithms

A) HITS Algorithm

HITS algorithm ranks the web page by dealing in links and out links of the web pages. In this algorithm if a web page is pointed by many hyper links then a web page is named as authority and if the page point to various hyperlinks then it is named as HUB. An image of HUB and authority are shown in Fig.3

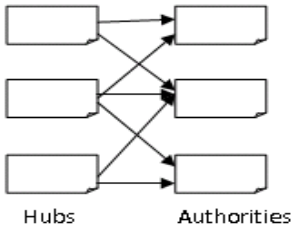

Fig 3: Illustration of Hub and Authorities

HITS algorithm is technically, a link based algorithm. In HITS [10] algorithm, ranking of the web page is determined by analyzing their textual contents against a given query. After collection of the web pages, the HITS algorithm focus on the structure of the web only, neglecting their textual contents. Priory HITS algorithm has some problems which are given below.

(i) High rank value is given to some well-known website that is not highly applicable to the given query.

(ii) Drift of the topic occurs when the hub has multiple topics as equivalent weights are given to all of the out links of a hub page. Fig 4 shows an Illustration of HITS process.

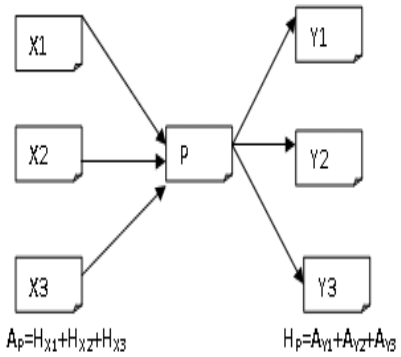

Fig 4: Illustration of HITS process

To minimize the problem of the original HITS algorithm, a clever algorithm is proposed by allusion [9]. Clever algorithm is the adjustment of standard original HITS algorithm. This algorithm provides a weight value to every link depending on the terms of queries and endpoints of the link. An anchor tag is combined to decide the weights to the link and a large hub is broken down into smaller parts so that every hub page is intensed only on one topic. Another limitation of standard HITS algorithm is that it assumes equal weights to all the links pointing to a webpage and it fails to identify the facts that some links may be more important than the other. To resolve this problem, a probabilistic analogue of the HITS (PHITS) algorithm is proposed by allusion [8]. A probabilistic clarification of relationship of term document is provided by PHITS. It is able to identify authoritative document as claimed by the author. PHITS gives better results as compared to original HITS algorithm. Other difference between PHITS and standard HITS is that PHITS can estimate the probabilities of authorities compared to standard HITS algorithm, which can provide only the scalar magnitude of authority [7].

B) Eigen Rumor Algorithm

As the number of blogging sites is increasing regularly, there is a challenge for service provider to provide good blogs to the users. Page rank and HITS shows potential in providing the rank value to the blogs but some limitations

occurs, if these two algorithms are applied directly to the blogs. The rank scores of blog entries as determined 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. To resolve these restrictions, an Eigen Rumor algorithm [5] is proposed for ranking the blogs. This algorithm provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of Eigen vector. This algorithm enables a higher score to be assigned to the blog entries submitted by a good blogger but not yet linked to by any other blogs based on acceptance of the blogger's prior work.

C) Weighted Links Rank Algorithm

Web page developers give more importance to some links using different HTML tags, because some Web resources are more significant than others. Hence, a link ranking technique that gives different weights to links may improve over uniform weight links [6]. This algorithm provides weight value to the link based on three parameters i.e. length of the anchor text, tag in which the link is contained and relative position in the page. Simulation results show that the results of the search engine are enhanced using weighted links. The length of anchor text seems to be the best attributes in this algorithm. Relative position, which reveal that physical position does not always in synchronism with logical position is not so result oriented. Future work in this algorithm allows, tuning of the weight factor of every term for further advancement.

D) Weighted Page Content Rank

Weighted Page Content Rank Algorithm (WPCR) is a anticipated page ranking algorithm which is used to give a sorted order to the web pages returned by a search engine in response to a user query.WPCR has a numerical value based on which the web pages are given an order. This algorithm handles web structure mining as well as web content mining techniques. The Web structure mining is used to calculate the weight of the web page and web content mining is used to find how much relevant a web page is? Here importance means the popularity of the page i.e. how many pages are pointing to that page or how many pages are referred by this particular page [2]. It can be calculated by depending on the number of in links and out links of the page. Here relevancy means matching of the page with the user fired query. If a page is maximally matched to the query, that becomes more relevant page than other.

E) Tag Rank Algorithm

An innovative algorithm named as Tag Rank [4] 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 heat of the tags by using time factor of the new data source tag and the annotations behavior of the web users. This algorithm provides a better authentication method for ranking the web pages. The results of this algorithm are very precise and this algorithm index new information resources in a better way. Future work in this track can be to utilize co occurrence factor of the tag to determine weight of the tag. This algorithm can also be improved by using semantic relationship among the co-occurrence tags.

F) Query Dependent Ranking Algorithm

Lian- Wang Lee, Jung- Yi Jiang, ChunDer Wu and Shie-Jue Lee [3] 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 consequent document. Whenever a query arises, then documents are extracted and ranked depending on the rank scores intended by the ranking model. The ranking form in this algorithm is the combination of various models of the similar training queries. Experimental results show that query dependent ranking algorithm is better than other algorithms.

G) Correlation Ranking Algorithm

In addition to relevance ranking, this algorithm also detects redundant documents. Removal of these redundant documents improves the quality of search results by providing unique relevant information. Normalized discounted cumulative gain method is used for evaluating this ranking algorithm [1]. Correlation analysis is used to find the related documents from the input document set of some particular category.

Comparison of various Page Ranking Algorithms

Based on the literature analysis, a comparison of some of various web page ranking algorithms is shown in Table 1 & Table 2.Comparison is done on the basis of some parameters such as main technique used, methodology, input parameter, relevancy, quality of results, importance of algorithms and limitations.

Algorithm	HITS	Eigen Rumor	Weighted Links Rank Algorithm	Weighted Page Content Rank
Main Technique	Web Content Mining	Web Content Mining	Web Content Mining	Web Content Mining
Methodology	It computes the hubs and authority of the relevant pages. It gives relevant as well as significant page as the result.	Eigen rumor use the adjacency matrix method, which is constructed from agent to object link not from page to page link.	It gives different weight to web links based on 3 attributes: Relative position in page, tag where link is contained, length of anchor text.	It gives sorted order to the web pages returned by a search engine as a numerical value in response to a user query.
I/P Parameter	Content, Back and Forward links	Agent/Object	Content, Back and Forward links	Back links, Forward links and content
Relevancy	More (this algo. Uses the hyperlinks so according to Henzinger, 2001 it will give good results and also consider the content of the page)	High for Blogs so it is mainly used for blog ranking.	More (it consider the relative position of the pages)	More relevant to a user given query.
Quality of Results	Less than Page Rank	Higher than Page Rank and HITS	Medium	High
Importance	Moderate.Hub&authorities scores are handled.	High for blog ranking.	Not specifically quoted.	It provide important information and relevancy about a given query by using web structure and web content mining
Limitations	Topic drift and efficiency problem	It is most purposely used for blog ranking not for web page ranking as other ranking like page rank, HITS.	Relative position was not so effective, representing that the logical position not always matches the physical position.	Extra calculations to find the weights of links

Table 1 Summary of various web page ranking algorithms

Algorithm	Tag Rank	Query Dependent Ranking	Correlation ranking algorithm
Main Technique	Web Content Mining	Web Content Mining	Web Content Mining
Methodology	Visitor time is used for ranking. Use of sequential clicking for sequence vector calculation with the uses of random surfing model.	This proposed the construction of the rank model by combining the results of similar type queries.	It is based on correlation method. Input datasets are preprocessed and then the term frequency for the common words between documents is computed, then correlation coefficient is computed
I/P Parameter	Popular tags and related bookmarks.	Training query.	Web documents.
Relevancy	Less as it uses the keyword entered by the user and match with the page title.	High (because the model is constructed from the training quires).	More
Quality of Results	Less	High	The quality of search results obtained through this approach is accurate.
Importance	High for social site.	High because it gives the results for user's query as well as results for similar type of query.	High document based searching
Limitations	It is comparison based approach so it requires more site as input.	Limited number of Characteristics are used to calculate the similarity.	It is Comparison based approach

Table 2 Summary of various web page ranking algorithms

Conclusions

The standard search engines usually result in a large number of pages in response to users queries, while the user always desires to get the best in a petite time. User generally spends a lot of time in sifting through the search results to find the relevant pages. The web page ranking algorithms, which are significance of web mining, play a major role in making the user search navigation easier in the results of a search engine. The paper presented a detailed comparison study of some prevalent page ranking algorithms. After going through exhaustive analysis of algorithms for ranking of web pages against the various parameters such as their methodology, i/p parameters, relevancy of results and importance of the outcome, it is concluded that these algorithms have limitations particularly in terms of time response, accuracy of results, importance of the outcome and relevancy of results. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global principles of web technology.

References

[1] G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G.V.Uma, K.Sarukesi ,"Relevance Ranking and Evaluation of Search Results through Web Content Mining" ,In proceedings of the International Multi Conference of Engineers and Computer Scientists 2012 vol.1,IMECS 2012,March 14-16,2012,HongKong.

[2] Pooja Sharma et al. ," Weighted Page Content Rank for Ordering Web Search Result",In proceedings of the International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7301-7310.
[3] 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.
[4] Shen Jie, Chen, Zhang Hui, Sun Rong-Shuang, Zhu Yan and HeKun, "TagRank: A New Rank Algorithm for Webpage Based on Social Web" In proceedings of the International Conference on Computer Science and Information Technology, 2008.
[5] Kos Fujimura, Takafumi Inoue and Masayuki Sugisaki,, "The Eigen Rumor Algorithm for Ranking Blogs", In WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem, 2005.
[6] 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.
[7] Wenpu Xing and Ali Ghorbani, "Weighted Page Rank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.
[8] D. Cohn and H. Chang, "Learning to Probabilistically Identify Authoritative Documents",. In Proceedings of 17th International Conference on Machine Learning, PP. 167–174.Morgan Kaufmann, an Francisco, CA, 2000.
[9] 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.
[10] Jon Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.
[11] http://webminingissues.blogspot.in
[12] http://www.mysoo.com.cn/news/2006/200610520.shtml
[13] http://paginas.fe.up.pt/~ec/files_0506/slides/06_WebMining.pdf
[14] http://www.infovis.net/printMag.php?num=172&lang=2 watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.