Ranking Algorithms based on Links and Contentsfor Search Engine: A Review

Charanjit Singh, Vijay Laxmi, Arvinder Singh Kang

Abstract—**The major goal of any website's owner is to provide the relevant information to the user. Due to huge growth of web pages on WWW, Web mining plays an important role to extract the useful information from the web pages. Web mining has three divided into three categories on the basis of web page feature i.e. their hyperlinks, contents and usage. To present the meaningful information to the user, various link analysis algorithm plays an important role in search engines. Some of commonly used link analysis algorithms are Page Rank, Weighted Page Rank, Weighted page rank, HITS etc. In this paper, we study various link analysis algorithm as per mining of web and compare these algorithms in context of performance has been carried out.**

Index Terms— **WWW; Data mining; Web mining; Search engine; Link based ranking algorithm; Content based ranking algorithm.**

I. Introduction

The popular segment of the internet that contains billions of documents called web pages including links, content (text, images, audio, video etc.) is WWW [1]. It is becoming unmanageable information of web page due to constantly increasing web pages days by days. Therefore, there is big challenge to retrieve meaningful information from the web world by the search engines. Some of commonly used search engines are Google, Bing, Yahoo, msn etc.

Web mining is the branch of data mining follows by search engines which are used to extracting the required information [2] by the user from the world wide web. In spites of all these, search engines return the numerous list of web pages as a result for any query entered by user.

Search engine accomplishes various tasks based on their respective designs to provide relevant information to the users. Major components of any web search engine are: Interface as user, Web Crawler, Parser, Databases andRanking Engine as shown in Fig. 1.Spider or a web crawler send by search engine to visit andcopy all the web and retrieve the data needed from them.

Using that data from the crawler, the determination of what the site about and index the information. But before user see the list of web pages, search engines use various ranking algorithm to order them as per some approach. That approaches are depends upon various features of the web pages i.e. their links, content, Hits etc. and

Finally, that ordered or re-ordered list send to the user as the result list on the basis of query entered by the user in search engine.

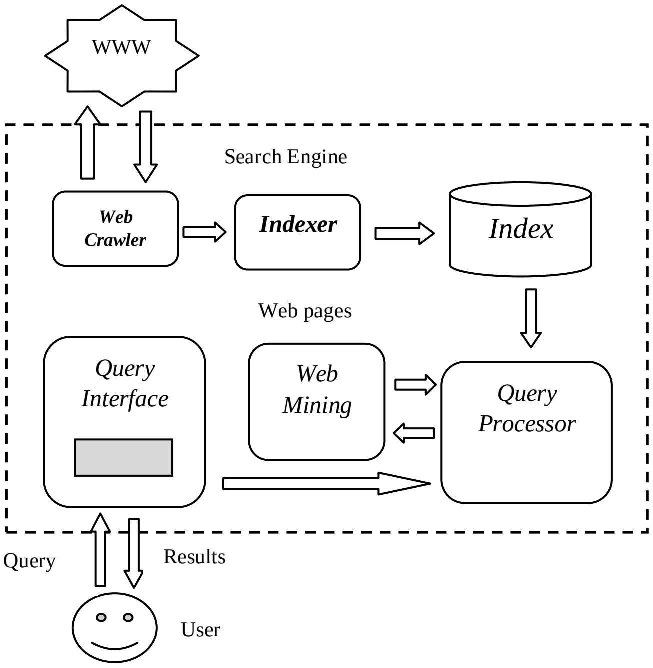

Figure 1: Architecture of Search Engine

In this study paper, various link based and content based has been reviewed and a fare comparison done. This paper is divided into four different section, in section I, concept introduction and working of search engines, section II shows concepts of web mining and their categories, in section III we are describe some link and content based ranking algorithms and finally in section IV we compare these algorithms based on their features.

A. Web Mining

It is application of data mining and used to automatically discover and fetch useful information from them. [4,5]The concept of web mining can be

Manuscript received Dec, 2017.

Charanjit Singh, Department of Computer Science &Engg.,Guru Kashi University, Talwandi Sabo, Bathinda, India.
Dr. Vijay Laxmi, Department of Computer Application, Guru Kashi University, Talwandi Sabo, Punjab, India.
Dr. Arvinder Singh Kang, Department of Computers, Chandigarh University, Gharaun, Punjab, India.

All Rights Reserved © 2017 IJARCET

divided into three broad categories (shown in figure 2) i.e.

i. Web Structure data- (contains hyperlinks, tags etc. of web pages)
ii. Web Content data- (contains text, images, records, etc. of web pages)
iii. Web Usage data-(contains http logs, app server logs, etc. of web pages)

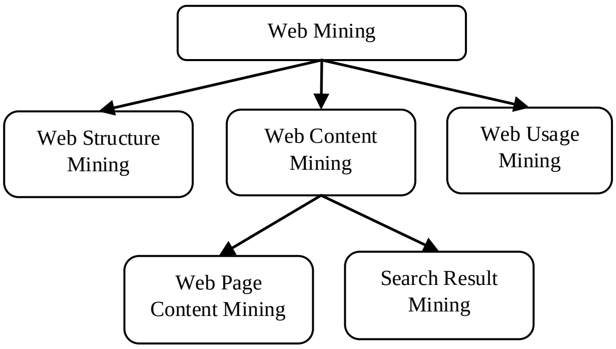

Figure 2: Structure of Web Mining

Web Structure Mining

Process of discovering structure information about the web page using graph theory is lies in web structure mining. This mining majorly applied either at the document level or hyperlink level. On the other hand, this may also determine as intra-page or inter-page, where we mining the information about the structure of web pages either interlinked webpages or within the webpages. The typical graph theory consists of web pages as nodes, and their hyperlinks as edges which are used to connect them with each other.

Web Content Mining

Extracting useful information from the contents of web documents is known as the process of content mining. Content data leads to collection of facts [6] having a web page that was designed to convey useful information to the user about the site. The available data or information on the web site may be in the form of text, video, audio, images or any structured data in the form of list and designed tables on the websites.

This may be applied on the web site or the results web page which is returned by the search engine. Another research activities such as information retrieval, natural language processing etc. are also involve in this type of mining. Web content mining is further divided into categories i.e.Web Page Content Mining and Search Result Mining

Web Usage Mining

When the web pages mine on the basis of discovering their significant pattern from data which is generated by the client-server model's transactions on one or more web localities.It can be further categorized [3] in finding the general access patterns or in finding the patterns matching the specified parameters.

Various web application such as Business intelligence, site improvement and modification, web personalization, the process of applying ranking by the search engine follows the above said categories of web mining. There is numerous ranking algorithm used by search engines based on web sites links and contents used by the search engines to provides the ranks to webpages and shows the web pages as the results to the users. Some of the link based and content based ranking algorithms are discussing in following section.

II. LINK AND CONTENT BASEDRANKING ALGORITHMS

A. Page Ranking Algorithms

The primary goal of search engines is to provide relevant information to users. Several Page Ranking Algorithms are used to rank the web pages by search engines and shown to user as a result of their query. Some algorithms rely only on the link structure of the document. i.e their popularity scores adapted by web structure mining, some commonly used page ranking algorithms have been discussed as follows:

Page Rank Algorithm

Page Rank is used to measure the importance of website pages by counting the number and quality of links to a page. Page Rank Algorithm was purposed by SurgeyBrin and Larry Page [7]. Larry Page also was cofounder of Google search engine and Page Rank was named of him. Usually used by the Google web search engine to rank websites in their search engine results.

This algorithm states that the Page Rank of a page is defined recursively and depends on the number and Page Rank metric of all pages that link to it (incoming links). If a page has some important incoming links to it than its outgoing links to other pages also become important [8]. A page that is linked to by many pages with high Page Rank receives a high rank itself.

A Page Rank Algorithm considers more than 25 billion web pages on the www to assign a rank score. A simplified version of Page Rank is defined in Eq.1:

$$PR(u) = C \sum_{V \in B(u)} PR(v) / N_v \qquad (1)$$

Here, represents 'u' as a web page, the set of pages B(u) that points to u, PR(u) and PR(v) respectively are rank scores of pages u and v respectively.

The number of outgoing links of pages v denoted as N_v, for normalization C is a factor used [9]. In Page Rank, the rank score of a page, p, is evenly divided among its outgoing links. The values assigned to the outgoing links of page p are in turn used to calculate

All Rights Reserved © 2017 IJARCET

the ranks of the pages to which page p is pointing.After further modification, algorithm was modified, observing that not all users follow the direct links on WWW. The modified version is given in Eq. 2:

$$PR(u) = (1-d) + d \sum_{V \in B(u)} PR(v) / N_v \qquad (2)$$

Here, a damping factor denotes as 'd' that is usually set as 0.85 and it can be thought of as the probability of users' following the links and (1-d) as the page rank distribution from non-directly linked pages.

Weighted Page Rank Algorithm

Algorithm assigns rank values to pages according to their importance or popularity rather than dividing it evenly. This algorithm was proposed [13], which is an extension of PageRank algorithm. The popularity is assigned in terms of weight values to incoming and as well as outgoing links which are denoted as $W^{in}(v, u)$ and $W^{out}(v, u)$ respectively.

Here, W^{in} (v, u) is the weight of link (v,u) calculated on the basis of incoming links to page u and the number of incoming links to all reference as outgoing linked pages of page v.

$$W^{in}_{(v,u)} = I_u / \sum_{p \in R(v)} I_P \qquad (3)$$

Here in expression 3, number of incoming links of page u as I_u and page p as $I_p$denoted and reference page list of page v denoted as R(v). Weight of link (v,u) calculated using W^{out}(v,u) on the basis of the number of outgoing links of page u and the number of outgoing links of all the reference pages of page v.

$$W^{out}_{(v,u)} = O_u / \sum_{p \in R(v)} O_P \qquad (4)$$

Here, number of outgoing links of page u as O_u and page p as O_p represented. Then the weighted Page Rank is given by a formula:

$$WPR(u) = (1-d) + d \sum_{V \in B(u)} WPR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \qquad (5)$$

HITS

Hyperlink- Induced Topic Search (HITS) algorithm was developed, which gives two forms of web pages known as hubs and authorities where hubs are the pages that act as resource lists on the other hand authorities are having important contents [11,12]. A fine hub page point to many authoritative pages on that context for a subject and a good authority page is pointed by many fine hub pages on the same subject. HITS assume that if the author of page p provides a link to page q, then p confers some authority on page q.A page may be a good hub and a good authority at the same time stated by Kleinberg.WWW as directed graph G(V,E) considered by HITS algorithm where V is a set of vertices representing pagesand E is a set of edges that match up to links. Fig. 3 shows the hubs and authorities in web.

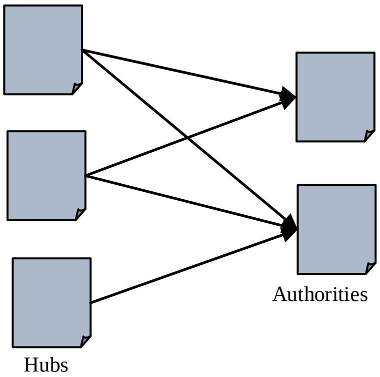

Figure 3: Association of Hubs and Authorities

The HITS algorithm works in two major steps:

- Sampling step: In this step, a set of relevant pages are collected for the given query.
- Iterative step: In this step, using the output of sampling step finds hubs and authorities. The scores of hubs and authorities are calculated as follows:

$$H_p = \sum_{q \in I(p)} A_q \qquad (6)$$

$$A_p = \sum_{q \in B(p)} H_q \qquad (7)$$

Here, representation of hub score as H_q and authority score as A_q of a page and the set of reference as I(p) and referrer pages of page p as B(p). The page's authority weight is proportional to the sum of the hub weights of pages that it links to.

B. Content based Ranking Algorithms

Page Content Rank Algorithm

This algorithm is implemented by combining a number of heuristics that seem to be important and used analyse the content of the web page. In this ranking method, page relevance ranking, which employ Web Content Mining technique and is known as Page Content Rank.

In this algorithm [13], the page importance of terms in the page is used to calculates the page importance. Whereas importance of a term is stated with respect to user's query *q*, entered as input. For its inner classification structure PCR used a neural network. Let for a given query *q* and a search engine, according to importance of pages a set R_q of ranked pages areresulted outin PCR. As in the vector model, represented pages [12] and frequencies of terms in which page is used.Working Principle of Page Content Rank Algorithm has divided in four steps which can be describes the PCR method, these steps are:

All Rights Reserved © 2017 IJARCET

Step 1: Term Extraction:A HTML parser extracted terms in R_q from each page, Afterward an inverted list is created and that will be used in step 4.

Step 2: Calculation of Parameters: Statistical parameters i.e. Term Frequency (TF) and occurrence positions. Other parameters called linguistic can calculate using the frequency of words and its synonyms in the natural language are identified.

Step 3: Classification of Terms: Using the calculation of parameters at step 2, there is process applied to determine the importance of each term. A classifier called neural network is used that to learnt training on set of terms. Each parameter agreed to excite of one neuron at the input level and excitation of output neuron achieved from the importance of a term in the time of termination propagation.

Step 4: Calculation of Relevance: Using the importance of various terms in a page, a page relevance score calculated in step 3 and achieved in this step. A new calculated score of page P is equal to the average importance of various terms in page P.

PCR states that importance of page is directly proportional to the importance of all terms in page P. Usually aggregation functions by this algorithm, i.e. *sum, Min, Max, Average, count* and another function called *Sec_moment* that is defined in Eq. 8.

$$Sec\_moment(S) = \sum_{i=1}^{n} \frac{x_i^2}{n} \quad (8)$$

Here, $S = \{X_i \mid i = 1...n\}, n = |S| .....$, *Sec_moment* used in PCR, and used in result to increases the influence of extreme values contrast to Average function.

SimRank

A new page rank algorithm which is based on similarity measure from the vector space model, called SimRank [14]. It is used to make in-order to rank the query results of web pages in an effective and efficient manner. Normally, Page Rank algorithm only employ the link relations among pages to calculate rank of each page but the content of each page cannot be ignored.

Actually, the accuracy of page scoring greatly depends on the content of the page which helps to provide the most relevant information to the users. To calculate the score of web pages, a page in vector space model is represented as a weight vector, in which each component weight is computed based on some variation of TF (Term Frequency) or TF-IDF (Inverse Document Frequency) scheme as mentioned.

- TF scheme: In TF scheme, the weight of a term denoted as t_i in page denoted as d_j is the number of times that t_i appears in document d_j, denoted as f_{ij}. The following normalization approach is applied:

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, ... f_{|V|j}} \quad (9)$$

Here, |V| is the number of terms in page and f_{ij} is the frequency count of term t_i in page j. The disadvantage as well as limitation of this scheme is that a term appears in several pages does not considered. To calculating the SimRank, the formula is applied as follows defined:

$$SimRank(p_j) = tconst * W_{ij}^{title} + bconst * W_{ij}^{body}$$
(10)

Here, (w_{1j}, w_{2j}, ... ,w_{mj}) could be denoted for p_j, term weight as W_{ij}, and some constants ‘t const’ and ‘b const’ represents between 0.1 to 1.

III. COMPARISON OF VARIOUS ALGORITHMS

By studying concept web mining and various link and content based algorithm in section I and section – II respectively. The comparison of algorithms on the basis their different features or attributes such as using Mining Techniques, Input parameters, Methodology, Relevancy, Quality of results, working levels, Importance and their limitation has been done and shown in Table 1. On the basis of these attributes/features, we can check or adapt the performance of each algorithm.

IV. CONCLUSION

The concept of web mining is an application of data mining, which is used to extract and mine the information from the web pages for enhance the performance of applications. Search engines are useful tools used to extract the useful information from the web matrix and followed various ranking algorithm based on different techniques web mining. This paper describes various ranking algorithm based on links as well as content of web pages.

Some of the ranking algorithms studied are Page Rank, Weighted Page Rank, HITS, Page Content Rank and SimRank. All algorithms may provide satisfactory performance in some cases but many times the user may not get the relevant information because some algorithms calculate the rank by considering only the links of web page but others compute by focusing content of web pages. By considering these comparison factors, a new technique can be proposed that will consider either content or link relation of a web page.

ACKNOWLEDGMENT

I would like to thanks Dr. Vijay Laxmi, Guru KashiUniversity for her spontaneous supervision

All Rights Reserved © 2017 IJARCET

and her valuedrecommendation with respect to this paper. I also thanks to Dr.Arvinder Singh Kang to help me to give shapemy ideas ofthis research paper and for being open person to discuss research work in different direction. They help me in all stages of this research article and make happen this research work efficient.

REFERENCES

[1] Companion slides for the text by Dr. M. H. Dunham, "Data Mining: Introductory and Advanced Topics", Prentice Hall, 2002.

[2] R.Cooley, B.Mobasher and J.Srivastava, "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), 1997*.

[3] Naresh Barsagade, "Web Usage Mining And Pattern Discovery: A Survey Paper", CSE 8331, Dec.8,2003.

[4] Bing Liu. "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)". *Springer-Verlag NewYork, Inc., Secaucus, NJ, USA.*

[5] Companion slides for the text by Dr. M. H. Dunham, *"Data Mining:Introductory and Advanced Topics"*, Prentice Hall, 2002.

[6] Jaroslav Pokorny, Jozef Smizansky, *"Page Content Rank: An Approachto the Web Content Mining"*.

[7] 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.*

[8] Ridings and M. Shishigin, *"Pagerank Uncovered"*. Technical report,2002.

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

[10] Qiao, S., Li, T., Li, H., Zhu, Y., Peng, J., Qin, J., "SimRank : A Page Rank Approach based on Similarity Measure", *Published in IEEE*, Print ISBN No: 978-1-4244 -6793-8, 2010, pp. 390-395.

[11] Lizorkin, D., Velikhov, P., Grinev, M., Turdakov, D., "Accuracy estimate and optimization Techniques for Simrank Computation", *Published in ACM*, Print ISBN No: 978-1-60558-305-1, on 24-30 Aug 2008, pp. 422-433.

[12] Zhao, C., Zhang, Z., Li, H., Xie, X., "A Search Result Ranking Algorithm Based on Web Pages and Tags Clustering", *Published in IEEE*, Print ISBN No: 978-1-4244-8728-8, 2011, pp.609-614.

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

[14] Jain, A., Sharma, R., Dixit, G., Tomar, V., "Page Ranking Algorithm in Web Mining, Limitations of existing methods and a new method for Indexing Web Pages", *Published in IEEE*, Print ISBN No: 978-0-7695-4958-3,2013, pp. 640-645.

[15] Kleinberg J., "Authorative Sources in a Hyperlinked Environment", *Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

Algorithm	***Page Rank***	***Weighted Page Rank***	***HITS***	***Page Content Rank***	***SimRank***
Mining Technique used	Web Structure Mining	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Content Mining	Web Content Mining
I/P Parameters	Backlinks	Backlinks, forward links	Backlinks, forward links, content	Content	Content
Methodology	Computes scores at indexing time not ay query time. Results are sorted according to the importance of pages.	Computes scores at indexing time, unequal distribution of score, pages are sorted according to importance.	Computes hub and authority scores of n highly relevant pages on the fly.	Compute New Score of top 'n' pages on the fly. Pages returned are related to the query. i.e. relevant documents are returned.	Computes scores at query time. Results are calculated dynamically.
Relevancy	Less	Less (higher than PR)	More	O(m^*)	More
Quality of results	Medium	Higher than PR	Less than PR	Approximately equal to WPR	Approx equal to WPR
Working levels	N^*	1	<N	1	1
Importance	More	More	Less	More	Less
Limitations	Query Independent	Query Independent	Topic drift and efficiency problems	Importance of pages is ignored totaly	Importance of page links is totally ignored

Table 1:Comparison of Link based and Content based Ranking Algorithms

All Rights Reserved © 2017 IJARCET