Web Mining: A Survey on Various Web Page Ranking Algorithms

Saravaiya Viralkumar M.[1], Rajendra J. Patel[2], Nikhil Kumar Singh[3]

[1]M.Tech. Student, Information Technology, U. V. Patel College of Engineering, Kherva, Gujarat, India.
[2]Assistant Professor, Computer Engineering, U. V. Patel College of Engineering, Kherva, Gujarat, India.
[3]Assistant Professor, Information Technology, U.V.Patel College of Engineering, Kherva, Gujarat, India.

---***---

Abstract -*Web Mining is very active area where research is going on actively today's. The Uses of websites are day by day increases. So user in the search engine fined the websites for their own purpose. In this situation the website owner has faces challenges to provided exact information to their respective web users. The Users are Uses the search engine is to find various kinds of information on the internet or World Wide Web. Sometimes it is very difficult to user that finds the high quality of information or exact information as they are wanted. The Duties of Page Rank Algorithm is to provides highest rank on their important pages on the internet. This Survey Paper gives the idea of various Page Rank Algorithms and gives the comparison those algorithms used for information getting from internet.*

***Key Words*:Web mining, web content mining, web structure mining, web usage mining, and page rank.**

1. INTRODUCTION

The World Wide Web (WWW) is the most likely and usable resource for getting various kinds of information which users are wants.User usually needed that user wants the exact information which they are find out on the internet sources. Search Engine Optimization Process is the well defined process for website page ranking on the search engine and this process usually uses for finding page rank on various websites or web pages. Search Engine Optimization is also used for improved website page rank on search engine.The Highest Page Rank meaning is that more number of users is uses those web sites or web pages.The hot and newest technique is search engine optimization for finding page rank for respective web site or respective web pages on the internet.The World Wide Web is most useful resource for finding multimedia resource and other relevance data.

1.1 Web Mining

In 1996, the famous scientist Etzioni who is given the first time definition of web mining. Etzioni says that various kinds of information are available on the World Wide Web and the information is structured on their respective resource [1]. Etzioni refers to that the various kinds of information is most usefully and the information is previously unknown from the others websites.

A. Web Mining Process: The Web mining Process is used for getting useful knowledge from web data is given in below Figure [2].

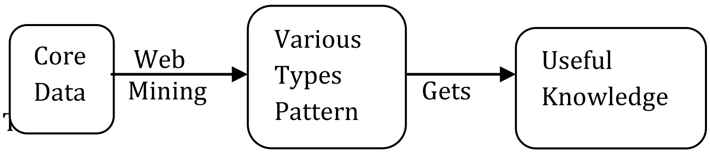

Fig.1: Web Mining Process

The various steps for web mining process are given in below:

a) **Resource Identification:** It is the process that getting the useful resource from the web site or web pages on the internet.

b) **Information Selection**: Useful Information sources are gets selected on the web resources.

c) **Generalization Patterns:** Systematically general patterns at Individual web sites on the web resources.

d) **Analysis Sources:** Analysis on various resource and get exact information or knowledge.

1.2 Web Mining Category

In Web Mining have three types of categories. They are given in below:

I. Web Content Mining
II. Web Structure Mining
III. Web Usage Mining

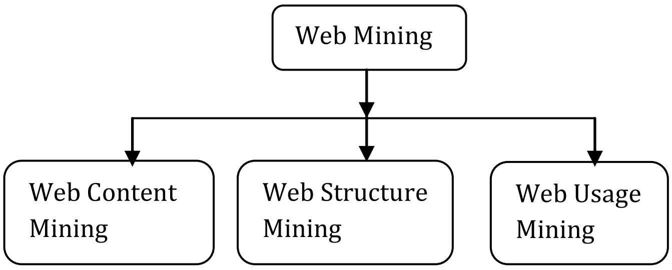

Fig.2: Web Mining Category

I. Web Content Mining

The web content mining is the process of getting various kinds of information or data from the content of the web which is available on the internet resources. The web content are text, images, audio, video etc. formats. Mostly web pages are open access for the web resource.

II. Web Structure Mining

The web structure mining is used for getting appropriate resource and information which is provide in the structural manner on the internet. The web structure mining is work on the two levels. They are the document level and the hyperlink level. The Structure of a web graph is contains two attribute. That is node and graph.

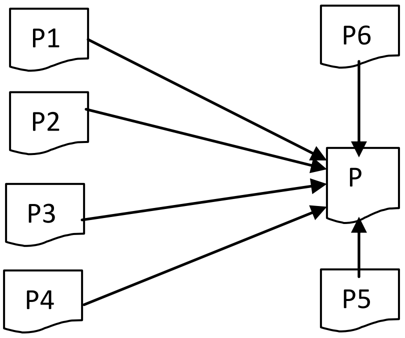

Fig.3: Web Graph Structure [3]

III.Web Usage Mining

Web usage mining [4] is process of finding out what users are looking for on the internet. Web usage mining are always prefers those techniques which are provide result in favors for users. Web usage mining duties that they are collected web records from the log files and provide exact usage patterns which user exactly needed.

2. LITERATURE REVIEW

The Page Rank Algorithm is the Google's heart. Google was designed in March 1996 as a research project by Larry Page and Sergey Brin, Ph.D. Student at Stanford University, USA. Both are Works on the Stanford Digital Library Project. The goal of SDLP was "To develop the technologies for a single, integrated and universal or unique digital library for the Stanford University" and this project was funded by the National Science Foundation and other federal agencies. The various pages are focus on the respective information for the web resources and they are finds available pattern for their web usage data on the respective web document. [5].

The Famous scientists Wenpu Xing and Ali Ghorbani are proposed the standard Page Rank which is called as a Weighted Page Rank (WPR) in 2004. Weighted Page Rank working theme is that in various pages provides highest priority on those web pages which are mostly visited by the users. This algorithm important on different pages for their rank value [6].

By Using Visit on links the Gyanendra Kumar, Neelam Duhan, A. K. Sharma was proposed PageRank. They are proposed based on Visit of Links in year 2011. Generally search engine does not give exact response for the large number of queries. Generally most of the page rank algorithm is used by links and their respective data or content. But Gyanendra Kumar, Neelam Duhan, A. K. Sharma, uses page ranking algorithm called as a Page Ranking based on Visits of Links. It will consider the page rank and inbound or outbound links. This visit of links concept are use for providing highest rank on the top most position among the other web page or web sites [7].

According to various topics the page rank are distributed for their respective resources [8]. There are various algorithms are proposed based on link analysis. Three important algorithms PageRank [9], Weighted PageRank [10] and HITS (Hyper-link Induced Topic Search) [11] are discussed in below.

A. PageRank Algorithm

Brin and Larry Page [9] was developed The Page Rank algorithm at Stanford University based on their Ph.D. Research Work. PageRank algorithm is mostly uses by the world famous search engine, Google. Another search engines are also used PageRank algorithm for ranking the different kinds of pages in the web-sites.

The Page Rank algorithm is usually depends on the link structure of the various kinds of web pages or websites which are available on the internet. The PageRank algorithm is based on the in links and out links. Thus, if back link are high than the page rank of respective is also high. Page Rank Equation [12] is -

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

Where:

PR (A) → Page Rank of page A,
PR (Ti)→ in link Page Rank of Ti pages linked to page A
C (Ti) → Out Link Page Rank of Ti pages linked to page A
d → Damping factor (between 0 and 1)

Damping factor d is mostly set as a 0.85. So it is easy to each & every page that it is distributes 85% of its original Page Rank [12].

To calculate the page rank of any web page or web site we must to know the page rank of each page. Without knowledge of respective page rank we are not able to do that point to it and number of the inbound or out bound links from each of those pages.

Now, Let us consider a simple example of three web pages.They are A, B and C.

1. Page A contains 1 out link that is pointing to Page B.
2. Page B contains 2 out links that is pointing to Page A and Page C.
3. Page C contains 1 out link that is pointing to Page A.
4. Initial page rank of each page is considered to be 1.

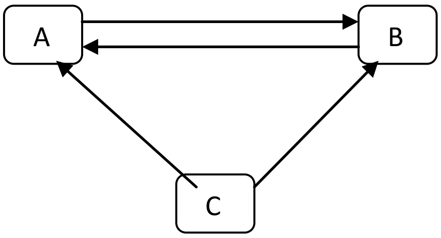

Fig.4: Three Web Pages links between each other

The Page Rank of each page is calculated by following equation

PR (A) = (1-0.85) + 0.85(2/1)
PR (B) = (1-0.85) + 0.85(2/1)
PR (C) = (1-0.85) + 0.85 (2/2)

The Result of above equation is given in below:

PR (A) = 1.85
PR (B) = 1.85
PR (C) = 1.00

The Page Rank Algorithms is used by world famous search engine Google, where the most important web pages are to be displayed at the top position in the Google search engine and rest of the others web pages are set on the bottom side of their web pages in the web search engine.

Advantages of Page Rank [13]:

a) The Most Important or relevant web sites web pages are put on the top position and irrelevant pages are put on the bottom position.

b) It is representation of web structure mining.

c) The Page Rank Algorithm is Representation is simple.

Problems of Page Rank Algorithm [14]:

a) The Page Rank Algorithm is Static Algorithm.

b) The Page Rank Algorithm is not fast enough.

c) The Page Rank Algorithm needs number of in links.

d) The Page Rank Algorithm needs numbers of out links.

B. Weighted PageRank Algorithm

The Wenpu Xing and Ali Ghorbani [10] is design a Weighted Page Rank (WPR) algorithm which is an addition of the PageRank algorithm. This algorithm assigns a higher rank values to the more important pages on websites web pages. It will separate the rank on that web site web pages value of a page rank on its outgoing linked on their respective web pages.

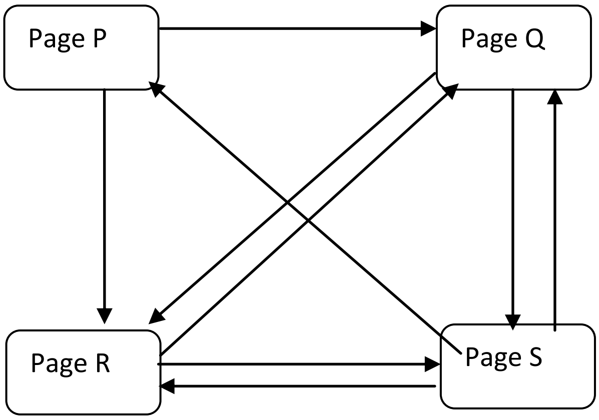

Fig.5: Structure for four pages Page Rank Algorithm

In the weighted page rank algorithm the weight values are depends upon to the incoming in link and outgoing out links. They are denoted as Win (m, n) and Wout (m, n) respectively. Win (m, n) as shown in given below equation is the weight of link (m, n). After applying the weighted page rank algorithm the calculation is based on the number of incoming in links of page n and the number of incoming in links of all pages of page m.

The Equation is proposed by The Wenpu Xing and Ali Ghorbani for the Weighted Page Rank is as shown in given below equation which a modification of the PageRank formula is given in below:

$$WPR (n) = (1-d) + d \sum WPR (m) * Win (m, n) * Wout (m, n)$$

The Page Rank Algorithms and Weighted Page Rank Algorithms both are used for provided ranking for the pages. They are based on the given query which is applied by the user for searching those kinds of things. Apply different types of weight on the website web pages on the internet resource.

Comparison of Weighted Page Rank and Page Rank [10]:

To compare the Weighted Page Rank from the standard Page Rank, they are classified into four categories. They are:

a) **Very Relevant Pages (VRP):** The Pages are consists very important information from to a given query.

b) **Relevant Pages (RP):** The Pages are relevant but it is not containing important or exact information from a given query.

c) **Weak Relevant Pages (WRP):** The Pages consists the query key-words but they do not have contain the exact data or information.

d) **Irrelevant Pages (IRP):** The Pages are not contains any kinds of relevant information and query keywords from data source.

C. HITS Algorithm

The Kleinberg [15] was developed Web Structure Mining. Web structure mining is based on algorithm which is named As Hyperlink-Induced Topic Search (HITS). It is used to that for each & every query which is given by the user request.

In HITS Algorithm there are hub are set as a web pages which is contain various link structure. Hubs are including the authorities The Hubs pages are directly or indirectly depend or it will be connected to the authority's pages. Hubs and Authorities are shown in Fig. 5. That's represent in given below:

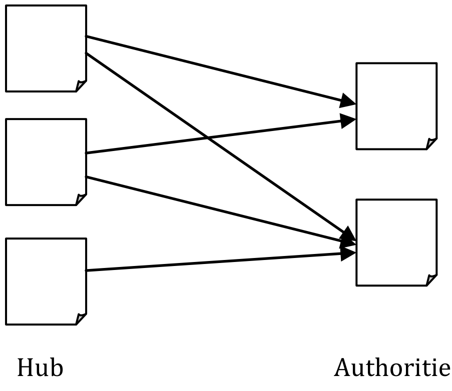

Fig.6: Hubs and Authorities [15]

The Kleinberg says that the various pages which may be a best hub and best authority at the same time on the same situation. The HITS algorithms are gives World Wide Web as a directed graph G (V, E), where V is a set of Vertices that is to representing various pages on websites and E is a set of their edges. Sometimes the edges are match up to links.

The Web page authorities are directly or indirectly depends upon hub pages. The authority's pages are proposed to the hub pages. Similarly that the hub pages are proposed to the authorities pages. The given Fig. 7 is shows an example of the calculation of authority and hub for HITS Algorithm.

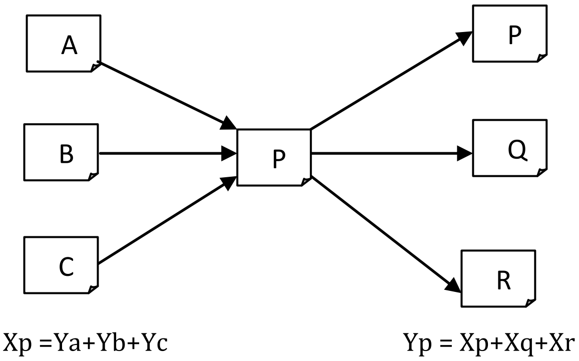

Fig.7: Calculation of Hubs and Authorities [15]

Drawbacks of HITS algorithm [16]:

a) **Hubs and authorities:** In many sites there are the hubs and the authorities are same. Sometimes that is different also.

b) **Topic Drift:** Due to equivalent weights sometime HITS Algorithm are not given exact information as the user are wanted. It means proper information some time miss.

c) **Links are generated automatically:** The HITS are gives equivalent importance of all the links which are not provides relevant result for respective user query.

D. The Intelligent Surfer QD PageRank algorithm

The Intelligent surfer QD PageRank algorithm [17] is the improved upon the standard PageRank algorithm. It will introduce a more intelligent surfer model. The web surfers are mostly uses for the page to pages, they are also depends on the respective pages and their related contents or data. The user is request for a query for given websites web pages. When choosing between multiple out-links from a respective web page from the web resource that the intelligent surfer will select a random link from the set of pages where apply the query relevant, instead of one at random from the entire set of page out-links. It may be consider as an out links also for their purpose.

Now, Assume that a condition that when a page has no out-links then the links are added from the page to all pages in the dataset which are design or develop.

The Calculation of The Intelligent surfer model in the Use of the intelligent surfer algorithm. The formula is in given below:

$$P_q(j) = (1-d)P_q'(j) + d \sum_{i \in B_j} P_q(i) P_q(i \to j)$$

E. SQDPageRank algorithm

Simultaneous Query Dependent Page Rank model [17] is probabilistic distribution model. There uses are that the terms to guided to large number of steps in the web site or different web pages. The main drawback of this model is that it is used to the combination of different single word Query Dependent PageRank to calculate the Query Dependent PageRank for a different multiple word query is a worse output other than results.

F. Query Dependent Ranking Algorithm

Lian- Wang Lee, Jung- Yi Jiang, ChunDer Wu and Shie-Jue Lee [18] are design a query dependent ranking algorithm or different techniques for search engine like Google, yahoo. In this approach it will check the similarities between the queries. Whenever a query is comes at that time, the documents are extracted. The page rank is depends on the rank scores calculated by the different ranking model. Various Experimental results come better than other algorithms.

G. Topic Sensitive PageRank

The Topic Sensitive PageRank algorithm has a several ranks which are computed. In internet there are some ranks for each page that is contents other different topics. There topics are some time related to each other area. Sometime they are matching their usual patterns.

Table -1: Comparison of Page Ranking Algorithms

Algorithm Criteria	**Page Rank Algorithm**	**QD Page Rank Algorithm**	**SQD PageRank Algorithm**	**HITS**
Web Mining Classification	Web Structure Mining	Web Content Mining	Web Content Mining	Web Structure Mining & Web Content Mining
Parameter Used	Inlinks	Query & Inlinks	Query & Inlinks	Query & Inlinks, Out links
Model	Random Surfer Model	Intelligent Surfer Model	Intelligent Surfer Model	Depends on Content & links
Method	PageRank computed	Selects pages based on relevance & computes PageRank	Measure relevance for all term in query & computes PageRank	Compute score based on content on the fly
Limitation	Query Independent	Not Work in Multi-term Query.	Identifying useful terms in the given query.	Not Efficient in real time

The PageRank are probability distribution their rank on each pages. Without knowing final value of a page rank the page rank has been calculated by each page by page. It is an iterative algorithm which is follows the principle of normalized link matrix of web. The PageRank of a page depends on the number of pages which are pointing to a page [19].

CONCLUSIONS

The Page Rank algorithm played wide role for user accessing records and find the relevance result for their respected data or content. Generally The Web mining is the Data Mining technique that automatically discovers use full the information from web by using some techniques or algorithms. The Page Rank Algorithm is used in Web Structure Mining which is used to rank the relevant pages In World Wide Web. In World Wide Web there are many advanced web searching and Data mining techniques have been worked. The Page ranking algorithms are used to ranking technical information for their respective data or contents, so that the user are able to get the most relevant information in the top of the result list which they wanted and give the proper information which has strong content.

REFERENCES

[1] Kaur, Chintandeep, and Rinkle Rani Aggarwal. "Web mining tasks and types." International Journal of Research in IT & Management (IJRIM) Vol 2, No. 2 (2012).

[2] Aggarwal, B. B. D. S., and Shivangi Dhall. "Web mining:Information and pattern discovery on the world wide web." International Journal of Science, Technology & Management (2010).

[3] Haveliwala, Taher H. "Topic-sensitive page rank: A context-sensitive ranking algorithm for web search." *Knowledge and Data Engineering, IEEE Transactions on* 15.4 (2003): 784-796.

[4] Masseglia, Florent, Pascal Poncelet, and Rosine Cicchetti. "An efficient algorithm for web usage mining." *Networking and Information Systems Journal* 2.5/6 (2000): 571-604.

[5] http://en.wikipedia.org/wiki/PageRank

[6] Duhan, Neelam, A. K. Sharma, and Komal Kumar Bhatia. "Page ranking algorithms: a survey." *Advance Computing Conference, 2009. IACC 2009. IEEE International*. IEEE, 2009.

[7] Kumar, Gyanendra, Neelam Duhan, and A. K. Sharma. "Page ranking based on number of visits of links of Web page." *Computer and Communication Technology (ICCCT), 2011 2nd International Conference on*. IEEE, 2011.

[8] Zhang, Yong, Long-bin Xiao, and Bin Fan. "*The Research about Web Page Ranking Based on the A-PageRank and the Extended VSM."Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on*. Vol. 4. IEEE, 2008.

[9] Brin, S., and L. Page. "The anatomy of a large-scale hyper textual web search engine, world wide web conference, 7." (1998).

[10] Xing, Wenpu, and Ali Ghorbani. "Weighted page rank algorithm." *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*. IEEE, 2004.

[11] Kleinberg, Jon M. "Authoritative sources in a hyper linked environment." *Journal of the ACM (JACM)* 46.5 (1999): 604-632.

[12] Sangeetha, M., and K. Suresh Joseph. "Page ranking algorithms used in Web Mining." *Information Communication and Embedded Systems (ICICES), 2014 International Conference on*. IEEE, 2014.

[13] El-Shewy, Samir, Abd El-Fatah Hegazy, and Ayman E. Khedr. "An Adaptive Web Mining Approach to Improve Customer Loyalty via Cloud Computing." (2015)

[14] N.V.Pardakhel and Prof.R.R. Keole, "Analysis of Various Web Page Ranking Algorithms in Web Structure Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013.

[15] Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." *Journal of the ACM (JACM)* 46.5 (1999): 604-632.

[16] Chakrabarti, Soumen, et al. "Mining the Web's link structure." *Computer* 32.8 (1999): 60-67.

[17] Frikh Bouchra, Ahmed Said Djaanfar, *"An Intelligent Surfer Model Based On Combings Web Contents and Links",* International Journal of Engineering Science and Technology 2011

[18] Lee, Lian-Wang, et al. "A query-dependent ranking approach for search engines." *Computer Science and Engineering, 2009. WCSE'09. Second International Workshop on*. Vol. 1. IEEE, 2009.

[19] Jain, Rekha, and Dr GN Purohit. "Page ranking algorithms for web mining."*International journal of computer applications* 13.5 (2011): 22-25.

[20] Cheng, Su, et al. "PageRank, HITS and impact factor for journal ranking."*Computer Science and Information Engineering, 2009 WRI World Congress on*. Vol. 6. IEEE, 2009.

BIOGRAPHIES

Saravaiya Viralkumar M.
M.Tech. Student, Information Technology, U.V.Patel College of Engineering, Ganpat University, Kherva, Mehsana, Gujarat, India.

Rajendra J. Patel
Assistant Professor, Computer Engineering, U.V.Patel College of Engineering, Ganpat University, Kherva, Mehsana, Gujarat, India.

Nikhil Kumar Singh
Assistant Professor, Information Technology, U. V. Patel College of Engineering, Ganpat University Kherva, Mehsana, Gujarat, India.