IOSR Journal of Computer Science (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727
PP 47-51
www.iosrjournals.org

Application of Page Ranking Algorithm in Web Mining

A.M. Sote[1], Dr. S. R. Pande[2]

[1]*Lecturer, Department of Computer Science,ACS College, Arvi, India,*
[2]*Associate Professor and Head, Department of Computer Science,SSESA's Science College, Nagpur, India*

***ABSTRACT* :** *The World Wide Web is a popular and interactive medium to disseminate information today. It is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia, and navigate between them via hyperlinks. It is very difficult for a user to find the high quality information which he wants to need. When we search any information on the web, the number of URL's has been opened. User wants to show the relevant on the top of the list. So that Page Ranking algorithm is needed which provide the higher ranking to the important pages. In this paper, we discuss the Page Ranking algorithm to provide the higher ranking to important pages.*

Keywords* -HITS, Page Rank Algorithm, Web mining, Weighted page Rank,*

I. INTRODUCTION

The World Wide Web is a very useful and interactive resource of information like hypertext, multimedia etc. When we search any information on the Google, there are many URL's has been opened. The bulk amount of information becomes very difficult for the users to find,extract and filter the relevant information. Sothat web mining techniques are used to solve these problems.

Web mining is the application of Data Mining technique to find useful information from web data. With the help of web, we can access multiple data. In the distributed information environment, document or objects are usually linked together to facilitate interactive access to that we can easily access information. There are some following tasks: [1]

1. *Resource finding:* It is the process which involvesextracting data from either online or offline text resource available on the web.
2. *Information selection and pre-processing:* The automatic selection and pre-processing of particularinformation from retrieved web resources. This process transforms the original retrieved data into information. The transformation could be renewal of stop words, stemming or it may be aimed for obtaining the desired representation such as finding phrases in the training corpus and representing the text in the first order logic form.
3. *Generalization:* It automatically discovers specific patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization.

4. *Analysis:* It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web.

In this paper we present applicability of Page Ranking algorithm in web mining. This paper is organized as follows: Section II presents Web Mining Methodology, section III presents Page Rank algorithm, section IV presents Weighted Page rank Algorithm, section V presents HITS algorithm, section VI presents comparisons of these algorithms and finally in section we presents conclusion on this papers.

I. WEB MINING METHODOLOGIES

In this section we discus web mining methodologies in short. Web mining, the application of machine learning(data mining) techniques to web based data for the purpose of learningor extracting

knowledge. Web mining methodologies can generally be classified into one of three distinct categories.

1) Web Content Mining

2) Web Structure Mining

3) Web Usage Mining

1) Web Content Mining: Web Content Mining is the process of retrieving the information from web document into more structure forms. It is related to Data Mining because many Data Mining techniques can be applied in Web Content Mining.

2) Web Structure Mining: Web Structure Mining deals with the discovering and modelling the link structure of the web. This can help in discovering similarity between sites or discovering web communities.

3) Web Usage Mining: Web Usage Mining deals with understanding user behaviour in interacting with the web site. The aim is to obtain information that may assist web site recognition to better suit the user. The logs include information about the referring pages, user identification, time a user spends at a site and the sequence of pages visited.

II. PAGE RANKING ALGORITHM

With the increasing number of Web pages and users on theWeb, the number of queries submitted to the search enginesare also increasing rapidly. Therefore, the search enginesneeds to be more efficient in its process. Web miningtechniques are employed by the search engines to extractrelevant documents from the web database and provide thenecessary information to the users. The search engines becomevery successful and popular if they use efficient Rankingmechanism. Google search engine is very successful becauseof its PageRankalgorithm.

Page ranking algorithms are usedby the search engines to present the search results byconsidering the relevance, importance and content score andweb mining techniques to order them according to the userinterest. Some ranking algorithms depend only on the linkstructure of the documents i.e. their popularity scores (webstructure mining), whereas others look for the actual content inthe documents (web content mining), while some use acombination of both i.e. they use content of the document aswell as the link structure to assign a rank value for a givendocument. If the search results are not displayed according tothe user interest then the search engine will lose itspopularity. So the ranking algorithms become very important.

Brin and Page developed PageRankalgorithm during theirPh. D. at Stanford University based on the citation analysis [2].PageRankalgorithm is used by the famous search engine,Google. They applied the citation analysis in Web search bytreating the incoming links as citations to the Web pages.However, by simply applying the citation analysis techniquesto the diverse set of Web documents did not result in efficientoutcomes.

Therefore, PageRankprovides a more advancedway to compute the importance or relevance of a Web pagethan simply counting the number of pages that are linking to it(called as “backlinks”). If a backlink comes from an“important” page, then that backlink is given a higherweighting than those backlinks comes from non-importantpages. In a simple way, link from one page to another pagemay be considered as a vote. However, not only the number ofvotes a page receives is considered important, but the“importance” or the “relevance” of the ones that cast thesevotes as well.

Assume any arbitrary page *A* has pages *T1* to *Tn*pointing toit (incoming link). *PageRank* can be calculated by thefollowing (1).

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

The parameter *d* is a damping factor, usually sets it to 0.85(to stop the other pages having too much influence, this totalvote is "damped down" by multiplying it by 0.85). *C(A)* isdefined as the number of links going out of page *A.*

ThePageRanksform a probability distribution over the Webpages, so the sum of all Web pages' *PageRank* will be one.*PageRank* can be calculated using a simple iterative algorithm,and corresponds to the principal eigenvector of the normalizedlink matrix of the Web.

III. WEIGHTED PAGE RANK ALGORITHM

Wenpu Xing and Ali Ghorbani [3] proposed a WeightedPageRank(WPR) algorithm which is an extension of thePageRankalgorithm. This algorithm assigns a larger rankvalues to the more important pages rather than dividing therank value of a page evenly among its outgoing linked pages.Each outgoing link gets a value proportional to its importance.The importance is assigned in terms of weight values to theincoming and outgoing links and are denoted as *Win*(*m, n*) and*Wout*(*m, n*) respectively. *Win*(*m, n*) as shown in (2) is the weightof *link*(*m, n*) calculated based on the number of incoming linksof page *n* and the number of incoming links of all referencepages of page *m*.

$$W_{(m,n)}^{in} = \frac{I_n}{\sum_{p \in R(m)} I_p} \quad (2)$$

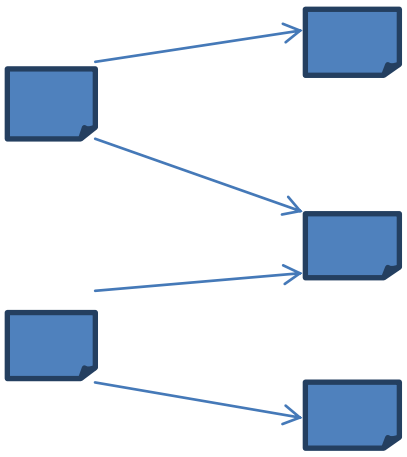
$$W_{(m,n)}^{out} = \frac{O_n}{\sum_{p \in R(m)} O_p} \quad (3)$$

Where *In* and *Ip*are the number of incoming links of page *n*and page *p* respectively. *R*(*m*) denotes the reference page listof page *m. Wout*(*m, n*) is as shown in (3) is the weight of*link*(*m, n*) calculated based on the number of outgoing links ofpage *n* and the number of outgoing links of all reference pagesof *m*. Where *On* and *Op* are the number of outgoing links ofpage *n* and *p* respectively. The formula as proposed by Wenpuet al.[3] for the *WPR* is as shown in (4) which is a modification ofthe PageRankformula.

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

IV. HITS ALGORITHM

Kleinberg [4] identifies two different forms of Web pagescalled *hubs* and *authorities*. Authorities are pages havingimportant contents. Hubs are pages that act as resource lists,guiding users to authorities. Thus, a good hub page for asubject points to many authoritative pages on that content, anda good authority page is pointed by many good hub pages onthe same subject. Hubs and Authorities are shown in Fig. 1[5].In [4] Kleinberg says that a page may be a good hub and a goodauthority at the same time. This circular relationship leads tothe definition of an iterative algorithm called HITS (HyperlinkInduced Topic Search).Hubs Authorities



Hubs Authorities

Fig. 6 Hubs and Authorities[5]

The HITS algorithm treats WWW as a directed graphG(*V*,*E*), where *V* is a set of Vertices representing pages and *E*is a set of edges that correspond to links.

There are two major steps in the HITS algorithm. Thefirst step, Samplingstep, a set of relevant pages for the givenquery are collected i.e. a sub-graph *S* of *G* is retrieved which ishigh in authority pages. This algorithm starts with a root set *R*,a set of *S* is obtained, keeping in mind that *S* is relativelysmall, rich in relevant pages about the query and contains mostof the good authorities.

The second step, Iterativestep, findshubs and authorities using the output of the sampling stepusing (5) and (6).

$$H_p=\sum_{q\in I(p)} A_q \quad (5)$$

$$A_p=\sum_{q\in B(p)} H_q \quad (6)$$

Where *Hp*is the hub weight, *Ap*is the Authority weight, *I*(*p*)and *B*(*p*) denotes the set of reference and referrer pages ofpage *p*. The page's authority weight is proportional to the sumof the hub weights of pages that it links to it, Kleinberg [6].Similarly, a page's hub weight is proportional to the sum ofthe authority weights of pages that it links to.

HITS algorithm hassome constraints which are explain below [7]:

- Hubs and authorities: It is not easy to distinguishbetween hubs and authorities because many sites arehubs as well as authorities.
- Topic drift: Sometime HITS may not produce themost relevant documents to the user queries becauseof equivalent weights.
- Automatically generated links:HITS gives equalimportance for automatically generated links whichmay not produce relevant topics for the user query.

- Efficiency: HITS algorithm is not efficient in realtime.

HITS was used in a prototype search engine called Clever[7] for an IBM research project. Because of the aboveconstraints HITS could not be implemented in a real timesearch engine.

V. COMPARISON OF ALGORITHMS

Table I shows the comparison [8] of all the algorithmsdiscussed above. The main criteria used for comparison aremining techniques used, working method, input parameters,complexity, limitations and the search engine using thealgorithm. Among all the algorithms, PageRank and HITSare most important ones. PageRank is the only algorithmimplemented in the Google search engine. HITS is used in theIBM prototype search engine called Clever. A similaralgorithm is used in the Teoma search engine and later it isacquired by Ask.com. HITS cannot be implemented directlyin a search engine due to its topic drift and efficiency problem.That is the reason we have taken PageRank algorithm andimplemented in a Java program.

Table I. Comparison of algorithms [8]

Algorithms/Criteria	PageRank	Weighted Page Rank	HITS
Mining techniques used	WSM	WSM	WSM and WCM
Working	Computes scores at indexing time.Results are sorted according to importance of pages.	Computes scores at indexing time.Results are sorted according to page importance.	Computers hub and authority scores of n highly relevant pages on the fly.
I/P Parameters	Backlinks	Backlinks, Forward links	Backlinks, Forward links and content
Complexity	O(log N)	< O(log N)	< O(log N)
Limitations	Query independent	Query independent	Topic drift and efficiency problem
Search Engine	Google	Research model	Clever

VI. CONCLUSION

In this paper we studied basic of web mining and its methodology. Special purpose of this paper is that we explain three most important algorithms of web mining Page Rank, Weighted Page Rank and HITS algorithm with formulas.

We also explains comparisons of these algorithms with different criteria such as mining techniques, I/P parameters, complexity, limitations and search engine used.

REFERENCES

[1] R. Kosala, and H. Blockeel, Web Mining Research: A Survey,SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.
[2] S. Brin, L. Page, "The Anatomy of a Large Scale Hypertextual Web search engine," *Computer Network and ISDN Systems*, Vol. 30, Issue 1- 7, pp. 07-117, 1998.
[3] W. Xing and Ali Ghorbani, "Weighted PageRank Algorithm", *Proc. Of the Second Annual Conference on Communication Networks and Services Research (CNSR '*04), *IEEE*, 2004.
[4] J. Kleinberg, "Authoritative Sources in a Hyper-Linked Environment", *Journal of the ACM* 46(5), pp. 604-632, 1999.
[5] P Ravi Kumar, and Singh Ashutoshkumar, Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of applied sciences, 7 (6) 840-845 2010.
[6] J. Kleinberg, "Hubs, Authorities and Communities", *ACM Computing Surveys*, 31(4), 1999.
[7] 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.
[8] N. Duhan, A. K. Sharma and K. K. Bhatia, "Page Ranking Algorithms: A Survey, *Proceedings of the IEEE International Conference onAdvance Computing*, 2009.