Developing an approach for hyperlink analysis with noise reduction using Web Structure Mining

[1]Mamta M. Hegde, [2]Prof. M.V. Phatak
[1] Research Associate, Department of Computer Engineering, MIT, Pune
[2] Research Scholar, Department of Computer Engineering, MIT, Pune

[1]mamtamh@gmail.com

[2]madhura.phatak@mitpune.edu.in

Abstract **– With the huge volume of web pages that exist in today's world, search engines play a vital role in the current Internet. But even if they allow finding relevant pages for any search topic, nowadays the number of results returned is often too big to be carefully explored. The need of the users varies, so that what may be interesting for one may be completely irrelevant to another. Ranking the returned webpage's such that the useful ones appear in the top of the ranked list is a critical task in the web information retrieval. The role of ranking algorithms is thus crucial, select the pages that are most likely to satisfy the user's need, and bring them in the top positions. This paper covers the popular ranking algorithm used today by the search engines: HITS. The objective of the paper is to provide right solutions for knowledge discovery and extraction on the web and then identify and eliminate noise hyperlinks from the web pages thus efficient mining can made be possible.**

Keywords **– Web mining, Web Structure mining, Hyperlink analysis, Noise Reduction.**

I. INTRODUCTION

The dramatic growth of the world-wide web, now exceeding a million pages, is forcing web search engines to look beyond simply the content of pages in providing relevant answers to queries. Recent work in utilizing the link structure of the web for improving the quality of search results is promising. The explosively growing number of Web contents, services requires an elaborate framework that can provide easy user navigation. Let's look at some of the challenges faced while locating relevant data to users search. Different kinds of web contents can offer valuable information to user. Only a part of information is useful and the remaining information is noises. How from this sea of web pages will the user find useful information needed? These metrics must be carefully selected, clearly defined so that user specific data can be provided.

A. Motivation

When using search engines to look for information on the Internet, we often find that, much of what we find is useful, also great deal of it is irrelevant to our queries, making it difficult to separate the useful information from the useless. Moreover, search results also sometimes include irrelevant commercial web pages masquerading as relevant sources of information, further creating difficulty. Thus, search engines are charged with a major task: to decide quantitatively what content is relevant and authoritative, making it easier to find useful information. HITS algorithm assigns authority rankings to every page in World Wide Web. This paper discusses HITS.

B. Page Rank Algorithm

An efficient ranking algorithm is important in any information retrieval system. Due to the magnitude of the current web, and the needs of the users, its role becomes critical. It is common for simple search queries to return thousands or even millions of results. Internet users do not have the time and patience to go through all them to find the ones they are interested in; they don't even look beyond the first page of results. Therefore, it is crucial for the ranking algorithm to output the desired results within the top few pages, otherwise, the search engine could be considered useless.

Page Rank Algorithm - L. Page and S. Brin proposed the Page Rank algorithm to calculate the importance of web pages using the link structure of the web. The Page Rank algorithm is defined as: "*We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85. Also C (A) is defined as the number of links going out of page A. The Page Rank of a page A is given as follows:*

PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn)). "*The d damping factor is the probability at each page the "random surfer" will get bored and request another random page."* Page Rank is an iterative algorithm that determines the importance of a web page based on the importance of its parent pages [1].

This paper is organized as follows: Section II describes the literature review. Section III describes the related work. The detailed design of the approach is described in section IV. Section V mentions the expected results and Section VI concludes this work.

II. LITERATURE REVIEW

The main goal of information retrieval is to find the documents relevant to a user query. Before the Web came into existence, the retrieval algorithm in information retrieval systems were usually based on the analysis of the text in the document but web changed it all. With the emergence of web, the concept of hypertext and hyperlinks came into existence. Typically, a link between two pages infers that either the content of web pages is good or the pages might be similar. So, Link analysis plays an important role in search engines. It is being used in search engine for deciding which web pages to add to the collection of documents (i.e., which pages to crawl), to order the documents matching a user query (i.e., how to rank pages), to find degree of similarity between web pages etc. In this paper, a literature survey of different papers is done for ranking web pages and finding similar pages to a given page in search engines is provided.

In paper [1], the authors have described the working and use of HITS algorithm to find the structure of web. They have defined what authoritative web pages are, and how to find authorities.

In paper [3], link mining is introduced and review of two popular methods applied in web structure mining: HITS and Page Rank are done. The authors say that web structure mining plays a vital role with various benefits including quick response to the web users.

In paper [4], an algorithm is proposed to extract the structure of a web site automatically based on hyperlink analysis. The algorithm identifies and filters noise hyperlinks by patterns of web pages these hyperlinks connected, instead of patterns of the hyperlinks. This paper addresses how to filter the "noise" hyperlinks. The author concludes by saying that because the approach is based on hyperlink analysis, it does not utilize any information of hypertext in Web pages. The next step of research would address to combine the information from hyperlinks and information from hypertext of Web pages to improve the performance of the algorithm.

In paper [5], the authors provide overview of web mining, its categories and also discuss two algorithms HITS and Page Rank. With the growing interest in Web mining, the research of structure analysis had increased and resulted in a newly emerging research area called Link Mining. Some of the possible tasks of link mining which are applicable in Web structure mining are - link-based classification, link-based cluster analysis, link type, link strength. The authors conclude paper by saying that web structure mining plays a crucial role in structural analysis.

In paper [6], the author proposed a technique to clean web pages for web data mining. Observing that the web pages in a given web site usually share some common layout or presentation style, they propose a new tree structure, called Style Tree (ST) to capture those frequent presentation styles and actual contents of the Web site. The site style tree (SST) provides with rich information for analyzing both the structures and the contents of the Web pages. They also proposed information based measure to evaluate the importance of element nodes in SST so as to detect noises. To clean a page from a site, simply map the page to its SST.

In paper [7], the authors proposed a method to eliminate noisy information from the web page and present valuable information to the user. Web page information is divided into various blocks from which the duplicate blocks are removed. For each block 3 parameters are considered –

a. Keyword redundancy
b. Linkword percentage
c. Titleword relevancy; the importance of each block is calculated.

Based on a predefined threshold value the noise blocks are removed and the remaining blocks are the useful information. Valuable knowledge from the structure of hyperlinks is being located by Web structure mining. Recognition and elimination of noise are the vital problem for extraction of information from the web.

. The different stages of proposed work are as follows:

1. Block splitting

2. Selecting distinct blocks

3. Finding block importance

4. Extracting keywords for web content mining.

The authors conclude the paper by proposing an approach for removing noises from web pages to improve the performance of web mining by the above described methods**.**

In paper [8], the authors propose a fast and efficient page ranking mechanism for web crawling and retrieval. HITS concept and Page Rank method are elaborated in this paper. The author concludes by saying this paper is one of the most successful implementation of Page Rank Algorithm.

In paper [9], the authors provided a technique to identify significant pages in the web by computing hubs and authorities. They have described the HITS algorithm for this purpose. The author hopes that this overview provides a starting point for fruitful discussion.

III. RELATED WORK

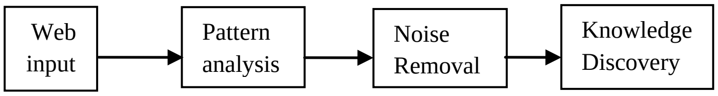

Figure 1. Block diagram for knowledge discovery

Webpage's, will be the input, these are according to the users query. User navigation logs will be collected by the way they access the links and web pages. Important links can be found based on the hit count, and links having less hit count would be treated as noise hyperlinks which can be eliminated and useful knowledge will be retrieved. This work bridges the gap between data mining and web mining disciplines. It consists of-

1. Cleaning the web pages.
2. Discovering interesting knowledge.
3. Eliminating noise hyperlinks by categorizing them as excellent, medium and weak.

For this following is required-

1. Web documents are taken as input Data cleaning removes entries unhelpful to data analysis and mining. It has to remove log entries that have status code as "failure" or "error". Some of other common indicators such as (a) the repeated request for the same URL from the same host;(b) a time interval between requests too short (c) a series of requests from one host all of whose referrer URLs are empty.

2. Pattern Analysis - Challenge of pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users. Classify Web pages in the web site in three different ways i.e. Excellent -the web pages with highest hit count. Medium - the web pages with average hit count. Weak - the web pages with least hit count.

3. Knowledge Discovery - The above statistics collected from the web site can help discovering useful knowledge. This knowledge collected can be used to take decision on various factors like: The web pages with highest hit count will be the popular pages, finding the possible navigation patterns of users?, The time spent on each web page by a user which tells about importance of the page, If time spent on particular web page is negligible it indicates that the web page does not contain useful information, the web pages with no user request indicates that page must be modified. After the pattern analysis is done on web pages, the important decision can be done regarding structure of the website. The outcome would be that the excellent web pages will be moved very near to the home page, at next level medium class web pages moved and so on. The pages with more hit count can be given the preference to be brought closer to the home page.

IV. DETAILED DESIGN

Given a specific topic T and starting set of pages S, it is necessary to find as more T on-topic pages as possible in a predefined number of steps. By step is meant visiting (and downloading and indexing) a page reachable from S following hyperlinks from pages in S. In other words it is important to estimate whether an outgoing link is promising or not. In any case when crawler decides to take into account page for link expansion, all links from the page are inserted into the crawl frontier (links that are to be visited). But many of them are not interesting at all. Sometimes links that belong to menus or footer are also misleading. Can we measure the importance of the link according to link position in the page? Links in the center of the page are probably more important than links in the down left corner. Links that are surrounded by "more" text are probably more important to topic than links positioned in groups, but groups of links can signify we are on the hub page that can also be important to our focused crawler. Can we learn positions of interesting links for some topics? In any case, information about position and belonging to a certain area can help to infer if link is promising or not!
HITS Algorithm- Kleinberg identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to Kleinberg, "Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is page that is pointed to by many good hubs". HITS associates a

nonnegative authority weight x<p> and a non-negative hub weight y<p>.

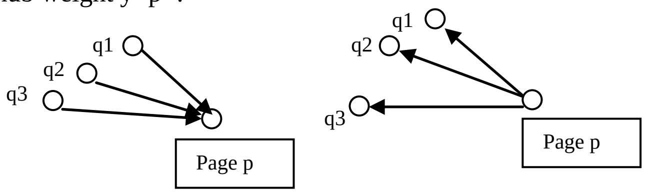

Figure 2. Operations of HITS

$x<p> = \sum y<q>$, for all q pointing to p.
$y<p> = \sum x<p>$, for all q pointed by p.

Basic operations of HITS- According to Kleinberg, "Numerically the reinforcing relationship can be expressed as follows: if p points to many pages with large x-values, then it should receive a large y-value; if p is pointed to by many pages with large y-values, then it should receive a large x-value. Given weights x<p>, y<p>, then the x-weights and y-value are as follows

$x<p> \leftarrow \sum y<q>$
$y<q> \leftarrow \sum x<p>$

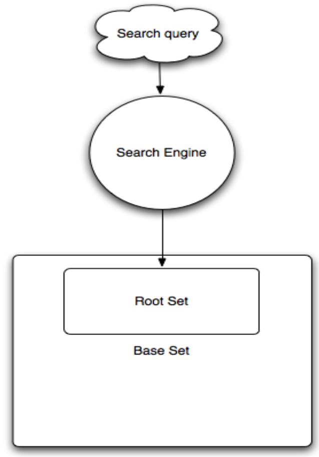

Figure 3. The Hits Algorithm

Consider a dataset that contains many web pages with many hyperlinks. Group the web pages according to their contents; web pages having similar contents will be collected in one group (clusters) and the remaining in other groups.

1. When the user fires a query many links related to the search will be returned, the user would select one of the links that will have also hyperlinks related to the search.
2. HITS Algorithm will be used to find good authorities and hubs. These authorities provide good source of contents and hubs will give a good set of hyperlinks.
3. The hyperlinks will be classified as either noise hyperlinks or recommended hyperlinks on the grounds of hit count returned by the algorithm. The user will get most recommended link for his query. In this way the user would never be misguided with noise hyperlinks and these hyperlinks can be eliminated.

Steps for HITS algorithm [11]:

1. Start with each node (web page) having a hub score and authority score of 1.

2. Run the Authority Update Rule.

3. Run the Hub Update Rule.

4. Normalize the values by dividing each Hub score by the sum of the squares of all Hub scores, and dividing each Authority score by the sum of the squares of all Authority scores.

V. EXPECTED RESULT

To test the effectiveness of search (query) matching in determining an accurate measure of the hyperlinks in web pages, extensive sets of experiments using the web based structure analysis and content similarity measure are conducted. The experimental setup consists of standard data set containing lots of web pages with many hyperlinks. In the data sets, the web page directly is analyzed. This clearly demonstrates the effect of using hyperlink structure on the web mining process. The pages will be classified as excellent, medium and weak depending on the valuable contents it contains. HITS algorithm is used to compute authorities and hubs, Authorities are the pages with good source of contents and hubs point to many good authorities.

VI CONCLUSION

Algorithms harnessing the link structure of the web are becoming increasingly useful tools to present relevant results to search queries. Web structure mining is a new area of research with rapidly growing set of research results. Due to lots of information available on the web, web mining technologies are right solutions for knowledge discovery/extraction on the web. The important decision can be done regarding structure of the website i.e. the excellent webpage's will be moved very near to the home page. The pages with more hit count can be given the preference to be brought closer to the home page. This topic focuses on the knowledge issues on web using web structure mining. For this popular algorithm HITS (to filter uninteresting information i.e. reduce noise) is studied.

REFERENCES

[1] Soumen Chakrabarti, Byron E. Dom, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, David Gibson, Jon Kleinberg,'Mining the Web's Link Structure',IEEE 1999.

[2] J.Han , M.Kamber,'Data Mining Concepts and Techniques', Book 2001.

[3] Miguel Gomes da Costa, Júnior Zhiguo Gong, Av. Padre Tomás, S.J., Taipa, Macao S.A.R., China, 'Web Structure Mining: An introduction', Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China.

[4] Feng Li, 'Extracting Structure of Web Site Based on Hyperlink Analysis', IEEE 2008.

[5] Sekhar Babu Boddu, V.P Krishna Anne, Rajesekhara Rao Kurra, Durgesh Kumar Mishra,' Knowledge Discovery and Retrieval on World Wide Web Using Web Structure Mining', Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation 2010.

[6] Guohua Hu, Qingshan Zhao,' Study to Eliminating Noisy Information in Web Pages based on Data Mining', 2010 Sixth International Conference on Natural Computation (ICNC 2010), IEEE 2010.

[7] P. Sivakumar, R. M. S Parvathi, 'An Efficient Approach of Noise Removal from Web Page for Effectual Web Content Mining', European Journal of Scientific Research ISSN 1450-216X Vol.50 No.3 (2011), pp.340-351 © EuroJournals Publishing, Inc. 2011.

[8] Lili Yan, Yingbin Wei, Zhanji Gui , Yizhuo Chen,' Research on PageRank and Hyperlink-Induced Topic Search in Web Structure Mining', IEEE 2011.

[9] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar,' Web Mining – Accomplishments & Future Directions', 200 Union Street SE, 4-192, EE/CSC Building University of Minnesota, Minneapolis, MN 55455, USA.

[10] Miloš Kovaevi, Michelangelo Diligenti, Marco Gori2, Marco Maggini, Veljko Milutinovi,' Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification', pp 1-8

[11] **http://en.wikipedia.org/wiki/HITS_algorithm. Last referred - 30/4/2012**