International Journal of Computer Applications Technology and Research
Volume 2– Issue 6, 771 - 774, 2013, ISSN: 2319–8656

A Survey on Enhancing the Efficiency of Various Web Structure Mining Algorithms

Arun Kumar Singh
Department Of Computer Science
IIMT Engineering College
Meerut, India

Avinav Pathak
Department Of Computer Science
IIMT Engineering College
Meerut, India

Dheeraj Sharma
Department Of Computer Science
IIMT Engineering College
Meerut, India

Abstract - With the fast pace in internet technology, users get easily confused in large hyper text structure. Providing the relevant information to user is primary aim of the website owner. In order to achieve this goal, they use the concept of web mining. Web mining is used to categorize users and pages by analyzing the users' behavior, the content of the pages, and the order of the URLs that tend to be accessed in order [1]. Web structure mining executes very important role in this approach. It's defined as the process of analyzing the structure of hyperlink using graph theory. There are many proposed algorithms for web structure mining such as PageRank Algorithm, HITS, Weighted page rank Algorithm, Topic Sensitive Pagerank Algorithm (TSPR), weighted page content rank Algorithm (WPCR) etc. In this paper, we have described the outline of all the algorithms and identify their strengths and limitations with a broader survey and description.

Keywords: HITS, Page rank algorithm, TSPR, Web mining weighted pagerank

I. INTRODUCTION

Web mining [2] is an application of data mining [9]. As we know data mining is the process of extracting the useful information from large amount of data present in any organization. Web mining is defined as the process of extracting relevant information from World Wide Web data. Two different approaches are taken to define the web mining. 1) "Process-centric view" which defines web mining as sequence of task. 2) "Data-centric view" which defines web mining in terms of the types of web data that is being used in the mining process. Mainly web mining is divided into three types: 1) Web Content Mining 2) Web usage mining 3) Web Structure Mining. Web content mining is the process of extracting the useful information from web document. To carry out this task there are two main methods like agent based approach and database approach. Web usage mining is the process of mining the relevant information from web history. Web usage mining process can be divided into three stages: 1) Pre-processing 2) Pattern discovery 3) Pattern analysis. In preprocessing stage, data is cleaned and partitioned into set of user's transaction that represents the activity of each user during the visiting of different sites. In Pattern discovery stage database, machine learning and statistical operations are performed to obtain the hidden patterns that reflect the behavior of user. Pattern analysis: In this, discovered patterns are further processed, filtered and analyses in user model that can be used in model that can be used as input to applications such as visualization tools and report generation tools. Web structure mining is very complex task. It is the process of analyzing the hyperlink and extract relevant information from it. It is also used to mine the structure of a document, analyze the structure of page to describe the HTML or XML usage. The goal of the Web Structure Mining is to generate the structural summary about the Web site and Web page. Web Structure mining will further categorize the Web pages and generate the information like the similarity and relationship between different Web sites. This type of mining can be either performed at document level that is referred to as intra-page [11] or at hyperlink level that referred to as inter-page mining. Due to the significance of this mining technique, there have been several algorithms proposed to solve this. In this paper we will describe and analyze web structure mining algorithms and identify their strengths and limitations. The rest of this paper is organized as follows. In Section 2 Process of web structure mining is described and several proposed algorithms are introduced. Section 3 gives the detailed of these algorithms is given. In Section 4 we provide a Comparison of the different web structure mining algorithms. In Section 5 we conclude.

2. PROCESS OF WEB MINING

Web structure mining is also known as "Link Analysis" function. It is a past area of research but with the increasing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area also called Link Mining. The Web contains a multiple variety of objects with almost no unifying structure, with differences in the style and content much larger than in traditional collections of text documents. Link mining is divided into four parts and is shown in following figure:

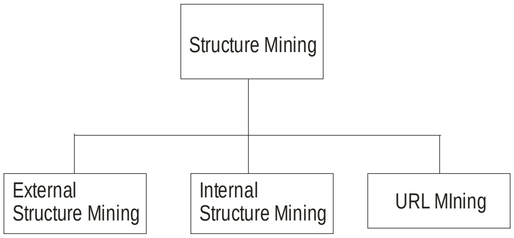


Fig 1.0

The objects in the WWW are web pages, and links are in-, out- and co-citation i.e. two pages that are both linked to the same page. There are some possible tasks [2] of link mining which are applicable in Web structure mining and are described as follows: **1. Link-based Classification:** - is the most recent upgrade of a classic data mining task to linked Domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page. **2. Link-based Cluster Analysis.** The aim in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data. **3. Link Type.** There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link. **4. Link Strength.** Links could be associated with weights. **5. Link Cardinality.** The main task here is to predict the number of links between objects.

There are some uses of web structure mining like it is:

1. Used to rank the user's query
2. Deciding what page will be added to the collection
3. page categorization
4. finding related pages
5. finding duplicated web site and also to find out similarity between them

3. PROPOSED ALGORITHMS

A. Pagerank algorithm

Pagerank algorithm is link analysis algorithm[8] that was discovered by Larry page. This algorithm is used by Google internet search engine. In this algorithm numerical weight is assigned to each element of hyperlink set of document such as World Wide Web, with the purpose of measuring the relative importance of that particular set in that hyperlink. Pagerank is a probability distribution algorithm used to represent the person's randomly clicking on links will arrive at any particular page. A probability is expressed as a numeric value between 0 and 1. That numerical value is defined as damping factor. It is represented as d and usually its value set to be 0.85. Also C (A) is the number of link going out of that particular page and is known as backlink. Pagerank of any page is evaluated by: PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn)) (1) Where PR (A) is pagerank of particular web page A D is damping factor. PR (T1) is page link with main page PR (A) C is outlink. Page Rank can be calculated using a simple iterative algorithm, and corresponds to the principal Eigen vector of the normalized link matrix of the web. Page Rank algorithm needs a few hours to calculate the rank of millions of pages.

B. HITS

This algorithm [9] was given by Kleinberg in 1997. According to this algorithm first step is to collect the root set. That root set hits from the search engine. Then the next step is to construct the base set that includes the entire page that points to that root set. The size should be in between 1000-5000. Third step is to construct the focused graph that includes graph structure of the base set. It deletes the intrinsic link, (the link between the same domains). Then it iteratively computes the hub and authority scores. In HITS concept, he 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 him, A good hub is a page that points to many good authorities; a Good authority is a page that is pointed to by many good hubs". Although HITS provides good search results for a wide range of queries, HITS did not work well in all cases due to the following three reasons: 1 Mutually reinforced relationships between hosts. Sometimes a set of documents on one host point to a single document on a second host, or sometimes a single document on one host point to a set of document on a second host. 2. Automatically generated links. Web document generated by tools often have links that were inserted by the tool. 3. Non-relevant nodes. Sometimes pages point to other pages with no relevance to the query topic.

C. Weighted pagerank algorithms [6]:

Wenpu Xing and Ali Ghorbani proposed a Weighted Pagerank algorithm which is an extension of the Pagerank algorithm. This algorithm assigns a larger rank values to the more important pages rather than Dividing the rank value of page evenly among its outgoing linked pages, each outgoing link gets a value proportional to its importance. In this algorithm weight is assigned to both backlink and forward link. Incoming link is defined as number of link points to that particular page and outgoing link is defined as number of links goes out. This algorithm is more efficient than pagerank algorithm because it uses two parameters i.e. backlink and forward link. The popularity from the number of in links and out links is recorded as Win and Wout respectively. Win (v, u) is the weight of link (v, u) calculated based on the number of in links of page u and the number of in links of all reference pages of page v.

D. Weighted page content rank algorithm:

Weighted Page Content Rank Algorithm (WPCR) [7] is a proposed 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 is a numerical value based on which the web pages are given an order. This algorithm employs web structure mining as well as web content mining techniques. Web structure mining is used to calculate the importance of the page and web content mining is used to find how much relevant a page is? Importance here means the popularity of the page i.e. how many pages are pointing to or are referred by this particular page. It can be calculated based on the number of in links and out links of the page. Relevancy means matching of the page with the fired query. If a page is maximally matched to the query, that becomes more relevant. This algorithm is better than the pagerank as well as weighted pagerank algorithm because its complexity is less than both the algorithm and is < (Ologn).

E. Topic Sensitive PageRank Algorithm [5]:

In this algorithm, separate scores are evaluated, multiple important scores for each page under several topics that form a composite Pagerank score for those pages matching the query. At query time, the similarity of the query is compared to each of these vectors or topics; and subsequently, instead of using a single global ranking vector, the linear combination of the topic-sensitive vectors is weighed using the similarity of the query to the topics. This method yields a very accurate set of results relevant to the context of the particular query. For each web document query sensitive importance score. The results are ranked according to this composite score. It provides a better scalable approach for search rankings using Link analysis. For each Web page, compute an importance score per topic. At query time, these importance scores are combined based on the topics of the query and associated context to form a composite Page rank score for those pages matching the query. This score can be used in conjunction with other scoring schemes to produce a final rank for the result pages with respect to the query. This algorithm will improve the order of web pages in the result list so that user may get the relevant pages easily.

4. COMPARISON OF DIFFERENT ALGORITHMS

By analysing the literature review of significant web page ranking algorithms, it is concluded that each algorithm has some relative strengths and limitations. A tabular summary is given below in table 1.0, which summarizes the techniques, Advantages and limitations of some of important web page ranking algorithms:

5. CONCLUSION

This paper described various proposed web structure mining algorithms like pagerank algorithm, weighted pagerank algorithm, weighted content pagerank algorithm (WCPR), HITS etc. We examined their strengths and limitations and provide comparison among them. Hence, we can say that this paper may be used as a reference by researchers when deciding which algorithm is suitable. The present status defines that the algorithms have worked efficiently well and can be improved. Slight change in the parameters may fulfill the present day conditions but it has to checked that the goal is not disturbed.

6. REFERENCES

[1] R. Kosala and H. Blockeel, "Web Mining Research: A survey", In ACM SIGKDD Explorations,

[2] Q. Lu and L. Getoor. Link-based classification. In *Proceedings of ICML-03*, 2003.

[3] 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.

[4] S. Pal, V. Talwar, and P. Mitra. Web mining in soft computing framework: Relevance, state of the art and future directions *IEEE Trans. Neural Networks*, 13(5):1163–1177 2002.

[5] Mishra Shesh Narayan et al (2012), "An Effective algorithm for web mining based on Topic sensitive pagerank algorithm", International journal of computer science and software engineering.

[6] Wenpu Xing and Ali Ghorbani, "Weighted Pagerank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

[7] Bhatia Tamanna (2011), "Link analysis algorithm for web mining", International journal of computer science and software engineering.(0976-8491).

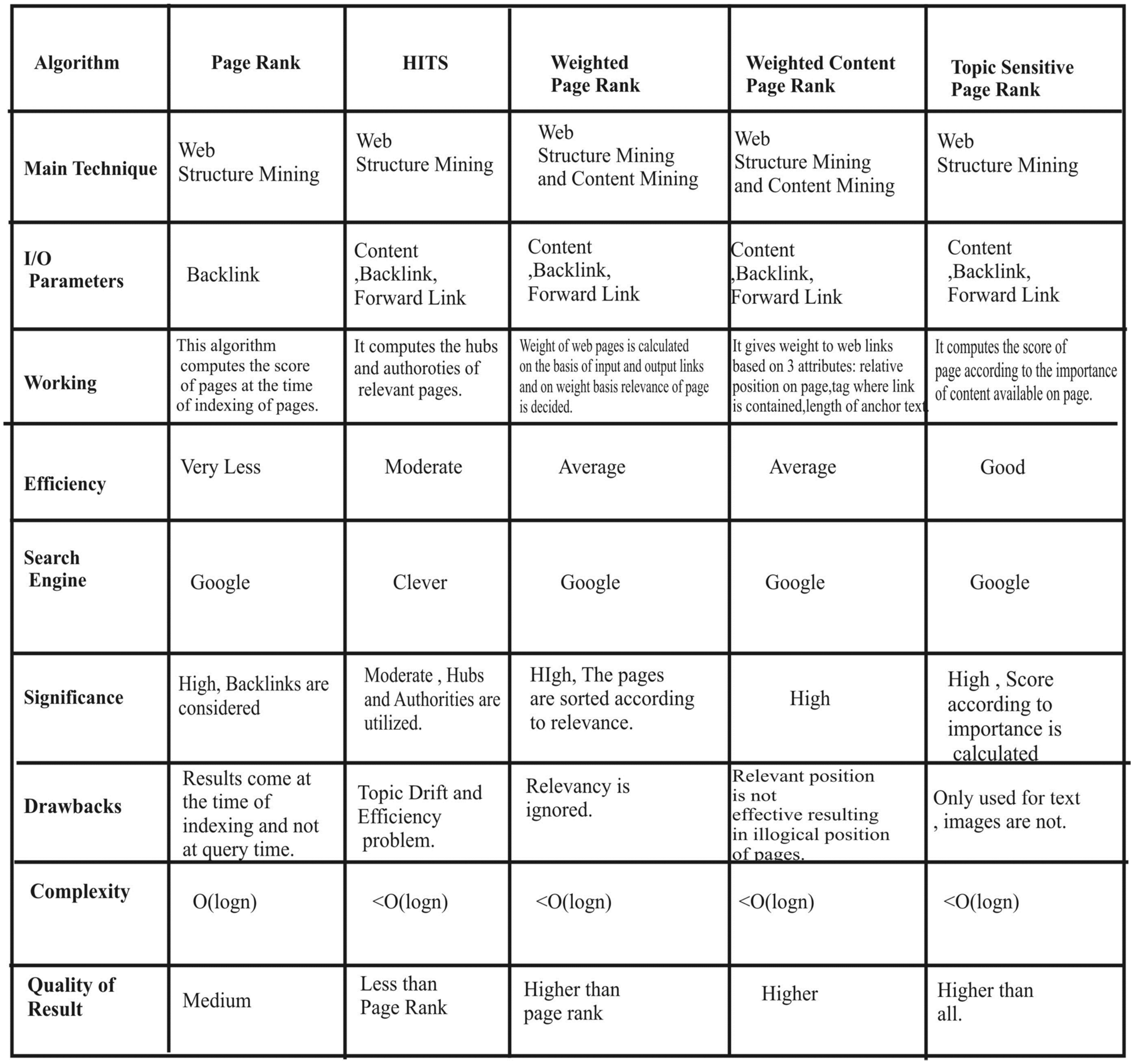
[8] Munibalaji T and , Balamurugan C (2012), "Analysis of Link Algorithms for Web Mining", International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 2,

[9] Mukhopadhyay et al.(2006)," A Syntactic Classification based Web Page Ranking Algorithm", 6th International Workshop on MSPT Proceedings.

[10] http://www.csbdu.in/econtent/Data%20Mining%20&%20 Warehousing/Unit%20III.pdf

[11] http://www.ieee.org.ar/downloads/Srivastava-tut-pres.pdf

Table 1 Comparison of Algorithms



Algorithm	Page Rank	HITS	Weighted Page Rank	Weighted Content Page Rank	Topic Sensitive Page Rank
Main Technique	Web Structure Mining	Web Structure Mining	Web Structure Mining and Content Mining	Web Structure Mining and Content Mining	Web Structure Mining
I/O Parameters	Backlink	Content ,Backlink, Forward Link	Content ,Backlink, Forward Link	Content ,Backlink, Forward Link	Content ,Backlink, Forward Link
Working	This algorithm computes the score of pages at the time of indexing of pages.	It computes the hubs and authoroties of relevant pages.	Weight of web pages is calculated on the basis of input and output links and on weight basis relevance of page is decided.	It gives weight to web links based on 3 attributes: relative position on page,tag where link is contained,length of anchor text	It computes the score of page according to the importance of content available on page.
Efficiency	Very Less	Moderate	Average	Average	Good
Search Engine	Google	Clever	Google	Google	Google
Significance	High, Backlinks are considered	Moderate , Hubs and Authorities are utilized.	HIgh, The pages are sorted according to relevance.	High	High , Score according to importance is calculated
Drawbacks	Results come at the time of indexing and not at query time.	Topic Drift and Efficiency problem.	Relevancy is ignored.	Relevant position is not effective resulting in illogical position of pages.	Only used for text , images are not.
Complexity	O(logn)	<O(logn)	<O(logn)	<O(logn)	<O(logn)
Quality of Result	Medium	Less than Page Rank	Higher than page rank	Higher	Higher than all.