ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol.2, Issue 12, December 2013

Analysis of Various Web Page Ranking Algorithms in Web Structure Mining

N. V. Pardakhe[1], Prof. R. R. Keole[2]

Department of Computer Science and Engineering

H.V.P.M's college of Engineering and Technology,

Amravati University, India

Abstract: With the rapid increase in internet technology, users gets easily confused in large hyper text structure. Providing the relevant information to user is primary goal 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 analysing the users' behaviour, the content of the pages, and the order of the URLs that tend to be accessed in order. Web structure mining plays very important role in this approach. It's defined as the process of analysing the structure of hyperlink using graph theory. There are many proposed algorithms for web structure mining such as PageRank (PR), Weighted PageRank (WPR), and Hyperlink-Induced Topic Search (HITS) etc. This paper studied about web mining and its various techniques. Different web page ranking algorithms are also compared based on their methodology, relevancy, quality of results and limitations etc.

Keywords: HITS, Page Ranking, web structure mining, weighted Page Ranking

I. INTRODUCTION

Today, the World Wide Web is the popular and interactive medium to disseminate information. The Web is huge, diverse and dynamic. The Web contains vast amount of information and provides an access to it at any place at any time. Most of the people use internet for retrieving information. But most of the time, they gets lots of insignificant and irrelevant document even after navigating several links. For retrieving information from the Web, Web mining techniques are used.

A. *Web Search Engine*

Web Search Engine is a tool enabling document search, with respect to specified keywords, in the Web and returns a list of documents where the keywords were found.

Fig.1 Variousn Web Search Engines

Components of Web Search Engine

1. User Interface
2. Parser
3. Web Crawler
4. Database
5. Ranking Engine

1) User Interface -It is the part of Web Search Engine interacting with the users and allowing them to query and view query results.

2) Parser- It is the component providing term (keyword) extraction for both sides.The parsers determines the keywords of the user query and all the terms of the Web documents which have been scanning by the crawler. Term extraction procedure includes the following subprocedures:

1. Tokenization
2. Normalization
3. Stemming
4. Stop word handling

3) Web Crawler - A web crawler is a relatively simple automated program, or script, that methodically scans or "crawls" through Internet pages to create an index of the data it is looking for. Alternative names for a web crawler include web spider, web robot, crawler, and automatic indexer.

When a web crawler visits a web page, it reads the visible text, the hyperlinks, and the content of the various tags used in the site, such as keyword rich meta tags.

Using the information gathered from the crawler, a search engine will then determine what the site is about and index the information. Lastly, the website is included in the search engine's database and its page ranking process.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol.2, Issue 12, December 2013

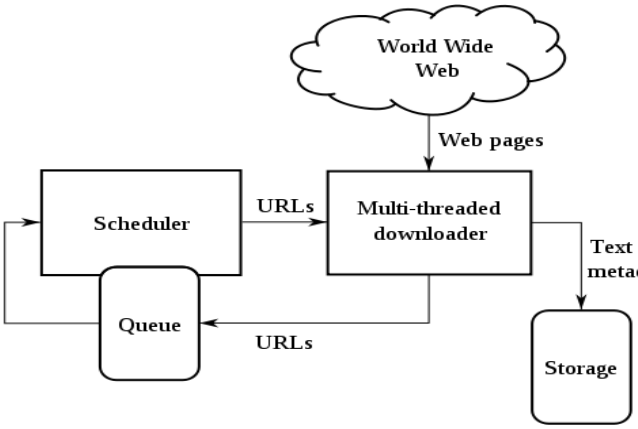

Fig.2 Web Crawler Architecture

4) Database - It is the component that all the text and metadata specifying the web documents scanned by the crawler.

5) Ranking Engine - The component is mainly the ranking algorithm operating on the current data, which is indexed by the crawler, to be able to provide some order of relevance, for the web documents, with respect to the user query.

The Web is perhaps the single largest data source in the world. Due to the heterogeneity and lack of structure, mining and integration are challenging tasks. Much of the Web mining is about Data/information extraction from semi-structured objects and free text, and Integration of the extracted data/information. Problems Faced By Information Users are:

- Finding Relevant Information
- Personalization of the information available on the web
- Learning about customers or individual users

B. *Web Mining*

Web Mining is the use of data mining techniques to automatically discover and extract Information from web documents and services. The World Wide Web, www or web is becoming a complex universe. Naturally, deriving something valuable out of it is targeted use of web mining. Three sub categories:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

1) Web Content Mining

Web Content Mining refers to the discovery of useful information from the web content, here Content refers to Text, Audio Video etc. that numerous websites are holding. Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data.

Example of Web Content Mining is:

Typical Google or Yahoo or Microsoft Bing search that we do, and the resultant links listing page we get is an example of content mining. The process of extracting useful information from the web content happens here.

2) Web Usage Mining

Web usage mining process involves the log time of pages. The world's largest portal like yahoo, msn etc., needs a lot of insights from the behaviour of their users' web visits. Without this usage reports, it will be difficult to structure their monetization efforts. Usage mining has direct impact on businesses.

The challenges involved in web usage mining could be divided in three phases:

A. Pre-processing. The data available tend to be noisy, incomplete and inconsistent. In this phase, the data available should be treated according to the requirements of the next phase. It includes data cleaning, data integration, data transformation and data reduction.

B. Pattern discovery. Several different methods and algorithms such as statistics, data mining, machine learning and pattern recognition could be applied to identify user patterns.

C. Pattern Analysis. This process targets to understand, visualize and give interpretation to these patterns.

Example of Web Usage Mining is:

A particular feature of website that is used by the visitors frequently, that we want to enhance and pronounce so as to increase the usage that can appeal more to users of the website Simply by understanding the movement of the guests and the behaviour of surfing the net, you can look forward to meet the preferences and the needs in a better manner and popularize your website among the masses in the internet arena.

3) Web Structure Mining

Web structure mining is done at the hyper link level. This kind of mining tries to discover the model underlying the link structure of the web. A relevant example can be Google's Page rank.

The goal of the Web Structure Mining is to generate the structural summary about the Web site and Web page. It tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web Structure mining will categorize the Web pages and generate the information like similarity and relationship between different Web sites. This type of mining can be performed at the document level (intra-page) or at the hyperlink level (inter-page). It is important to understand the Web data structure for Information Retrieval. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents.

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 of link mining which are applicable in Web structure mining and are described as follows: [2] [13]

A. 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,

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol.2, Issue 12, December 2013

based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

B. Link-based Cluster Analysis. The goal 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.

C. 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.

D. Link Strength. Links could be associated with weights.

E. 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:

- Used to rank the user's query
- Deciding what page will be added to the collection
- Page categorization
- Finding related pages
- Finding duplicated web sites
- And also to find out similarity between them

TABLE 1
WEB MINING CATEGORIES

	WEB MINING			
	Web Content Mining		**Web Structure Mining**	**Web Usage Mining**
	IR View	**DB View**		
View of Data	-Unstructured -Structured	- Semi Structured - Web Site as DB	- Link Structure	-Interactivity
Main Data	-Text Documents-Hypertext Documents	-Hypertext Documents	- Link Structure	-Server Logs-Brower Logs
Representation	-Bag of Words-Phrases, concept of ontology-Relational	-Edge labelled Graph-Relational	-Graph	-Relational Table-Graph
Method	-Machine Learning -Statistical (including NLP)	-Proprietary Algorithms -Association Rules	-Proprietary Algorithms	-Machine Learning-Statistical-Association Rules
Application Categories	-Categorization -Clustering -Finding extract rules -Finding Patterns in text	- Finding frequent sub structures - Web site schema discovery	-Categorization -Clustering	-Site Construction -Adaptation and management -Marketing -User Modelling

II. WEB PAGE RANKING ALGORITHMS

The search engines on the Web need to be more efficient because there are extremely large number of Web pages as well queries submitted to the search engines. Web mining techniques are employed by the search engines to extract relevant documents from the web database and provide the necessary information to the users. Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some ranking algorithms depend only on the analysis of the link structure of the documents i.e. their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining), while some use a combination of both i.e. they use content of the document as well as the link structure to assign a rank value for a given document. There are number of algorithms proposed based on link analysis. Three important algorithms, such as PageRank, Weighted PageRank and HITS (Hyper-link Induced Topic Search) are discussed below.

A. PageRank Algorithm (Google)

The "PageRank" algorithm, proposed by founders of Google Sergey Brin and Lawrance Page, is one of the most common page ranking algorithms that is also currently used by the leading search engine Google. In general the algorithm uses the linking (citation) info

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol.2, Issue 12, December 2013

occurring among the pages as the core metric in ranking procedure. Existence of a link from page p1 to p2 may indicate that the author of is interested in page.
The PageRank metric PR(p), defines the importance of page p to be the sum of the importance of the pages that point to p. More formally, consider pages p1,…,pn, which link to a page pi and let cj be the total number of links going out of page pj. Then, PageRank of page pi is given by:

PR(pi)=d+(1-d)[PR(p1)/c1+…+PR(pn)/cn]

where d is the damping factor.

This damping factor d makes sense because users will only continue clicking on links for a finite amount of time before they get distracted and start exploring something completely unrelated. With the remaining probability (1-d), the user will click on one of the cj links on page pj at random. Damping factor is usually set to 0.85. So it is easy to infer that every page distributes 85% of its original PageRank evenly among all pages to which it points.

Problems of PageRank Algorithm are:

- It is a static algorithm that, because of its cumulative scheme, popular pages tend to stay popular generally.
- Popularity of a site does not guarantee the desired information to the searcher so relevance factor also needs to be included.
- In Internet, available data is huge and the algorithm is not fast enough.
- It should support personalized search that personal specifications should be met by the search result.

B. HITS (Hyper-link Induced Topic Search) Algorithm (IBM)

It is executed at query time, not at indexing time, with the associated hit on performance that accompanies query-time processing. Thus, the hub(going) and authority(coming) scores assigned to a page are query-specific. It is not commonly used by search engines. It computes two scores per document, hub and authority, as opposed to a single score of PageRank. It is processed on a small subset of 'relevant' documents, not all documents as was the case with PageRank.
This algorithm 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. 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".

Problems of HITS Algorithm are
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][13]

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 Algorithm

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 W_{in} and W_{out} respectively. W_{in} (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. [2][3]

III COMPARISON

Based on the literature analysis, a comparison of some of various web page ranking algorithms is shown in table 2. Comparison is done on the basis of some parameters such as main technique use, methodology, input parameter, relevancy, quality of results, importance and limitations.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol.2, Issue 12, December 2013

TABLE 2
COMPARISON OF VARIOUS WEB PAGE RANKING ALGORITHMS

Algorithm	Page Rank	HITS	Weighted Page Rank
Main Technique	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Structure Mining
Methodology	This algorithm computes the score for pages at the time of indexing of the pages.	It computes the hubs and authority of the relevant pages. It relevant as well as important page as the result.	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided.
Input Parameter	Back links	Content, Back and Forward links	Back links and Forward links.
Relevancy	Less (this algo. Rank the pages on the indexing time)	More (this algo. Uses the hyperlinks so according to Henzinger, 2001 it will give good results and also consider the content of the page)	Less as ranking is based on the calculation of weight of the web page at the time of indexing.
Quality of results	Medium	Less than PR	Higher than PR
Importance	High. Back links are considered.	Moderate. Hub & authorities scores are utilized.	High. The pages are sorted according to the importance.
Limitation	Results come at the time of indexing and not at the query time.	Topic drift and efficiency problem	Relevancy is ignored.

IV CONCLUSION

In this paper it has been mentioned the introduction of web mining and its related techniques such as web content mining, web structure mining and web usage mining and are also tabulated**.** The goal of search engines is to provide relevant information to the users to cater to their needs. Therefore, finding the content of the Web and retrieving the users' interests and needs have become increasingly important. The different algorithms used for link analysis like PageRank (PR), Weighted PageRank (WPR), Hyperlink-Induced Topic Search (HITS) algorithms are discussed and compared depending on which the aim to discover an efficient and better system for mining the web topology to identify authoritative web pages.

REFERENCES

[1] Ashish Jain, Rajeev Sharma, Gireesh Dixit, Varsha Tomar ," Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages", 2013 IEEE International Conference on Communication Systems and Network Technologies.

[2] Seifedine Kadry , Ali Kalakech ," On the Improvement of Weighted Page Content Rank", *Journal of Advances in Computer Networks, Vol. 1, No. 2, June 2013.*

[3] Rashmi Rani, Vinod Jain ," Weighted PageRank using the Rank Improvement" International Journal of Scientific and Research Publications, Volume 3, Issue 7, July 2013.

[4] Preeti Chopra, Md. Ataullah ,"A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013.

[5] B.Aysha Banu, Dr.M.Chitra.," A Novel Ensemble Vision Based Deep Web Data Extraction Technique for WebMining Applications", 2012 IEEE Intemational Conference on Advanced Communication Control and Computing Technologies (ICACCCT).

[6] P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar," Content Based Ranking for Search Engines",Proceedings of the International MultiConference of Engineers and Computer Scientists 2012 Vol I, Hong Kong.

[7] Dilip Kumar Sharma, A. K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2670-2676.

[8] Mohamed-K HUSSEIN, Mohamed-H MOUSA ," An Effective Web Mining Algorithm using Link Analysis", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (3) , 2010, 190-197.

[9] Shesh Narayan Mishra, Alka Jaiswal, Asha Ambhaikar ," Web Mining Using Topic Sensitive Weighted PageRank", International

Journal of Scientific & Engineering Research Volume 3, Issue 2, February-2012 , ISSN 2229-5518.

[10] Faustina Johnson , Santosh Kumar Gupta," Web Content Mining Techniques: A Survey", *International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012.*

[11] Shesh Narayan Mishra ,Alka Jaiswal,Asha Ambhaikar ," An Effective Algorithm for Web Mining Based on Topic Sensitive Link Analysis ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012 ISSN: 2277 128X.

[12] V. Lakshmi Praba , T. Vasantha," EVALUATION OF WEB SEARCHING METHOD USING A NOVEL WPRR ALGORITHM FOR TWO DIFFERENT CASE STUDIES "Ictact Journal on Soft Computing, April 2012, Volume: 02, Issue: 03.

[13] Miguel Gomes da Costa, Júnior Zhiguo Gong," 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.

[14] Neelam Tyagi, Simple Sharma," Comparative study of various Page Ranking Algorithms in Web Structure Mining (WSM)" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-1, June 2012.

[15] Ms.N.Preethi , Dr.T.Devi," New Integrated Case And Relation Based (CARE) Page Rank Algorithm" 2013 International Conference on Computer Communication and Informatics (ICCCI - 2013), Jan. 04 – 06, 2013, Coimbatore, INDIA.