[VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY]
Vol.4. No. 1, July, 2015
77 - 82,ISSN : 0975-5446

Relative Study Of Various Web Structure Mining Algorithms

PANCHOLI Vishal R.
Assistant Professor,
Narmada College Of Science And Commerce
Bharuch
vishal.pancholi@yahoo.com

Dr. CHAUDHARI Vimal
Assistant Professor,
Department Of Computer Science
Veer Narmad South Gujarat University,Surat
vimal@vnsgu.ac.in

Abstract

With the rapid escalation of internet technology gradually, people rely on the search engines to explore the web. In this situation, the primary goal and also challenge for website owner is to provide appropriate information to users as per needs and fulfill their requirements. In order to achieve this goal the web mining concept is used. It is used to classify users and pages by analyzing users' activities, content of pages, and order of URLs that tend to be accessed in order. Web structure mining plays very vital role in this approach. It's defined as the process of analyzing the structure of hyperlink. There are many proposed algorithms for this such as PageRank, Weighted PageRank and Hyperlink-Induced Topic Search etc. This paper studied the introduction of web mining and its different techniques, a review of page ranking algorithms and comparison of some important algorithms in context of performance has been carried out
.

Keywords : Web Mining, PageRank, Weighted PageRank, HITS

1. Introduction

The *World Wide Web* is the collection of information resources on the Internet that are using the Hypertext Transfer Protocol. It is a repository of many interlinked hypertext documents, accessed via the Internet. Web may contain text, images, video and other multimedia data. In order to analyze such data, some techniques called web mining techniques are used by various web applications and tools[3]. In the extremely competitive world and with the extensive use of the Web in e-commerce, e-learning, and e-news, finding users' needs and providing useful information are the primary goals of website owners[4]. The Web is vast, varied and dynamic. The Web contains huge 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[2]. Web mining is used to discover the contents of web, the user's behavior in the past and the web pages that the users want to view in the future[4].

Web mining is categorized into three parts: 1) Web content mining (WCM) 2) Web structure mining (WSM) 3) Web usage mining (WUM). Web Content Mining deals with discovering useful information or knowledge from web page contents. Web content mining analyzes the content of web resources. Content data is the collection of facts that are contained in a web page. It consists of unstructured data such as free texts, images, audio, video, semi structured data such as HTML documents and a more structured data such as data in tables or database generated HTML pages[3]. Web structure mining discovers associations between web pages by analyzing web structures. 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[2].

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, WSM classify web pages and generates related patterns, such as the similarity and the relationships between different Web sites. Web Usage Mining determines user profiles and the users' behavior recorded within the web log file. 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[2].Technically, WCM focuses mainly on the structure within a document (the inner-document level) while WSM tries to discover the link structure of the hyperlinks between documents (the inter document level). The numbers of inlinks (links to a page) and of outlinks (links from a page) are important information in mining[4].

2. Web Page Ranking Algorithms

There are number of algorithms based on the link analysis. Out of them three algorithms Page Rank, weighted pager rank and HITS are discussed in this paper.

[A] Page Rank algorithm

This algorithm is used to determine the importance of website pages and it is developed by Brin and Page at Stanford University [11]. It works by counting the number and quality of links to determine a rough estimate of how important the website is. It allocates a numerical weight to each element of a hyperlinked set of documents. The link from one page to another page is considered as a vote. Not only the number of votes that a page receives is important but the importance of pages that casts the vote is also important. Page and Brin proposed a formula to calculate the page rank of a page A as stated below:

$$PR_{(u)} = \sum_{v \in Bu} \frac{PR(v)}{L(v)}$$

i.e. the page rank value for a page *u* is dependent on the pagerank values for each page *v* contained in the set B_u (the set containing all pages linking to page *u*), divided by the number

L(v) of links from page v.

Later on PageRank observed that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor.

PR(A)=(1-d)+d(PR(T1)/L(T1)+…..+PR(Tn/L(Tn))

Here PR (Ti) is the page rank of the pages Ti which links to page A, L(Ti) is number of outlinks on page Ti and d is damping factor. It is used to stop other pages having too much

influence. The total vote is "damped down" by multiplying it 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. 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.

[B] HITS (Hyper -link Induced Topic Search) algorithm

It is a link analysis algorithm that rates web pages developed by Jon Kleinberg [8]. It is also knows as Hubs and Authorities. A good hub represented a page that pointed to many other pages and a good authority represented a page that was linked by many different hubs. The twitter social network uses a HITS algorithm to suggest user accounts to follow.

This algorithm assigns two scores for each page:

1) Authority – estimates the value of the content of the page
2) Hub – estimates the value of its links to other pages

It is processed on a small subset of "relevant" documents, not all documents as was the case with PageRank. The first step of algorithm is to retrieve the most relevant pages to the search query which is called root set. The next step is to augment the root set with all the web pages that are linked from it and some of the pages that link to it which is known as base set.

The web pages in the base set and all hyperlinks among those pages form a focused subgraph. The HITS computation is performed only on this focused subgraph. 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.

[C] Weighted PageRank algorithm

It is develop by Wenpu Xing and Ali Ghorbani [4]. It is an addition of PageRank algorithm. PageRank and HITS algorithms treat all links equally when distributing rank scores. But this algorithm considers the importance of both inlinks and outlinks of the pages and distributes rank scores based on the popularity of the pages. The popularity from the number of inlinks and outlinks is recorded as $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$ respectively. $W^{in}_{(v,u)}$ is the weight of *link*(*v*, *u*) calculated based on the number of inlinks of page *u* and the number of inlinks of all reference pages of page *v*.

$$W^{in}_{(v,u)} = \frac{Iu}{\sum_{p \in R(v)} Ip}$$

Where I_u and I_p represent the number of inlinks of page *u* and page *p*, respectively. *R(v)* denotes the reference page list of page *v*.

$W^{out}_{(v,u)}$ is the weight of *link*(*v*, *u*) calculated based on the number of outlinks of page *u* and the number of outlinks of all reference pages of page *v*.

$$W^{out}_{(v,u)} = \frac{Ou}{\sum_{p \in R(v)} Op}$$

Where O_u and O_p represent the number of outlinks of page *u* and page *p*, respectively. *R(v)* denotes the reference page list of page v.

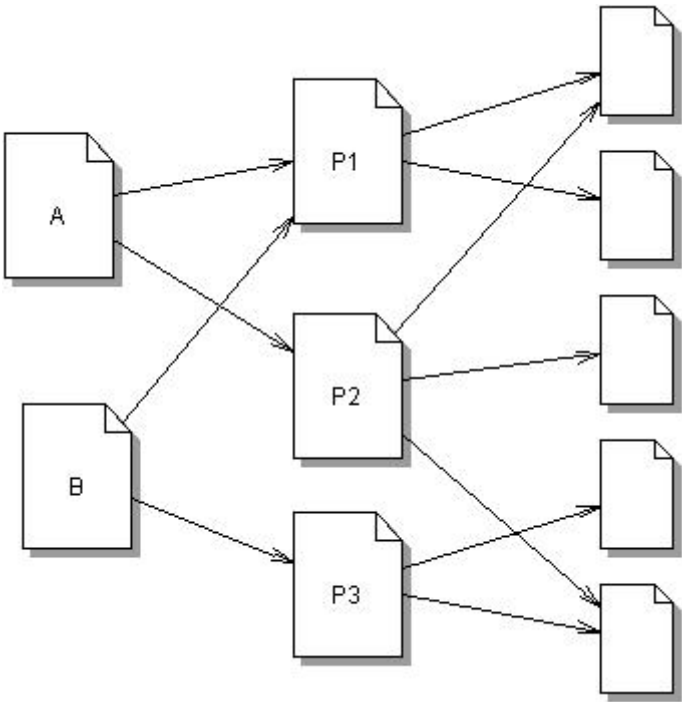

Figure 1 Example for page links

In this example, page A has two reference pages P1 and P2. The inlinks and outlinks of these two pages are I_{p1} =2, I_{p2}=1, O_{p1} =2, O_{p2}=3. So,

$$W^{in}_{(A, P1)} = \frac{Ip1}{(Ip1+Ip2)} = 2/3 \quad \text{and}$$

$$W^{out}_{(A, P1)} = \frac{Op1}{(Op1+Op2)} = 2/5$$

Considering the importance of pages, the original PageRank formula is modified as,

$$PR(u) = (1-d) + d \left(\sum_{v \in Bu} PR(v)\, W^{in}(v,u) W^{out}(v,u) \right)$$

Table 1 Comparison of various web page ranking algorithms [1][2]

Algorithm	**Page Rank**	**HITS**	**Weighted Page Rank**
Mining method	Web structure mining	Web structure mining, Web content mining	Web structure mining
Input parameter	Inlink	Content, inlink and outlink	Inlink and outlink
Working method	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.
Quality of result	Medium	Less than PageRank	Higher than PageRank
Advantages	It is a global measure and	It has the ability to	The pages are sorted

	query independent.	rank pages according to the query topic and provide more relevant Hub and authority values.	according to the importance.
Disadvantages	New pages have less page rank and they take much time to be getting listed and gain high ranks.	The query evaluation time is slow. Collecting the root set, expanding it and performing computation are all expensive tasks.	Relevancy is ignored
Search engine	Google	Twitter	Google

3. Conclusion

Web mining is the Data Mining technique that automatically finds out or extracts the information from web documents. Page Rank and Weighted Page Rank algorithms are applied in Web Structure Mining to rank the relevant pages. In this paper it has been stated the preface of web mining and its related techniques such as web content mining, web structure mining and web usage. The objective 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 more and more 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 categorize authoritative web pages.

Acknowledgement

We acknowledge UGC for Special Assistance Program (SAP) for Natural Language Processing and Data Mining (file number is F.3-48/2011) under which this work has been done.

References

[1] 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

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

[3] Gurpreet Kaur, Shruti Aggrawal "A SURVEY- LINK ALGORITHM FOR WEB MINING" Gurpreet Kaur et al , International Journal of Computer Science & Communication Networks,Vol 3(2), 105-110

[4] Wenpu Xing and Ali Ghorbani "Weighted Page Rank Algorithm" Faculty of Computer Science University of New Brunswick Fredericton, NB, E3B 5A3, Canada

[5] Naresh Barsagade, "Web Usage Mining And Pattern Discovery: A Survey Paper", CSE 8331, Dec.8,2003.

[6] N. Duhan, A. K. Sharma and K. K. Bhatia, "Page Ranking Algorithms: A Survey", Proceedings of the IEEE International Conference on Advance Computing, 2009.

[7] P Ravi Kumar, and Singh Ashutosh kumar, "Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval", American Journal of applied sciences, 7 (6) 840-845 2010.

[8] Kleinberg, “Hubs, Authorities and Communities”, ACM Computing Surveys, 31(4), 1999.
[9] Rashmi Rani, Vinod Jain ,” Weighted PageRank using the Rank Improvement” International Journal of Scientific and Research Publications, Volume 3, Issue 7, July 2013.
[10] 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.
[11] S. Brin, and L. Page, “The Anatomy of a Large Scale Hypertextual Web Search Engine”, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998