Survey on Web Structure Mining

Hiep T. Nguyen Tri, Nam Hoai Nguyen
Department of Electronics and Computer Engineering
Chonnam National University
Republic of Korea
Email: tuanhiep1232@gmail.com

***Abstract*—Nowadays, World Wide Web becomes huge information resource. Research engine also becomes popular tool that help user find needed information quickly. Because of the huge number of web sites and also web pages, search engines play an important role in these days. One of the main factors that create the difference of a search engine with other is the ranking mechanism (page rank algorithm). In this paper, we will summarize some prominent page rank algorithms. ANd then we will present our implementations that implement two page rank algorithms PageRank and Weighted PageRank.**

I. Introduction

Nowadays, World Wide Web(WWW) becomes huge information resource. Along with the development of WWW is the development of tools that support data mining, such as search tools like google, yahoo, etc. Because of the huge number of web pages, finding the useful information is not easy with users. Web search tools plays an important role in these days.

Basically, the operator of search engines is as follows. Firstly, they read the web pages, extract content of web pages and find the links. By that links the crawler can follow and process another web pages. After collecting the web pages, index module will parse the content of web pages and build index table based on the key words used in those web pages. When an user fire a search query, search engines will match the keywords in the query and in the index table and return the related web pages. Before return the results to the user, a ranking mechanism is executed to order the results and give the best web page order to the user. Nowadays, there are many search engines, the competition is very high. Therefore, ranking mechanism of search engines becomes the main factor that determines the success of search engines.

In this paper, we will survey some main ranking technique, and then we will present a study in that we try to implement some techniques with some real web page. The organize of this paper is as follows. Section II present the Web Mining and its classification. In section III, we will summary some prominent techniques. Section IV discuss the comparison between those techniques and few discussion. In section V, we will show our implementation of two techniques. Section IV concludes the paper.

II. Web Mining

Ranking mechanism is tightly related to web mining. Hereby, we will introduce Web Mining and its classificationin this section.

Web mining is the use of data mining techniques to automaticcally discover and extract the information of the WWW. Web Mining consists of the following task [1]:

- Resource finding: the taskof retrieving intended web documents.
- Information selection and pre-processing: automatically selecting and pre-processing specific information from obtained web resources.
- Generalization: automatically discovers general patterns at individual web sites as well as across multiple sites.
- Analysis: validation and/or interpretation of the mined patterns.

Resource finding is the process of retrieving the online or offline data which is text resource available on the web such as email, content of HTML document, etc. Information selection and pre-processing is the transformation task in order to get the main content of document. For example, HTML document is removed HTMl tag to get the main content of HTML document. In the third step, machine learning or data mining technique are usually used to discover the general pattern of single site or multiple sites. Analysis validate the mined pattern and might interpret it.

There are three type of Web mining: Web Content Mining, Web Usage Mining and Web Structure Mining. Web Content Mining is the process of extracting useful information from the contents of web document. Web Content Mining concern with the retrieval of information from content of WWW. Web Content Mining can be differentiate from two different views : Information Retrieval view and Database view. The goal of Web Mining from the Information Retrieval view is to assist or to improve the finding or filtering information. While the goal of Web Mining from Database view is trying to model data on the web and integrate them.

WEb Usage Mining is the process of extracting useful information from the secondary data derived from the interaction of user while interacting with the web. The web usage data includes web server access logs, user profiles, user session, etc.

Web Structure Mining is the process of generating the structural summary about the web site and web page. The challenge for Web Structure Mining is to focus on the hyperlink structure of the Web. In other words, Web Structure Mining is thought to be a process by which the model of link structures and web pages are discovered [2]. The ultimate

purpose of Web Structure Mining is to generate structural summary about the Web site and Web page. This model can be used to categorize web pages or generate useful information such as the relationship between the web sites. It is used to think that Web Structure Mining and Link analysis are one. With the growing interest of Web Mining, Web Structure Mining research is developing into different techniques. Inside itself, Web Structure Mining is categorized into different sub-categorizations according to other researchers[3], [4]. It is recommended Web Structure Mining to be categorized into two sub-categorizations: Document Structure Mining and Link Mining. While Link Mining is aimed to generate the information of Web pages, such as the similarity and relationship between different Web sites, Document Structure Mining opens another direction research: reveal the structure (schema) of Web pages in order to compare or integrate Web page schemes [5]. However in the scope of this study, the sub-category Link mining is concentrated on its techniques and issues.

Ranking mechanism can use the techniques of three categories above. But, almost techniques related to the Web Structure Mining to evaluate the importance of web pages. There are number of algorithms proposed to solve issues in Link mining topic. In the next part, four important algorithms: Pagerank algorithm, Weighted pagerank algorithm, Weighted content pagerank algorithm (WCPR), Hyperlink-Induced Topic Search (HITS) are discussed and compared to make it clear about their techniques and issues.

III. Page Ranking Algorithms

A. PageRank Algorithm

L.Page and S.Brin developed page rank algorithm named PageRank [6]. PageRank assumes that a page that has high rank if the sum of the ranks of its backlinks is high. It means that if a page has backlinks from the other high rank pages or has many backlinks will have high rank. PageRank algorithm is utilized by Google. After user request a search query, Google combines pre-computed static PageRank scores with content matching score to obtains an overall ranking score for each web page. The PageRank equation is defined in 1.

$$R_{(u)} = (1-d) + d \sum_{v \in B_{(v)}} \frac{R_{(v)}}{N_v} \tag{1}$$

here, u is a web page that we want to calculate rank score, $B_{(v)}$ is a set of web pages which point to u, R is PageRank score of a web page. N_v is number of web page which is pointed by web page v. d is a damping factor that can be thought of as the probability of user's following direct links i.e the web graph and is usually set to 0.85.

$$\begin{aligned} R_A &= (1-d) + d(\frac{R_B}{2}) \\ R_B &= (1-d) + d(\frac{R_A}{2} + R_C) \\ R_C &= (1-d) + d(\frac{R_A}{2} + \frac{R_B}{2}) \end{aligned} \tag{2}$$

In order to illustrate the working of PageRank algorithm. Let consider the example in the figure 1. The Rank score

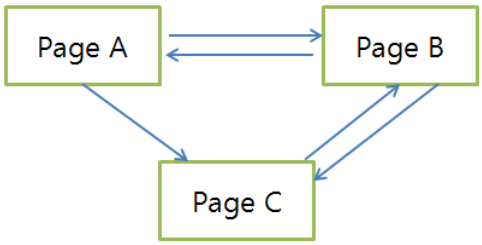

Figure 1. Hyperlink Structure for 3 pages

of each page are calculated as equation system 2. Solve the system of equation with d = 0.82 we can get R_A = 0.702, R_A = 1.298, R_C = 1. However, with the huge number of web pages in reality, solving the equation system is impossible. Another solution for calculating the Rank score is use iterative calculation. In this method, each page is assigned a starting rank value of 1 and then Rank score is iteratively calculated by new values. The method is illustrated in Table I.

Table I. PageRank Iteration Method

Iteration	R_A	R_B	R_C
1.0	1.0	1.0	1.0
0	0.575	1.425	1.0
1	0.756	1.244	1.0
2	0.679	1.321	1.0
3	0.711	1.289	1.0
4	0.698	1.302	1.0
5	0.704	1.296	1.0
6	0.701	1.299	1.0
7	0.702	1.298	1.0

The PageRank algorithm has two main features. Firstly, PageRanl considers 3 factors the rank of web pages that point to the web page, number of it's outgoing links and number of incoming link of the web page. Secondly, the convergence time could be large.

In order to improve the pageRank method, some other methods were proposed. In [7], Wenpu Xing and Ali Ghorbani extended PageRank algorithm named Weighted PageRabk Algorithm. We will discuss this algorithm in section III-B.

In [8], the authors proposed a new algorithm named to improve the performance of PageRank algorithm based on the optimized normalize technique. In each iteration, after rank of each page is recalculated. A mean value is calculated by dividing summation of rank of all web pages by the number of web pages. And then, the rank of each page is normalized by dividing previous rank by mean value. The algorithm is depicted in figure 2

B. Weigted PageRank Algorithm

Weighted PageRank Algorithm assigns larger rank values to more popular pages instead of dividing the rank value of a page among its outlink pages. The popular wep page is the more linkages that other web page tend to have to them or are linked to by them. The rank score is calculated as equation 3.

$$R_{(u)} = (1-d) + d \sum_{v \in B_{(v)}} R_{(v)} W^{in}_{(v,u)} W^{out}_{(v,u)} \tag{3}$$

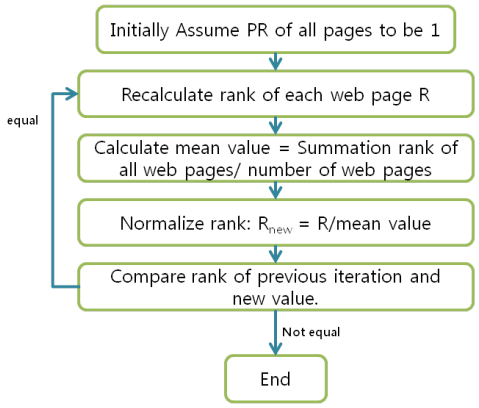

Figure 2. PageRank algorithm based on normalized technique

$W^{in}_{(v,u)}$ is the weight of link(v,u), which is calculated based on the number of inlinks of page u and the number of inlink of all reference pages of page v.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{v \in Rf_{(v)}} I_p} \tag{4}$$

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

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{v \in Rf_{(v)}} O_p} \tag{5}$$

Let's illustrate by example in figure 1.

$$W^{in}_{(B,A)} = \frac{I_A}{I_A + I_C} = \frac{1}{3}$$
$$W^{out}_{(B,A)} = \frac{O_A}{O_A + O_C} = \frac{2}{3}$$

Similarly, we have:

$$W^{in}_{(A,B)} = \frac{1}{2}; \quad and \quad W^{out}_{(A,B)} = \frac{2}{3}$$
$$W^{in}_{(B,C)} = \frac{2}{3}; \quad and \quad W^{out}_{(B,C)} = \frac{1}{3}$$
$$W^{in}_{(C,B)} = 1; \quad and \quad W^{out}_{(C,B)} = 1$$
$$W^{in}_{(C,A)} = \frac{1}{2}; \quad and \quad W^{out}_{(C,A)} = 1$$
$$W^{in}_{(A,C)} = \frac{1}{2}; \quad and \quad W^{out}_{(A,C)} = \frac{1}{3}$$

Solve the system equation we can get $R_A = 0.234$, $R_B = 0.284$, $R_C = 0.237$

C. Weighted content pagerank algorithm (WCPR)

Although Weighted Page Rank also takes the importance of the inlinks and outlinks of the pages, it was realized that the rank score to all links is not equally distributed, for example the unequal distribution is performed. Weighted content pagerank algorithm (WCPG) which based on web content mining and structure mining is introduced in order to shows the relevancy of the pages to a given query so it make users to be easily get the relevant and important pages in the list [9]. Weighted Page Content Rank Algorithm (WPCR) 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, e.g. how many pages are pointing to or are referred by this particular page. It can be calculated based on the number of inlinks and outlinks 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. The whole of this algorithm can be summarized as the two steps below: Input for the algorithm: Page P, inlink and outlink Weights of all backlinks of P, Query Q, d (damping factor). Output of the algorithm: Rank score Step 1: Relevance calculation:

a. Find all meaningful word strings of Q (say N)
b. Find whether the N strings are occurring in P or not?
c. Z = Sum of frequencies of all N strings.
d. S = Set of the maximum possible strings occurring in P.
e. X = Sum of frequencies of strings in S.
f. Content Weight (CW) = X/Z
g. C = No. of query terms in P
h. D = No. of all query terms of Q while ignoring stop words.
i. Probability Weight (PW) = C/D

Step 2: Rank calculation:

a. Find all backlinks of P (say set B).
b. Calculate ranks score as equation 6.
c. Output PR(P) as the Rank score

$$PR_{(u)} = (1-d) + d \sum_{v \in B_{(v)}} R_{(v)} W^{in}_{(v,u)} W^{out}_{(v,u)} (C_w + P_w) \tag{6}$$

D. Hyperlink-Induced Topic Search (HITS)

Hyperlink-Induced Topic Search (HITS) is a link algorithm, introduced by J. Klienberg [10], [11]. He mentioned webpages as two types: hubs and authorities. Hubs are the pages that act as resource lists. Authorities are pages having important contents. A good hub page is a page which is pointing to many authoritative pages on that content. A good authority page is a page which is pointed by many good hub pages on the same content. A page may be a good hub and a good authority at the same time.

Table II. WEB STRUCTURE MINING ALGORITHMS COMPARISON TABLE

Algorithm	PageRank	WPR	WPCR	HITS
Author/Year	S. Brin et al., 1998	Wenpu Xing et al, 2004	P. Sharmar et al., 2000	Jon Kleinberg, 1998
Mining Technique Used	WSM	WSM	WSM and WCM	WSM and WCM
Description	Computes scores at indexing time, not query time. Results are sorted according to importance of pages.	Assigns large value to more important pages instead of diving the rank value of a page evenly among its outlink pages	Gives sorted order to the web pages returned by a search engine as a numerical value in response to a user query	Computes hub and authority scores of n highly relevant pages on the fly. Relevant as well as important pages are returned.
Input / Output Parameters	Backlinks	Backlinks,Forward links	Backlinks,Forward links,Contents	Backlinks,Forward links,Contents
Complexity	O(logn)	¡O(logn)	¡O(logn)	¡O(logn)
Advantages	Providing important pages according to given query.	Providing important pages according to given query. Assigning importance in terms of weight values to incoming and outgoing links.	Providing important pages and relevant pages according to query by using web structure and web content mining	Providing more relevant authority and hub pages according to query
Limitation	Query independent	Query independent	Importance of page is ignored	Topic drift (topic unrelated to the original query)Cannot detect advertisements
Search Engine	Google	Research model	Research model	Clever

The HITS algorithm can be summarized two main steps as following: Input with a search topic, specified by one or more query terms. Step 1 - Sampling: A sampling component, which constructs a focused collection of several thousand Web pages likely to be rich in relevant authorities; and Step 2 - Weight propagation: A weight-propagation component, which determines numerical estimates of hub and authority weights by an iterative procedure. Outputs of HITS are hubs and authorities for the search. Some constraints of HITS algorithm are [11]:

a. Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.

b. Topic drift: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights.

c. Automatically generated links: HITS gives equal importance for automatically generated links which may not have relevant topics for the user query.

d. Efficiency: HITS algorithm is not efficient in real time

IV. WEB STRUCTURE MINING ALGORITHMS COMPARISON

See table II.

V. CASE STUDY: AN IMPLEMENTATION OF PAGERANK AND WEIGHTEDPAGERANK ALGORITHM

In order to understand clearly about PageRank algorithm and Weighted PageRank algorithm, we decide to implement them and apply it to reality web site. In this section we will present our implementation and the result. The website was chosen to analyse is ”altair.chonnam.ac.kr/ kbkim/”. In this section, the crawler we used crawler4j [12] an open source for Java language. The first experiment we compare the result of two algorithm. In this experiment we choose value of d is 0.85 and the threshold is 0.001. Table III shows 10 pages in which first 5 pages is top 5 page of page rank list that is obtained by PageRank algorithm and last 5 pages is top 5 pages obtained by Weighted PageRank algorithm. Figure 3 shows the relation between result of PageRank and Weighted PageRank. The order of x axis follows the rank order of PageRank algorithm. We can observe that, The difference between two algorithm results is very small in the left area. However, there is big difference between two results in the right area. It is because the difference of considering important page. PageRank considers three factors inlinks, source pages of inlink, outlink of source page. But Weighted Page Rank considers inlinks, source pages of inlinks, outlinks, relation between outlinks of source pages.

Table III. TOP TEN OF HIGH RANK PAGES

Short URL	PR		WPR	
	Score	Order	Score	Order
...3/overview-summary.html	25.942	1	2.651	4
...2/overview-summary.html	25.91	2	2.303	6
...3/deprecated-list.html	14.846	3	0.234	41
...3/help-doc.html	14.846	4	0.246	37
...3/index-files/index-1.html	14.846	5	0.725	19
...2/allclasses-frame.html	12.789	12	4.336	1
...3/allclasses-frame.html	12.308	14	3.898	2
.../package-summary.html	9.841	17	2.746	3
3/overview-summary.html	8.921	18	2.59	5
...game/package-summary.html	7.771	20	2.253	7

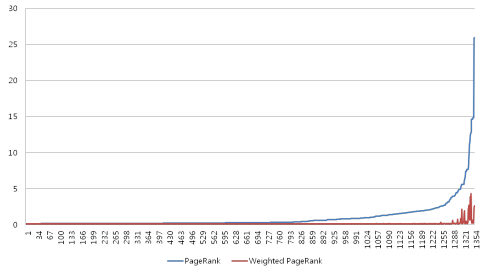

Figure 3. Relation Between PageRank Result and Weighted PageRank Result

In order to evaluate the performance of pagerank and weighted pagerank, we run two algorithms with different values of threshold and then we calculate the convergence

time and number of round. Figure 4 show the comparison of convergence time and iteration number of two algorithms. From the figure, we can observe that, the performance of Weighted Page Rank is better than PageRank. The number of iteration and the converennce time of Weighted PageRank is smaller than PageRank's.

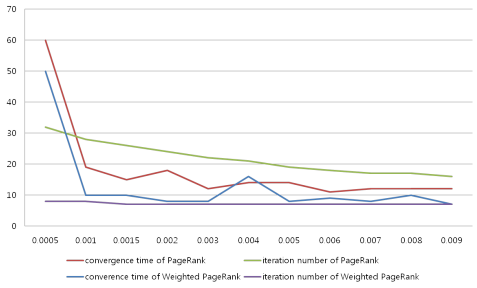

Figure 4. Comparison of convergence time and iteration number between two algorithms

In the next experiment, we run two algorithms with different values of d parameter and then we calculate the convergence time and number of round. Figure 5 show the comparison of convergence time and iteration number of two algorithms. The same result with the previous experiment. The performance of Weighted Page Rank is better than pageRank. And we can observe that the number of iteration get smaller when d parameter is smaller.

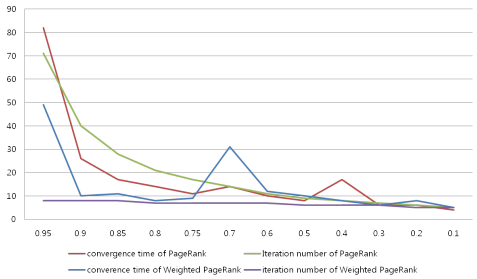

Figure 5. Comparison of convergence time and iteration number between two algorithms

VI. Conclusion

In this study, Web Structure Mining was reviewed in the relationship to Web Mining. Also, the two sub-categories of Web Structure Mining were asserted through literature analysis work. The four important algorithms in Web Structure Mining were examined and summarized condensed to make it easily understand the concept of those algorithms. The most important part of this study is the tabular comparison content about four important techniques used in Web Structure Mining. The comparison content was examined carefully through literature analysis work, and then it was tested by comparing the result gained from an implementation in practice some of those techniques. However, this study still has some limitation. We did not have enough time to implement all of those techniques to test the comparison content empirically. One thing we still missed in this study when we still could not contribute any novelty to those techniques.

References

[1] R. Kosala and H. Blockeel, "Web mining research: A survey," *SIGKDD Explor. Newsl.*, vol. 2, no. 1, pp. 1–15, Jun. 2000. [Online]. Available: http://doi.acm.org/10.1145/360402.360406

[2] T. Bhatia, "Link analysis algorithms for web mining," *IJCST*, vol. 2, no. 2 2 2 9 - 4 3 3 3, pp. 243–246, jun 2011.

[3] *Web structure mining: an introduction*, 2005. [Online]. Available: http://dx.doi.org/10.1109/icia.2005.1635156

[4] M. A. Preeti Chopra, "A survey on improving the efficiency of different web structure mining algorithms," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 2, no. 2249 8958, pp. 296–298, feb 2013.

[5] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim, "Research issues in web data mining." in *Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99*, 1999, pp. 303–312. [Online]. Available: http://www.springerlink.com/content/yar775kx05pgnj93

[6] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," in *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, 1998, pp. 161–172. [Online]. Available: citeseer.nj.nec.com/page98pagerank.html

[7] "Weighted pagerank algorithm," in *Proceedings of the Second Annual Conference on Communication Networks and Services Research*, ser. CNSR '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 305–314. [Online]. Available: http://dl.acm.org/citation.cfm?id=998669.998911

[8] H. Dubey and P. B. N. Roy, "An improved page rank algorithm based on optimized normalization technique," pp. 2183–2188, 2011.

[9] P. Sharma and P. Bhadana, "Weighted page content rank for ordering web search result," *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7301–7310, 2010.

[10] J. M. Kleinberg, "Hubs, authorities, and communities," *ACM Comput. Surv.*, vol. 31, no. 4es, Dec. 1999. [Online]. Available: http://doi.acm.org/10.1145/345966.345982

[11] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg, "Mining the web's link structure," *Computer*, vol. 32, no. 8, pp. 60–67, 1999.

[12] [Online]. Available: https://code.google.com/p/crawler4j/