Article can be accessed online at http://www.publishingindia.com

Research on Ranking Algorithms in Web Structure Mining

Suvarna Sharma*, Amit Bhagat**

Abstract

In the past few decades, the Web has emerged as a treasure of information and web mining is a technique to handle this treasure. During recent years web mining has been a well-researched area. Web mining is the application of the data mining which is useful to extract the knowledge from web. With the progress of web, more and more data are now available for users on web. Web structure mining deals with the contents and hyperlinks on web pages.In this review paper, we have focused on three basic algorithms for evaluating the importance of pages i.e. Page Rank, Weighted Page Rrank, and Hyperlink-Induced Topic Search and comparison of those algorithms. PageRank algorithm is based on back links of the page and it calculates the rank of web pages at indexing time. Weighted Page Rank algorithm scores pages according to their relevancies and rank of a page is calculated by its number of incoming and outgoing links. Hyperlink-induced topic search algorithm is an iterative algorithm developed to quantify each page's value as an authority and as a hub. This study was done basically to explore the link structure algorithms for ranking pages.

Keywords: Web Mining, Web Structure Mining, Page Rank, Weighted Page Rank, Hyperlink-Induced Topic Search

Introduction

The Web is not only a network where users can receive information, buy wares, and receive services. Due to the rapidly growing size of the information on the web, users are now barely able to access the information without the help of search engines.

In recent years, a number of papers (Eirinaki, 2004; Kosala, & Blockeel, 2000; Liu, Chen, & Song, 2001; da Costa Júnior, & Gong, 2005) have considered the use of web mining to automatically discover and extract information from web documents and services. In particular, these papers point out challenges and applications regarding the term web mining and three axes of web mining, namely web structure, web content, and web usage mining. Web mining is to apply special mining techniques on web data, because data on the web is different from traditional data. One way to apply web mining techniques is web searching. Searching on web means finding out most relevant and effective documents for users. In web structure mining this work is done by ranking algorithm (Duhan, Sharma, & Bhatia, 2009; Kleinberg, 1999; Page, Brin, Motwani, & Winograd, 2002; Xing, & Ghorbani, 2004; Yan, Wei, Gui, & Chen, 2011; https://en.wikipedia.org), different ranking algorithms are available which are different to each other because of parameters and methods they use. In this paper, we have discussed some most popular algorithms.

PageRank Algorithm: This algorithm works on the basis of number of in-links and out-links to a web page. Initially, the page rank is set to 1 for all. A page's rank is calculated recursively upon the rank of the pages that point to it, called in-links.

Weighted PageRank Algorithm: This algorithm assigns a larger rank to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance.

* Research Scholar, Mathematics and Computer Applications Department, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India. Email: suvarnasharma.mtech@gmail.com

** Assistant Professor, Mathematics and Computer Applications Department, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India. Email: am.bhagat@gmail.com

Hyperlink-Induced Topic Search Algorithm: The HITS algorithm works on the basis of relevant topic search. The algorithm classifies web pages into two types in order to calculate page scores.

SimRank: If a page provides trustworthy information for a given topic, then it is classified as authority pages. The web pages that give links to authority pages are called hub.

This paper shows how mining techniques can be applied to the web structure. It is organised into three sections as follows. Second section discusses the brief introduction of web mining categories. Third section discuses web structure mining. Fourth section reviews some well-known algorithms for web structure mining and last section concludes this paper.

Web Mining: Overview

Web mining is the new mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web data. This is more powerful and efficient than existing techniques. It consists of the following tasks (Kosala, & Blockeel, 2000):

1. Resource finding: It is the process by which users extract the data either from online or offline text resources available on web.
2. Information selection and pre-processing: This process formats the structure of the original retrieved data in terms of web mining algorithm.
3. Generalisation: It includes discovering usual patterns from individual websites as well as across multiple sites automatically.
4. Analysis: It involves the validation and interpretation of the discovered patterns. It plays an important role in pattern mining.

In general, web mining is divided into three categories:

Web Content Mining: Web Content Mining (WCM) focuses on the discovery of useful information from the contents or data or services available on the web.

Web Structure Mining: Web Structure Mining (WSM) deals with the structure of hyperlinks within the web itself. Structure represents the graph of the links in a site or between the sites.

Web Usage Mining: Web Usage Mining (WUM) tries to discover the useful information from the interactions of the users while surfing on the web.

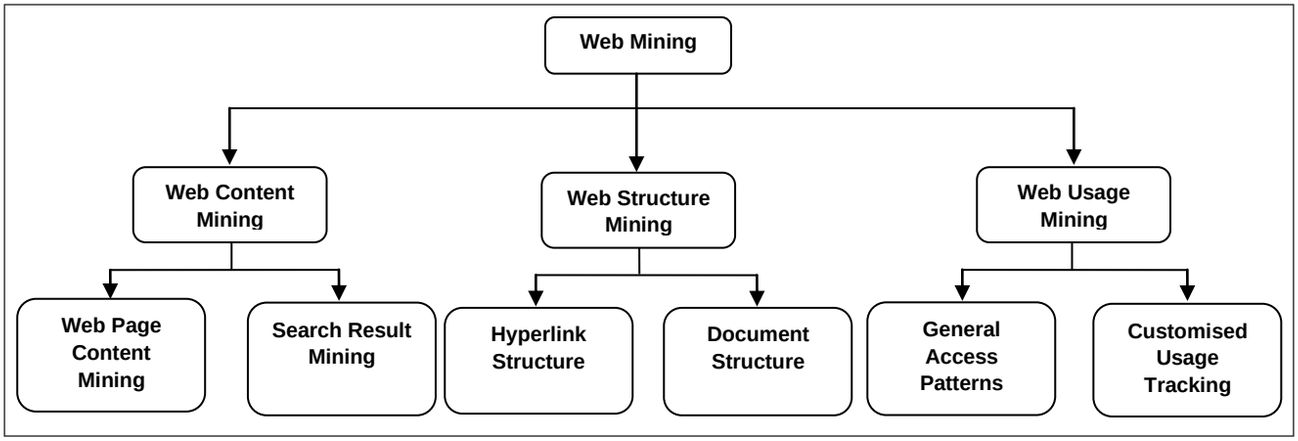
Web Structure Mining

The goal of the web structure mining is to generate the structural summary about the website and web page. Web structure mining is based on hyperlink analysis to analyze the dependencies between pages to find out relationship between the references pages to discovered interesting patterns, improve search engine optimisation.

Web structure mining uses the hyperlink structure of the web as an information source. It helps the users to retrieve the relevant documents by analysing the link structure.

This can be divided into two kinds based on the type of structure information used:

Fig. 1: Web Mining Classification



Hyperlink Structure: It is related to extract information from hyperlink in the web and design structure through analysing hyperlinks that connects the web pages.

Document Level Structure: It is related to mining the document structure and design tree like structure of page to describe the use of HTML or XML tag.

The web contains different kinds of objects as web pages, and links are in-out- and co-citation.

Link mining is broadly categorized into eight tasks that focus on objects, links or, graphs (Getoor, &Diehl, 2005):

1. Object-Related Tasks
   (a) Link-Based Object Ranking
   (b) Link-Based Object Classification
   (c) Object Clustering (Group Detection)
   (d) Object Identification (Entity Resolution)
2. Link-Related Tasks
   (a) Link Prediction
3. Graph-Related Tasks
   (a) Sub-graph Discovery
   (b) Graph Classification
   (c) Generative Models for Graphs

Algorithms

The network of web pages and links defines a directed graph G with web pages as nodes and links as edges connecting between pages. Standard definitions related to graph theory useful for algorithm are as follows (https://en.wikipedia.org):

Definition 4.1: A **directed graph** is **graph**, i.e., a set of objects (called vertices or nodes) that are connected together, where all the edges are **directed** from one vertex to another. A **directed graph** is sometimes called a digraph or a **directed** network.

Definition 4.2: The number of head endpoints adjacent to a node is called the ***indegree*** of the node.

Definition 4.3: The number of tail endpoints adjacent to a node is its ***outdegree*** .

Definition 4.4: An **adjacency matrix** of a graph described by an **n x n** matrix **A**, where **n** is the number of pages and $A_{ij} = 1$ if there is a link from page i to page j and otherwise $A_{ij} = 0$.

The commonly used algorithms in web structure mining, to access the web pages or websites effectively, have been discussed as follows:

Page Ranking

Page *et al*.developed a ranking algorithm used by Google, named PageRank (PR) after Larry Page (cofounder of Google search engine), that uses the link structure of the web to determine the importance of web pages.

PageRank algorithm is based on back links of the page. This algorithm calculates the rank of web pages at indexing time. PageRank algorithm computation of the web page score depends on the number of pages that are linking to it. These links are called as back-links. If there is more number of back-links to a web page it is of high importance. The key idea is that a page has high rank if it is pointed to by many highly ranked pages. So the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively till the rank of all the pages is determined.

A simplified version (Kleinberg, 1999) of PageRank is defined in Eq. 1:

$$PR(u) = c \sum_{V \in B(u)} \frac{PR(V)}{N_V} \quad (1)$$

where *u* represents a web page, *B(u)* is the set of pages point to *u*, *PR(u)* and *PR(v)* are rank scores of page *u* and *v* respectively, N_v denotes the number of outgoing links of page *v*, *c* is a factor used for normalisation.

Later this algorithm was modified and new version of PageRank is given in Eq. 2.

$$PR(u) = (1 - d) + d \sum_{V \in B(u)} \frac{PR(V)}{N_V} \quad (2)$$

where *d* is the damping factor and denotes the probability of deriving a particular node through a random traverse. A standard value for d is 0.85. Conversely, (1 - d) is the probability to stay on the current web page.

Demerits (Sangeetha & Joseph, 2015)

- **Rank Sink:** When pages are connected to each other forming a loop then computing PageRank for those pages is difficult.

- **Topic Bias:** Does not consider the web page content and the subject of the query.
- **Query Independent:** All PageRanks are pre-computed and thereby is not dependent on the query.
- Irrelevant Pages can also gain higher rank because of many numbers of in-links.
- PageRank algorithm emphasizes more on old pages.

Weighted PageRank Algorithm

Weighted PageRank is an extension of the PageRank algorithm. Weighted PageRank was introduced by Xing and Ghorbani (2004). This algorithm scores rank values to pages according to their relevancies rather than dividing it evenly. The rank of a page is calculated by its number of incoming and outgoing links.

The weighted PageRank algorithm considers the entire web as a graph in which the vertices are taken as V, and edges are taken as E. $W^{in}(v,u)$ is the weight evaluated by the number of incoming links of page u and the number of incoming links of all pages those are pointed to v and it assigns to *link* (v, u) as in-weight.

$$W^{\text{in}}(v,u) = \frac{I_u}{\sum_{p \in R(V)} I_P} \tag{3}$$

where I_u and I_P represent the number of in-links on the page u and page p respectively,

$R(v)$ denotes the reference page list of page v,

$W^{out}(v,u)$ is the weight evaluated by the number of outgoing links of page u and the number of outgoing links of all pages those are pointed to v and it assigns to *link*(v, u) as out-weight.

$$W^{\text{out}}(v,u) = \frac{O_u}{\sum_{p \in R(V)} O_P} \tag{4}$$

where O_u and O_P represent the number of outlinks on the page u and page p respectively,

$R(v)$ denotes the reference page list of page v.

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

$$PR(u) = (1-d) + d \sum v \in B(u)\, PR(v)\, W^{\text{in}}(v,u)\, W^{\text{out}}(v,u) \tag{5}$$

where d is the damping factor and same as the PageRank.

Demerits (Sangeetha, & Joseph, 2015)

- Certain irrelevant pages to the query are also included in the result set because of their popularity
- Topic Drift can occur.
- Weight-age distributed proportionately to pages based on popularity.
- Query independent.

Hyperlink-Induced Topic Search Algorithm

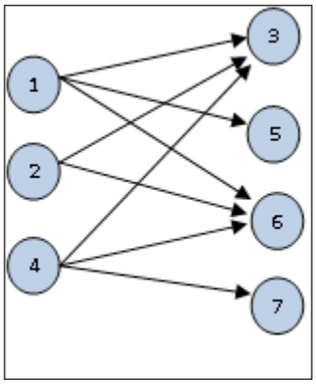
The HITS algorithm was introduced by Kleinberg (1999). The principle behind this algorithm is that a web page performs in two intentions either provide information on a topic or provide links to pages those contain information regarding that topic. This gives rise to two ways of categorising a web page.

Definition 4.5: A node is referred an **Authority** if it has many pages that link to it (i.e. it has a high indegree).

Definition 4.6:A node is referred a **Hub** if it points to many other nodes (i.e., it has a high outdegree).

The HITS algorithm iteratively performed operations for evaluating an authority and a hub value for each page. These concepts are explained in Fig. 2.

Fig. 2: Hub and Authority



In Fig. 2, nodes 1 and 4 are good hubs and 3 and 6 are good authorities.

The HITS algorithm is not applied to the graph representing the whole web, but rather to a subgraph derived from traditional text matching of the query terms in the search topic.

The HITS algorithm works in two major steps:

i) Sampling: The relevant pages are gathered based upon the given query.

ii) Iterative Step: The relevant pages that are found in the basis step are used in this iterative step. Hubs and Authorities are determined in these pages. This is done iteratively such that the Hub weight and the Authority weight results converge.

The Hub weight of a page p is given by h_p. The Authority weight of a page p is given by a_p. Initially the values of h_p and a_p are set to be 1. Then its values are calculated based upon the iterative formula.

$$h_p = \sum_{\forall q : p \to q} a_q \quad (6)$$

$$a_p = \sum_{\forall q : q \to p} h_q \quad (7)$$

The Hub weight is the sum of all authority weights that is pointed by the Huband the Authority weight is the sum of all hub weights that points to the authority.

Demerits (Sangeetha, & Joseph, 2015)

- **Mutual relationship:** Mutual relationship between hubs and authorities can lead to erroneous weights.
- **Hub and Authorities:** Cannot easily determine whether a page is a hub or authority.
- **Topic Drift:** May not produce relevant document as the algorithm weights all pages equally.
- **Efficiency:** Not efficient in real time.

Computation of Algorithms

Illustration of Ranking Algorithm

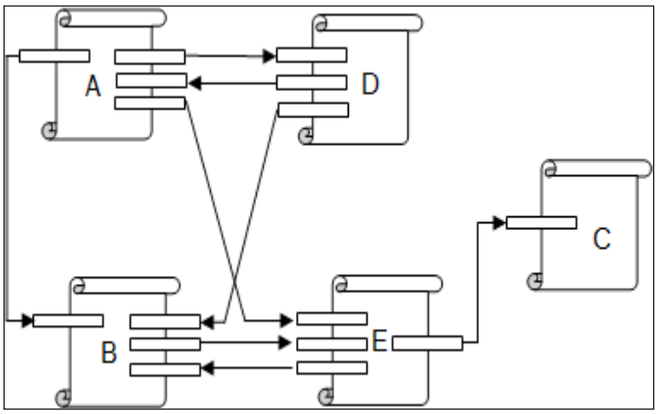
Let us consider a simple graph as shown in Fig.3.

$$M = \begin{vmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{vmatrix}$$

Out degrees=[3 2 0 1 2]

In degrees=[1 2 1 2 2]

Fig. 3: Simple Grapgh The Adjacency Matrix of the Graph is



PageRank Algorithm

By equation (2) PR of pages A, B,C, D and E are as follows:

$$PR(A) = .15 + .85\left(\frac{PR(D)}{N_D}\right)$$

$$PR(B) = .15 + .85\left(\frac{PR(A)}{N_D} + \frac{PR(E)}{N_E}\right)$$

$$PR(C) = .15 + .85\left(\frac{PR(E)}{N_E}\right)$$

$$PR(D) = .15 + .85\left(\frac{PR(A)}{N_A} + \frac{PR(B)}{N_B}\right)$$

$$PR(E) = .15 + .85\left(\frac{PR(A)}{N_A} + \frac{PR(B)}{N_B}\right)$$

Process stopped after 16th iteration because ranking of all nodes are similar as previous iteration. Node A has higher rank according to above calculation.

Weighted PageRank

For 1st iteration: By equation (5)

$WPR(A) = (1 - d) + d\ (WPR(D)\ W^{in}_{(D,A)}\ W^{out}_{(D,A)})$

By equation (3) and (4):

$W^{in}(D,A) = I_A/I_A = 2/2=1$

$W^{out}(D,A) = O_A/(O_B + O_D+O_E) = 2/(2+1+2)=2/5=0.4$

Similarly WPR of pages B,C,D and E are as follows:

Table 1: Iterative PageRanking of Nodes for Graph (Fig. 3)

NodesIteration	A	B	C	D	E
1	1.0	1.0	1.0	1.0	1.0
2	0.879	0.858	0.575	0.858	0.798
3	0.799	0.738	0.489	0.764	0.713
4	0.736	0.679	0.453	0.69	0.665
5	0.7	0.641	0.433	0.647	0.631
6	0.678	0.617	0.418	0.621	0.611
7	0.663	0.602	0.41	0.604	0.598
8	0.655	0.592	0.404	0.594	0.589
9	0.649	0.586	0.4	0.587	0.585
10	0.646	0.583	0.399	0.583	0.582
11	0.644	0.58	0.397	0.581	0.58
12	0.642	0.579	0.396	0.579	0.579
13	0.641	0.578	0.396	0.578	0.578
14	0.64	0.577	0.396	0.577	0.577
15	0.64	0.577	0.395	0.577	0.577
16	0.64	0.577	0.395	0.577	0.577

$$WPR(B) = (1 - d) + d\,(WPR(A)\,W^{in}_{(A,B)}\,W^{out}_{(A,B)} + WPR(E)\,W^{in}_{(E,B)}\,W^{out}_{(E,B)})$$

$$WPR(C) = (1 - d) + d\,(WPR(E)\,W^{in}_{(E,C)}\,W^{out}_{(E,C)})$$

$$WPR(D) = (1 - d) + d\,(WPR(A)\,W^{in}_{(A,D)}\,W^{out}_{(A,D)} + WPR(B)\,W^{in}_{(B,D)}\,W^{out}_{(B,D)})$$

$$WPR(E) = (1 - d) + d\,(WPR(A)\,W^{in}_{(A,E)}\,W^{out}_{(A,E)} + PR(B)\,W^{in}_{(B,E)}\,W^{out}_{(B,E)})$$

Process stopped after 8th iteration because ranking of all nodes are similar as previous iteration. Order of ranking is similar as PageRank Algorithm but it evaluated in less iterations.

HITS (Hyperlink-Induced Topic Search) Algorithm

Transpose Matrix of M is

$$M^T = \begin{vmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{vmatrix}$$

Table 2: Iterative PageRanking of Nodes for Graph (Fig.3)

Nodes Iteration	A	B	C	D	E
1	1.0	1.0	1.0	1.0	1.0
2	1.0	0.83	0.15	0.324	0.499
3	0.425	0.481	0.15	0.242	0.334
4	0.356	0.38	0.15	0.224	0.298
5	0.34	0.357	0.15	0.22	0.29
6	0.337	0.353	0.15	0.219	0.288
7	0.336	0.351	0.15	0.219	0.288
8	0.336	0.351	0.15	0.219	0.288

Assume that initial hub weight is $H = \begin{vmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{vmatrix}$

Hub

$H = M*A$

Authority

$A = M^T*H$

Table 3: Hub

A	*1*	*6*	*6*
B	1	4	4
C	1	0	0
D	1	1	1
E	1	3	3

Table 4: Authority

A	*1*	*2*	*0*
B	2	11	0
C	1	4	0
D	2	12	0
E	2	12	0

This already corresponds to our intuition that nodes B, C, D, and E are the most authoritative whereas node A is a good hub. Process stopped because we found same hub values, if we repeat the process further, we will obtain same values for hub and authority.

Conclusion

This is the survey paper of ranking algorithms in web structure mining, focusing on the introduction of web mining and taxonomy of web mining. We also discussed web structure mining, and introduced link mining. It also reviewed various pageranking algorithms i.e. PageRank, Weighted PageRank and Hyperlink-Induced Topic Search, used for web structure mining and shows how the ranking of pages is done. These algorithms are based on links and/or contents of the web pages. Since this is an upcoming area, there is a lot of potential for research. We are sure this paper would be a useful starting point for identifying opportunities for future research. Further the result of these algorithms would improve by using other factors present in the web page.

Table 5: Comparison of Ranking Algorithm

Algorithm Criteria	*PageRank*	*Weighted PageRank*	*HITS*
Area	Web Structure Mining	Web Structure Mining	Web Structure Mining, Web Content mining
I/P Parameters	In-links	In-links and out-links	In-links,out-links and query contents
Model	Random Surfer Model	Random Surfer Model	Query dependent model
Method	Computes scores at indexing time	Computes scores at indexing time	Computes hub and authority scores of n highly relevant pages on the fly
Accuracy of result	Medium	Higher than PR	Less than PR
Limitations	Query Independent	Query Independent	Not applicable in real time

References

da Costa Júnior, M. G. & Gong, T. Z. (2005). *Web structure mining: An introduction*. Proceedings of the 2005 IEEE International Conference on Information Acquisition.

Duhan, N., Sharma, A. K., & Bhatia, K. K. (2009). *Page ranking algorithms: A survey*. 2009 IEEE International Advance Computing Conference (IACC 2009).

Eirinaki, M. (2004). Web Mining: A Roadmap. Department of Informatics Athens University of Economics and Business.

Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *ACM SIGKDD Explorations Newsletter*, December, 7(2), 64-71.

Kleinberg, J. K. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, September, 46(5), 604-632.

Kosala, R., & Blockeel, H. (2000). Web Mining Research: A Survey. *ACM SIGKDD Explorations Newsletter*, July, 2(1), 1-15.

Liu, L., Chen, J., & Song, H. (2001). *The research of web mining*. Proceedings of the 4th World Congress on Intelligent Control and Automation, June, (pp. 10-14).

Page, L., Brin, S., Motwani, R., & Winograd, T. (2002). The Pagerank Citation Ranking: Bringing order to the Web. Technical Report, Stanford Digital Libraries.

Ridings, C., & Shishigin, M. (2002). *Pagerank Uncovered.* Technical Report, 2002.

Sangeetha, M., & Joseph, K. S. (2015). *Page ranking algorithms used in web mining* .ICICES 2014 - S. A. Engineering College, Chennai, Tamil Nadu, India.

Shivakumar, B. L. & Mylsami, T. (2014). Survey on web structure mining. *ARPN Journal of Engineering and Applied Sciences*, October, 9(10), 1914-1923.

Xing, W., & Ghorbani, A. (2004). *Weighted pagerank algorithm*. Technical report, 2 Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04) 2004 IEEE.

Yan, L., Wei, Y., Gui, Z., & Chen, Y. (2011). *Research on pagerank and hyperlink-induced topic search in web structure mining*. International Conference on Internet Technology and Applications.

Directed Graph. Retrieved from https://en.wikipedia.org.