The Pharma Innovation Journal 2023; SP-12(7): 1124-1128

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2023; SP-12(7): 1124-1128

www.thepharmajournal.com
Received: 12-05-2023
Accepted: 23-06-2023

Laxmi Choudhary
Department of Computer Science, Banasthali Vidyapith, Radha Kishnpura, Rajasthan, India

Rekha Jain
Department of Computer Science, Banasthali Vidyapith, Radha Kishnpura, Rajasthan, India

Corresponding Author:
Laxmi Choudhary
Department of Computer Science, Banasthali Vidyapith, Radha Kishnpura, Rajasthan, India

Various link algorithms in web mining

Laxmi Choudhary and Rekha Jain

Abstract
Web is a huge, massive, explosive, diverse, dynamic and mostly unstructured data repository. When user requests for a query on web then provide the relevant information to users for fulfill their needs. Everyone can store and retrieve the information from web. Extracting the important information from web is called web mining. Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data. It uses many data mining techniques because of heterogeneity and semi-structured or unstructured nature of the Web data so can easily improve the web services in fast way. Web mining is used to categorize users and pages by analyzing user's behavior, the content of pages and order of URLs accessed and then describe Web Structure mining. Web mining has three components: web content mining, web usage mining, web structure mining. This paper focuses on link Ranking algorithms in web structured mining and compares those algorithms which are used for information retrieval on web. Link algorithms are Page Rank (PR), WPR (Weighted Page Rank), HITS (Hyperlink-Induced Topic Search), Distance Rank, Eigen Rumor, and Dirichlet Rank algorithms.

Keywords: Web Mining, HITS, Page Rank, Weighted PageRank, WCM, WSM and WSM

1. Introduction
The World Wide Web is a popular and interactive medium to rich source of information today and continues to expand in size and complexity. Retrieving of the required web page on the web, efficiently and effectively, is becoming a Challenge. Whenever a user wants to search the relevant pages, he/she prefers those relevant pages to be at hand. The main aim of web mining is to extract the knowledge or information from the web. Web mining can be easily executed with the help of other areas. Web mining lies in between and copes with semi-structured data and/or unstructured data. Web mining calls for creative use of data mining and/or text mining techniques and its distinctive approaches. Mining the web data is one of the most challenging tasks for the data mining and data management. It includes data mining, machine learning, natural language processing, statistics, databases, information retrieval etc., there are several problems in web mining, and they are:

- To find relevant information
- To create new knowledge out of available information on web
- To personalize the information
- To learn about the customers

There are some challenges of web mining

- Web is huge
- Web pages are semi structured
- Web information stands to be diversity in meaning
- Degree of quality of the information extracted
- Conclusion of knowledge from information extracted.

Web Content Mining focuses on the discovery/retrieval of the useful information from the web contents/data/documents. For example, we can automatically classify and cluster web pages according to their topics.
These tasks are similar to those in traditional data mining. Web Structure Mining emphasizes to the discovery of how to model the underlying link structures of the web. For example, from the links, we can discover important web pages, which is a key technology used in search engines. We can also discover communities of users who share common interests. Web Usage Mining is relative independent, but not isolated, category, which mainly describes the techniques that discover the user‟s usage pattern and try to predicate user‟s behaviors. One of the key issues in Web usage mining is the pre-processing of click-stream data in usage logs in order to produce the right data for mining.

2. Web Mining Overview

In 1996, Etzioni who first invented the term web mining. That denotes the extension of data mining techniques to extract information from web resources, and uncover general patterns on the web. Etzioni starts by making a hypothesis that information on web is sufficiently structured and outlines the subtasks of web mining. According to Oren Etzioni and also it is based on the discovery of knowledge from the web. Web mining is to extract the knowledge in a proper way. It is useful to extract information, text, audio, video and multimedia document. We search any data or topic in the web and get accurate data about that topic. The main goal of the web mining is include the improvement of web design and structure and generation of dynamic recommendation. Web is huge and widely distributed data, all the information are interconnected in the repository. It provides some information service such as news, advertisements, customer information, financial management, education, government and e-commerce etc.

Web mining is categorized into three streams. They are:

a. Web Content Mining,
b. Web structure mining
c. Web usage mining

The web mining research relates to several research communities such a Database, Neural Networks, Information Retrieval and Artificial Intelligence.

The requirement of web mining is used to store information on World Wide Web (WWW). The information is growing rapidly on the web so this gives what users want.

2.1 Web Mining Process

The complete process of extracting knowledge from Web data is follows in Fig.1:

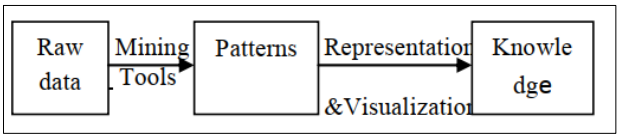

Fig 1: Web Mining process

A decomposition of Web mining in the following tasks:

1. **Resource finding:** It is the task of retrieving intended web documents.
2. **Information selection and pre-processing:** Automatically selecting and pre- processing specific from information retrieved Web resources.
3. **Generalization:** Automatically discovers general patterns at individual Web site as well as multiple sites.
4. **Analysis:** Validation and interpretation of the mined patterns.

2.2 Web Mining Categories

There are three main areas of web mining. Patterns followed by the users are evaluated by these three techniques of Web Mining and then these patterns are analyzed to get a user desired output. The categorization of web mining is found in fig. 2. The incorporating data mining to web-page ranking helps web search engines to find high quality web pages and enhances web click stream analysis, data semantics could substantially enhance the quality of keyword-based searches and indicate research problems (tremendous number of documents have not been indexed, which makes searching the data contains extremely difficult) to use data mining effectively in developing web intelligence.

It latter includes mining web search-engine data and analyzing web's link structure, classifying web documents automatically, mining web page semantic structures and page contents, and mining web dynamics.

These are classified in 3 categories

a. Web Content Mining
b. Web Usages mining
c. Web Structure Mining

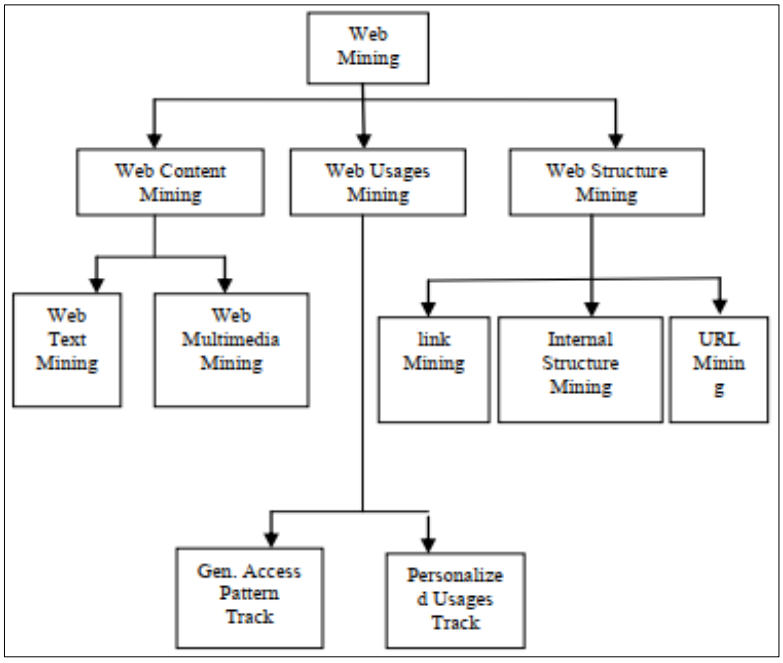

Fig 2: Web Mining Categories

2.2.1 Web Content Mining

Web content mining extracts or mines meaning information or knowledge from web page contents. For example, we can automatically classify and cluster web pages according to their topics. Web mining process is similar to those in traditional data mining. Similarly, we can also find patterns in web pages to extract meaning data such as descriptions of products, postings of forums, etc. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It can be text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues find out in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages.

Web/Text mining targets are classified into three major application categories:

1. Un-Structured documents-Text
2. Semi-Structured documents- HTML
3. Structured documents-XML

It is very typical task to discover appropriate knowledge from image and multimedia web contents.

2.2.2 Web Usage Mining

Web usage mining refers to the discovery of user access patterns. It applies many data mining algorithms. One main issues in Web usage mining is the pre-processing of click stream data in usage logs in order to produce the right data for mining.

The term web usage mining was introduced by Cooley *et al*. in 1997 [4] and in according with their definition: web usage mining is the automatic discovery of user access patterns from web servers. This information takes as input the usage data i.e. the data residing in the web server logs, recording the visits of the users to a web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

a) Web Server Data
b) Application Server Data
c) Application Level Data

2.2.3 Web Structure Mining

Web structure mining is the process of using graph theory to analyze the node and connection (hyperlinks) structure of a website to find meaning knowledge from hyperlinks. The goal is to categorized the web pages and
generate the information such as the similarity and relationship between them, taking the advantage of their hyperlink topology.

Then it focuses on the identification of authorities. This can be further divided into two kinds based on the kind of structure information used:

A). Hyperlinks B). Document Structure

2.2.3.1 Types of Web Structure Mining

a. Web graph mining
b. Web information extraction
c. Deep Web mining

2.2.3.2 Web Structure Mining Techniques

a. Link based classification
b. Link based cluster analysis
c. Link type
d. Link strength
e. Link cardinality

2.2.3.3 Web Structure Mining Algorithms

a. Page rank algorithm
b. HITS algorithm
c. Weighted page rank algorithm
d. Distance rank algorithm
e. Eigen rumor algorithm

3. Link Analysis Algorithms

Searching the web involves two main steps: Extracting the pages relevant to a query and ranking them according to their quality. Ranking is essential as it helps the user looks for "quality" pages that are related to the query. Different metrics have been proposed to rank web pages according to their linked quality. Web mining technique provides the additional information through hyperlinks where different documents are connected. We can view the web as a directed labelled graph whose nodes are the documents or pages and edges are the hyperlinks between them. This directed graph structure is known as web graph. The frequently used algorithms in web structure mining, to access the webpage or website effectively for user in online. There are number of algorithms proposed based on link analysis.

3.1 PageRank Algorithm

This algorithm was invented by Brin and Page at Stanford University which extends the idea of citation analysis. PageRank provides a better way to calculate the importance or relevance of a web page than simply counting the number of pages that are linking to it (called as backlinks).

If a backlink comes from an important page, then that backlink is given a higher weighting than those backlinks comes from non-important pages. In a simple way, link from one page to another page may be considered as a vote. So Page and Brin proposed a formula to calculate the PageRank of a web page A. Assume any arbitrary page A has pages T1 to Tn pointing to it (incoming link). PageRank can be calculated by the following equation (1).

$$PR(A) = (1 - d) + d(PR(T1) / C(T1) +.. + PR(Tn / C(Tn)) \quad (1)$$

Where d: damping factor usually sets it to 0.85 C(A): is defined as the number of links going out of page A.

The Page Ranks form a probability distribution over the web pages, so the sum of all web pages' PageRank will be one. PageRank is calculated by a simple iterative algorithm and corresponds to the principal eigenvector of the normalized link matrix of the Web.

3.2 Weighted PageRank Algorithm

Weighted PageRank (WPR) algorithm is proposed by Wenpu Xing and Ali Ghorbani, Which is an extended form of PageRank algorithm. This algorithm assigns a greater rank values 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. The incoming and outgoing links have assigned weight values by their importance and are denoted as $W^{in}_{(m, n)}$ and $W^{out}_{(m, n)}$ respectively. $W^{in}_{(m, n)}$ as shown in equation (2) is the weight of link(m, n) calculated based on the number of incoming links of page n and the number of incoming links of all reference

pages of page m.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \in R(m)} I_p} \quad (2)$$

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p \in R(m)} O_p} \quad (3)$$

Where I_n and I_p are the number of incoming links of page n and page p respectively. R(m) denotes the reference page list of page m. $W^{out}_{(m, n)}$ is as shown in equation (3) is the weight of link(m, n) calculated based on the number of outgoing links of page n and the number of outgoing links of all reference pages of m. Where O_n and O_p are the number of outgoing links of page n and p respectively. Another formula by Wenpu *et al* for the WPR is as shown in equation (4) which is a modification of the PageRank formula (equation 1).

$$WPR(n) = (1-d) + d \sum_{m \in B(n)} WPR(m) W^{in}_{(m,n)} W^{out}_{(m,n)} \quad (4)$$

3.3 Hyperlink Induced Topic Selection (HITS) Algorithm

HITS algorithm is proposed by Klienberg. He gives two forms of web pages called as hubs and authorities. Hubs are the pages that behave as resource lists. Authorities are pages having valuable contents. A good hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is denoted by many good hub pages on same content. A page may be a good hub and a good authority at the same time.

HITS is technically, a link based algorithm. In HITS algorithm, ranking of the web page is decided by analyzing their textual contents against a given query. After collection of the web pages, the HITS algorithm concentrates on the structure of the web only, neglecting their textual contents. Original HITS algorithm has some problems which are given below.

1. High rank value is given to some popular website that is not highly relevant to the given query.
2. Drift of the topic occurs when the hub has multiple topics as equivalent weights are given to all of the out links of a hub page.

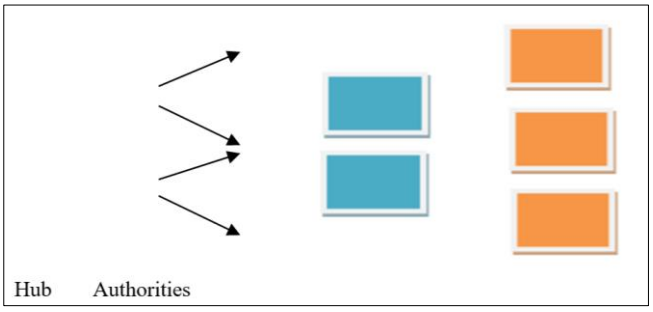

Fig 3: Hubs and Authorities

The HITS algorithm treats WWW as a directed graph G(V,E), where V is a set of Vertices representing pages and E is a set of edges that correspond to links. There are two major steps in the HITS algorithm. The first step is the sampling step and the second step is the Iterative step. In the Sampling step, a set of relevant pages for the given query are collected i.e. a sub-graph S of G is retrieved which is high in authority pages.

$$H_p = \sum_{q \in I(p)} A_q \quad (5)$$

$$A_p = \sum_{q \in B(p)} H_q \quad (6)$$

4. Comparison of Various Page Ranking Algorithms

Table1: Comparison of Algorithms

Algorithm / Criteria	PageRank	Weighted PageRank	HITS
Mining Technique	**WSM**	**WSM**	**WSM & WCM**
Working	Computes scores at index time. Results are sorted on the importance of pages.	Computes scores at index time. Results are sorted on the Page importance.	Computes scores of N highly relevant pages on the fly.
I/P Parameters	Backlinks	Backlinks, Forward links	Backlinks, Forward links and content
Complexity	O(log N)	<O(log N)	<O(log N)
Limitations	Query independent	Query independent	Topic drift & efficiency problem
Search Engine	Google	Research model	Clever

5. Conclusions

The web structure mining has very importance to deals with many algorithms that lead to fetch the data from any website in effective manner with best output. Now explored the hyperlink structure and understand the web graph in a simple way. This paper also focuses on the important algorithms used for hyperlink analysis, explore those algorithms and compare them. We survey the many research areas of web mining and focusing on many web structure mining algorithms like PageRank algorithm, Weighted PageRank algorithm, Weighted Content PageRank algorithm (WPCR), HITS etc. We analyzed their working. We noticed one thing that all the algorithms except HITS are based on Page Ranking. They are modifications in PageRanking algorithm. Based on the algorithm used, the ranking algorithm provides a definite rank to resultant web pages. Since this is a vast area, and there a lot of work to do.

6. References

1. Duhan N, Sharma AK, Bhatia KK. Page Ranking Algorithms: A Survey, Proceedings of the IEEE International Conference on Advance Computing; c2009.
2. Herrouz A, Khentout C, Djoudi M. Overview of Visualization Tools for Web Browser History Data.

IJCSI International Journal of Computer Science. 2012 Nov;9(6-3):92- 98
3. Kosala R, Blockeel H. Web Mining Research: A Survey, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining. 2000;2(1):1-15.
4. Cooley R, Mobasher B, Srivastava J. Web Mining: Information and Pattern Discovery on the World Wide Web". Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence; c1997.
5. Da Gomes Jr MG. Gong Z. Web Structure Mining: An Introduction, Proceedings of the IEEE International Conference on Information Acquisition; c2005.
6. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, *et al*. Graph Structure in the Web, Computer Networks: The International Journal of Computer and Telecommunications Networking. 2000;33:1-6.
7. Kleinberg J, Kumar R, Raghavan P, Rajagopalan P, Tompkins A. Web as a Graph: Measurements, models and methods, Proceedings of the International Conference on Combinatorics and Computing; c1999. p. 18.
8. Page L, Brin S, Motwani R, Winograd T. The Pagerank Citation Ranking: Bringing order to the Web. Technical Report, Stanford Digital Libraries SIDL-WP 1999-0120; c1999.
9. Xing W, Ali Ghorbani. Weighted PageRank Algorithm, Proc. Of the Second Annual Conference on Communication Networks and Services Research, IEEE.
10. Ridings and M. Shishigin, PageRank Convered. Technical Report; c2002.
11. Zareh Bidoki AM, Yazdani N. Distance Rank: An intelligent ranking algorithm for web pages Information Processing and Management. 2008;44(2):877-892.
12. Jon Kleinberg. Authoritative Sources in a Hyperlinked Environment, In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms; c1998.
13. Chakrabarti S, Dom BE, Kumar SR, Raghavan P, Rajagopalan S, Tomkins A, *et al*. Mining the Web's Link Structure, Computer. 1999;32(8):60-67.