International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 1 | Jan -2017 www.irjet.net p-ISSN: 2395-0072

A Study on Web Structure Mining

Anurag Kumar[1], Ravi Kumar Singh[2]

[1]Dr. APJ Abdul Kalam UIT, Jhabua, MP, India
[2]Prestige institute of Engineering Management and Research, Indore, MP, India

---

Abstract - ***As web is the largest collection of information and plenty of pages or documents, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. Web mining divides into web content, web structure and Web usage mining. In this paper, we focus on one of these categories: the Web structure mining. Web structure mining plays very significant role in web mining process. Within this category, we explain mining algorithms and review some popular methods applied in Web structure mining.***

***Key Words*: web mining, web structure mining, pagerank algorithms, Distance rank, Tag Rank , Time Rank .**

1. INTRODUCTION

Internet is a network of worldwide level, constantly changing and non-structured [1]. The Web is the largest data source in the world. Web mining aims to extract and mine useful knowledge from the Web. It is a multidisciplinary field involving data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia, etc. The amount of information on the Web is huge, and easily accessible. The knowledge does not come only from the contents of the web pages but also from the unique feature of Web, its hyperlink structure and the diversity of contents. Analysis of these characteristics often reveals interesting patterns and new knowledge which can be helpful in increasing the efficiency of the users, so the techniques which are helpful in extracting data present on the web is an interesting area of research. These techniques help to extract knowledge from Web data, in which at least one of structure or usage (Web log) data is used in the mining process. In this paper, we firstly provide a survey on overall Web mining concepts and technologies, then, focus on Web structure mining in detail.

2. WEB MINING

In general, Web mining tasks can be classified into three categories: *Web content mining, Web structure mining and Web usage mining.* All of the three categories focus on the process of knowledge discovery of implicit, previously unknown and useful information from the Web. Each of them focuses on different mining objects of the Web. Fig. 1 shows the Web categories and their objects. As follows, we provide a brief introduction about each of the categories.

2.1 Web Content mining

Web Content mining refers to the discovery of useful information from the contents of the webpage using text mining techniques. Webpage can be in traditional text form or in the form of multimedia document containing table, form, image, video and audio. Web content mining identifies the useful information from the Web contents. Web content mining could be differentiated from two points of view the agent-based approach or the database approach. The first approach aims on improving the information finding and filtering and could be placed into the following three categories [1]:

***Intelligent Search Agents*:** These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.

Information Filtering/ Categorization: These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.

***Personalized Web Agents*:** These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.

2.2 Web structure mining

The goal of web structure mining is to generate structural summary about web pages and web sites. It shows the relationship between the user and the web. It discovers the link structure of hyperlinks at the inter document level [6, 8]. Two algorithms that have been proposed to lead with those potential correlations: HITS [3] and PageRank [2], and Web structure mining itself will be discussed in the next section.

2.3 Web usage mining

It is also known as Web log mining, is used to analyze the behavior of website users. It tries to discover useful information secondary data derived from the interaction of

users while surfing web [8]. Web usage mining collects the data from Web log records to determine user access patterns of Web pages. This information is often gathered automatically into access logs via the Web server. Mainly there are four types of data sources present in which usage data is recorded at different levels they are: client level collection, browser level collection, server level collection and proxy level collection. It contains four processing stages including

- Data collection
- Preprocessing
- Pattern discovery and Analysis

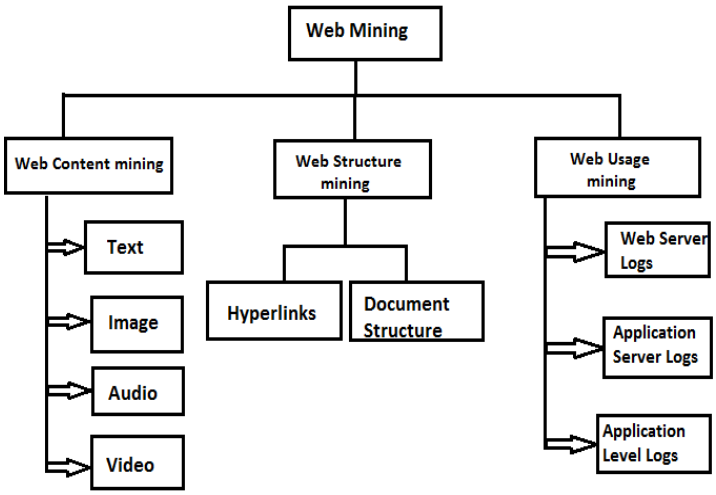


Fig -1: Web mining categories and objects

3. WEB STRUCTURE MINING

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [4], which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet.

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 (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor texts [4]. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the traditional data mining tasks. As follows, we summarize some of these possible tasks of link mining which are applicable in Web Structure Mining.

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

2. Link-based Cluster Analysis 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.

3. 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.

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

5. Link Cardinality. The main task here is to predict the number of links between objects, Page categorization, finding related pages, finding duplicated web sites and to find out similarity between them.

3.1 Web structure Mining Algorithms

Some of the popular web mining algorithms
The frequently used algorithms in web structure mining, to access the webpage or website effectively for user in online.

(i) Page rank algorithm.
(ii) HITS algorithms(Hyperlink-Induced Topic Search)
(iii)Weighted page rank algorithm.
(iv)Distance rank algorithm.
(v) Weighted page content rank algorithm.
(vi)Webpage ranking using Link attributes.
(vii) Eigen Rumor Algorithm
(viii) Time Rank Algorithm
(ix)Tag Rank Algorithm
(x) Query Dependent Ranking Algorithm

(i) Page rank algorithm

L. Page and S. Brin [2] has developed Page Rank algorithm during their Ph.D. research work at Stanford University based on the extract analysis. They applied the extract analysis in Web search by treating the incoming links as credentials to the Web pages. Page rank algorithm is the

most frequently used algorithm for ranking the various pages. Functioning of the page rank [6, 9, 10].

The Page Rank algorithm is defined as [2]: "*We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85.The Page Rank of a page A is given as follows:*

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

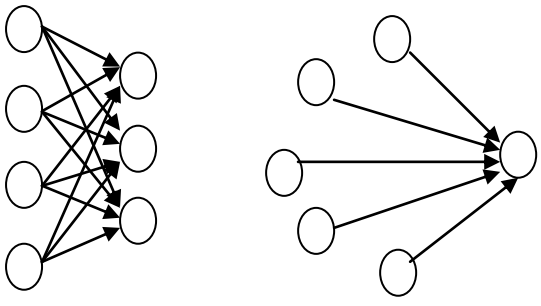
(ii) HITS: (Hyper-link Induced Topic Search) Algorithm

In HITS concept, Kleinberg [9] identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs.

According to Kleinberg [9], "*Hubs and authorities exhibit what could be called a mutually reinforcing relationship: 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*".

The HITS algorithm uses the Sampling and Iterative steps. In the Sampling step, a set of related pages for the given query are collected i.e. a sub-graph S of G is retrieved which is high in influence pages. This algorithm starts with a root set B, a set of S is obtained, keeping in mind that S is comparatively small, rich in relevant pages about the query and contains most of the good authorities. The second step, Iterative step, finds hubs and authorities using the output of the sampling. [6, 9, 14, 16].

The key objective of the algorithm is that by viewing the one mode web graph actually comprising two modes called hubs and authorities. A hub is a node primarily with edges to authorities and so a good hub has links to many authorities. A good authority is a page that is linked to by many hubs. Starting with a specific search objective, HITS algorithm performs a text based search to seed an initial set of results. An iterative relaxation algorithm will assigns hubs and authority weights using matrix power iteration. The CLEVER search engine is built primarily using the basics of HITS algorithm [12].



Hubs *Authorities*

Figure-2. Layout structure for webpage using hub and authority.

Hyperlink-induced topic search (HITS) is an iterative algorithm for mining the Web graphs to categorize the topic hubs and it symbolize the pages with good sources of content/data and authorities symbolize pages with good sources of links / hyperlink. Authorities are highly ranked pages for a given topic; hubs are pages with links to authorities. This algorithm takes as input search results returned by usual text indexing techniques and it reduces the results to identify hubs and authorities. The number value and weight of hubs pointing to a web page decide the page's authority. And also the algorithm assigns weight to a hub based on the validity of the pages it points to. Figure-2 shows the webpage layout structure which exists in HITS algorithm.

The HITS algorithm performs a chain of iterations, each consisting of two fundamental steps.

Authority update - This is to update each nodes authority score to be equal to the sum of the hub scores of each node that points to it. A node with high authority score is being linked to by pages that are recognized as Hubs for information.

Hub Update - This is to update each nodes Hub store to be equal to the sum of the authority scores of each node that it points to it. A node with high hub score is linking to nodes that are considered to be authorities on the subject.

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 [17]:

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. These situations could provide wrong definitions about a good hub or a good authority.

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.

(iii) Weighted page rank algorithm

This algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an extension of Page Rank algorithm. [20]. 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 back link 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. The algorithm is more efficient than page rank algorithm due to using of two parameters called back link and forward link. The status from the number of in links and out links is recorded and easy assign the label as Win and Wout. The meaning is assigned in terms of weight values to incoming and out coming links. This is denoted as Win (m, n)

and Wout (m, n) respectively. Win (m, n) is the weight of links (m, n) as given in the equation. Finally the calculation is based on the number of incoming links to page n and the number of incoming links to all references pages of page m [10].

$$W_{in}(m,n) = In \Sigma Ip$$
$$P E R (m)$$

In - is denoted as the number of incoming links of page n, Ip - is denoted as the number of incoming links of page p, R(m) is the reference page list page m. Wout(m,n) is the weight of links (m,n) as given the equation. The final value is calculated on the basis of outgoing links of page n and the number of outgoing links of all the reference pages of page m.

$$W_{out}(m,n) = In \Sigma Op$$
$$P E R (m)$$

On - is number of outgoing links of page n, Op is number of outgoing links of page p. Then the weighted Page rank is calculated as follow as

$$WPR(n) = (I\text{-}d) + d \Sigma WPR(m) W_{in}(m,n) W_{out}(m,n)$$

(iv) Distance rank algorithm

The distance between the two web pages treated as penalty, called ''Distance Rank'' to compute the ranks of web pages. It is denoted by the number of average clicks between the two web pages. The idea is to minimize the penalty or distance, and finally a web page with less distance value to be considered as higher rank value to be used. Like Page Rank algorithm, the rank of each page is denoted as the weighted sum of ranks and all the pages having links to the page. A page has a high rank and it has more back links or pages having links to one page have higher ranks. Specifically a web page is having as many as input links should have low distance and if pages pointing to this page have low distance then this page should have a low distance.

Two definitions have been applied to reach the better outcome of the webpage.

Definition-1. If page x points to page y then the weight of link between x and y is equal to Log_{10} O(x) where O(x) shows x's out degree (number of forward links).

Definition-2. The distance between two pages x and y is the weight of the shortest path i.e. the path is having the minimum value from x to y. We call this logarithmic distance and denote it with d_{xy}.

(v) Weighted page content rank algorithm

Weighted Page Content Rank Algorithm (WPCR) [18] 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. Importance here means the popularity of the page i.e. how many pages are pointing to or are referred by this particular page. It can be calculated based on the number of in links and out links of the page. If a page is maximally matched to the query, that becomes more relevant. This algorithm is better than the pagerank as well as weighted pagerank algorithm because its complexity is less than both the algorithm and is < (Ologn).

(vi) Webpage ranking using link attributes

In page algorithm concentrate on three factors like qualified position in the web page, tag where the link is contained and the length if the linked text. The same condition to be applied to the page rank concept with link attributes will leads to better result in performance. The page rank with link algorithm uses the page rank formula as base, it follows as **PR (p) = Q/T + (1-q) Σ PR (ri)/ L (ri)**, is used to calculate the probability of the page.

But the page rank algorithm with link attributes uses the above modified expression as follows **R(i) = q/T + (1-q) Σ W(j,i)R(j) / Σ W(j,k)**. From the expression R (i) corresponds to the probability to reach the page while searching the website.

(vii) Eigen rumor algorithm

The Eigen rumor algorithm gives a rank score to every blog in the webpage and it weights the scores of the hub and authority of the blogger based on the calculation of Eigen value. The Eigen algorithm concentrates on hub and authority value in each web page. This algorithm technique is applicable in web content mining and it uses the methodology like adjacency matrix and it creates an agent to object for linkage, not like web page to web page. The algorithm applied in the search engine in blog community model. The key objective of the algorithm is applicable in blog ranking concept. The performance in web site search is high and the limitation of the algorithm is not so better like other ranking algorithms.

(viii) Time rank algorithm

The time rank algorithm basically works on time based visiting model. This algorithm technique is applied in the field of web usage mining. The key function of the algorithm is that the visiting time is additional to the calculated score of the actual page rank value of that web page.

It uses the actual value of the page as given as input for processing and leads the average result as output. Relevancy of using this algorithm is high value when it compared to other ranking algorithms. It is more dynamic in nature with the attributes of duration of used by user and the basic structure of the links used in the webpage. The drawback of the algorithm is that omits the most relevant pages at the time search in web site. i.e. Avoid the recent accessed Webpage by the user and need more steps to reach out the same webpage.

(ix) Tag rank algorithm

The rag rank algorithm consist of two factors are initial probabilistic tag relevance estimation and random walk refinement. Collaborative tagging systems allow users to assign keywords so called tags.

Tags are used for routing, finding resources and unexpected browsing and thus provide an immediate benefit for users. These systems usually include tag recommendation mechanisms reduction the process of finding good tags for a resource, but also consolidating the tag vocabulary across users.

This algorithm consists of an adaptation of user based collaborative filtering and a graph-based recommender built on top of Folk Rank.

(x) Query dependent ranking algorithm

The query-dependent ranking analyzes the relationship between the query results and the tuples in the database. The role of query-dependent ranking by analyzing the user's browsing choices and comparing different queries in terms of their similarity with each other without requiring knowledge of the Web database. The query dependent uses the method of learning to rank based on a distributional similarity measure for gauging the similar data between the queries. Each webpage is accessed through query statement and its nature.

In general the queries describe the user's information need and play an eminent role in the context of ranking for information retrieval and webpage search. The users search intention is based on navigation, informational and transactional queries. In such required, the query categorization has high correlation with users different expectation on the result achieved through query dependent ranking.

Table-1. Comparison state of page rank algorithms.

Algorithms	Page rank	HITS	Weighted page rank
Main Technique	Web Structure mining	Web Structure mining Web content mining	Web Structure mining
Methodology	It score for pages at the time of Indexing	It uses hubs and authority of the relevant pages.	It performs on the basis of Input and output links.
Input parameter	Back links	Content, back and forward links	Back link and forward link
Quality	Medium	Low	Higher than PR
Mining technique used	Web Structure Mining	Web Structure Mining and web content mining.	Web Structure Mining
Search engine	Google	IBM search engine Clever.	Research model
Limitations	Query independent	Efficiency problem	Query independent.

4. CONCLUSIONS

In this paper we survey the research area of Web mining, focusing on the category of Web structure mining. We also discussed the different algorithms used in web structure mining. Web mining deals with retrieving the data from web with best output. The web structure mining also deals with many algorithms that lead to fetch the data from any website. In general web structure mining is that retrieves the data from website for online user in effective manner.

Since this is a vast area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research.

REFERENCES

[1] Herrouz, A., Khentout, C., Djoudi, M. Overview of Visualization Tools for Web Browser History Data, IJCSI International Journal of Computer Science Issues, Vol.9, Issue 6, No3, November 2012, pp. 92-98, (2012).

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. Technical report, Stanford University, 1998.

[3] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998, pages 668-677 – 1998.

[4] L. Getoor, Link Mining: A New Data Mining Challenge. SIGKDD Explorations, vol. 4, issue 2, 2003.dfsdfs

[5] Inamdar, S. A. and shinde, G. N. 2010. An Agent Based Intelligent Search Engine System for Web Mining. International Journal on Computer Science and Engineering, Vol. 02, No. 03.

[6] Preeti Chopra and Md. Ataullah. 2013. A Survey on improving thee Efficiency of different Web Structure Mining Algorithms. International Journal of

[7] Engineering and Advanced Technology. 2(3): 1-3. Johnson, F., Gupta, S.K., Web Content Minings Techniques: A Survey, International Journal of Computer Application. Volume 47 – No.11, p44, June (2012).

[8] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998, pages 668-677 – 1998.

[9] Arun Kumar Singh, Avinav Pathak and Dheeraj Sharma. 2013. A Survey on Enhancing the Efficiency of various web structure mining algorithms. International Journal of Computer Applications Technology and Research. 2(6): 771 -774.

[10] S. Sathya Bama, M.S. Irfan Ahmed and A. Saravanan. 2013. Improved Page Rank Algorithm for Web structure mining. International Journal of Computer and Technology. 10(9): 1969-1976.

[11] Ramesh Prajapati. 2012. A Survey paper on Hyperlink - Inducted Topic Search (HITS) Algorithms for Web Mining. International Journal of Engineering Research and technology. 1(2): 1-8.Dsff

[12] T. Munibalaji and C. Balamurugan. 2012. Analysis of Link Algorithms for Web Mining. International Journal of

Engineering and Innovative Technology (IJEIT). 1(1): 81-86.

[13] Kyung - Joong Kim and Sung-Bae Cho. 2007. Personalized mining of web documents using link structures and fuzzy concept networks. Applied Soft Computing. 7(1): 398-410.

[14] Miguel Gomes da Costa Junior and Zhiguo Gong. 2005. Web Structure Mining: An Introduction. International Conference on Information Acquistion, June 27- July 3 2005, China.

[15] T. Munibalaji and C. Balamurugan. 2012. Analysis of Link Algorithms for Web Mining. International Journal of Engineering and Innovative Technology (IJEIT). 1(1): 81-86.

[16] Ramesh Prajapati. 2012. A Survey paper on Hyperlink Inducted Topic Search (HITS) Algorithms for Web Mining. International Journal of Engineering Research and technology. 1(2): 1-8.

[17] A. Barfourosh, H.R. Motahary Nezhad, M. L. Anderson, D. Perlis, Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition, 2002.

[18] T.A. Runkler and J.C. Bezdek. 2003. Web Mining with relational clustering. International Journal of Approximate reasoning. 32(2-3): 217-236.

[19] Federico Michele Facca and Pier Luca Lanzi. 2005. Mining interesting knowledge from weblogs: a Survey. Data and Knowledge Engineering. 53(3): 225- 241.

[20] Haveliwala. T.H. 2003. Topic sensitive Page Rank: A Context sensitive ranking algorithm for Web Search. Knowledge and data engineering, IEEE Transactions. 15(4): 784-796.

[21] Chen Lihui and Chue Wai Lian. 2005. Using Web structure and summarization techniques for Web content mining. Information Processing and Management. 41(5): 1225-1242.

[22] B. L. Shivakumar and T. Mylsami 2014, Survey On Web Structure Mining, ARPN Journal of Engineering and Applied Sciences, 9(10):1914-1923

BIOGRAPHIES

Anurag Kumar received the M.Tech. degree in Information Architecture & Software engineering from School of Computer Science & IT,DAVV, Indore, India in 2016. Currently, he is an Asst. Prof. at Dr. APJ Abdul Kalam UIT, Jhabua, MP, India. His areas of interest are UX Design, Data Mining and Big data.

Ravi Kumar Singh received the M.Tech. degree in Information Architecture & Software engineering from School of Computer Science & IT,DAVV, Indore, India in 2016. Currently, he is an Asst. Prof. at Prestige Institute of Engg. Management & Research, Indore, MP, India. His areas of interest are Data Mining and Algorithm Design and analysis.