Parichay Maharaja Surajmal Institute Journal of Applied Research

A Study of Web Structure Mining Algorithmic and its Application

Harjender Singh*

Abstract: ***Web Data Mining is an important area of data mining which focuses on the extraction of interesting information from WWW and can be classified as a web content mining, web structure mining, web usage mining and other types of information. In each of the three different forms of web mining, web structures mining is the aim of the paper is to provide an up to date and past evaluation and update and outlines future research guidelines. Web structure mining is a tool used to identify the link between web pages connected by information or direct links. It provides information on how different pages are linked to this vast website. Web Structure Mining finds hidden basic structures for web application search and uses hyperlinks for other application. This paper presents an overview of research development and several important research topics in terms of comparisons and summaries of web structure data mining with its applications.***

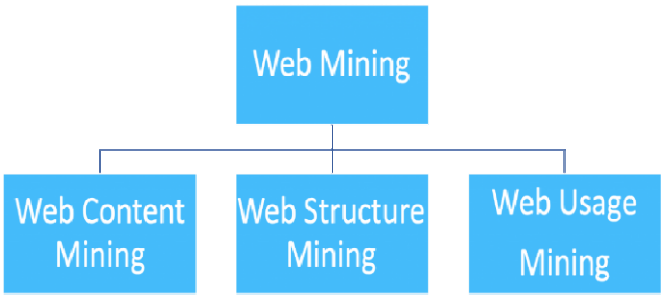
Keyword: ***Web mining, Web structure mining, Web content mining***

1. INTRODUCTION

Motivation

The motivation for writing this paper is primarily an interest in undertaking a challenging task in an interesting area of research (Web Structure Mining). The opportunity to learn about a new area of Structured mining and its techniques.

Problem definition



The application to discover structural information on the Web is web structural mining. The web graph structure consists of web pages as nodes and hyperlinks as borderswhich connect related pages. Structure mining essentially shows the structured summary of a specific website. It identifies relationships between linked web pages or a direct link. Web structure mining can be very useful in determining the connection between two commercial websites.

Web applications are characterised by hypertext links and certain procedures, which permit real-time discussions between client and server. The hyperlinks are symbolised with a different mark from the rest of the document of word, picture or icon document which causes the browser to be placed on the web, regardless of where the document is located. Electronic documents which refer to each other have been gathered and led to the Web name.

Objective

The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps.

Literature survey

This type of mining approach is used to identify the link between different web pages linked through information or direct links. This type of mining process is primarily concerned with deriving an existing type of unknown associations between the different web pages. The mining usage type can examine log data stored in the web server, proxy server and client caches in indifferent formats. Thus, because of the vast amount of information, the mining approach oriented towards the structure may reduce the two major issues posed by the WWW. And the first problem is not at all related to the search results.

**Assistant Professor, Department of Computer Applications*
Maharaja Surajmal Institute, Affiliated to Guru Gobind Singh Indraprastha University, New Delhi
harjendersingh@msijanakpuri.com

The relevance of the information searched has been misunderstood by the problems which the search engines apply mainly to the low accuracy conditions. And secondly, the inefficiency to create the large quantity index when the data is provided via the internet. This can reduce the reminder volume along with the mining approach in terms of its content. This sort of minimisation has therefore been used to find models of this type of mining technique in the web-based hyperlink architecture.

There are some issues for Web Structure mining process which is related with the architecture of the hyper links presented over the web. Earlier these hyper-links get analysed by the researchers. But these emerging interest within the web mining mechanism and various researches on analysis of structure have also increased which leads to latest growing research domain which is known as link-mining. Link mining is at communication level of various works like – analysis of link, web mining process, relational-learning process, graph mining approach etc..

WEB Structure Mining

The web structure mining can help users to get their relevant documents just by analysing link oriented structure of any type of web content. Also they give some issues in order to work with structure of any given hyperlinks in the web. Analysis of these links is also a domain for researchers and the web consist of not only some pages but also of hyperlinks directing any page to any other type of pages. It finds the specific structure of links that is hyperlink for any other inter document, also generate structural result regarding any web page or websites.

This concept of web structure mining is used for retrieving pages that are not only relevant also of high quality or authoritative on any topic. Though by the increase in attention on web mining concept, the analysis of structure have also been increased. The hyper text is placed and performed the web mining through some learning and inductive type of approach and through graph mining process.

2. TYPES :

Hyperlinks:

Hyperlinks is unit of structure which can link a particular location within a web page to other location of same or different web pages. There are majorly two types of Hyperlinks which are, the hyperlink that connects different sections of similar page is called intra-document hyperlink and the hyperlink that connects the two type of pages is referred as an inter-document hyperlink. There exist a significant body that work on hyperlinks analysis and provide up-to date survey.

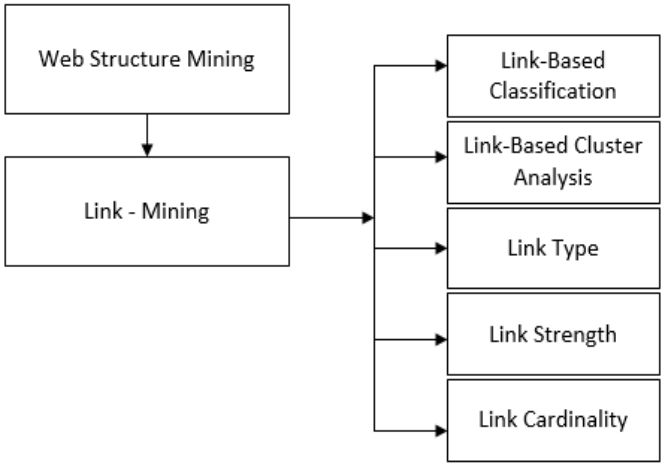
Document Structure

The content present in web pages may also be arrange in the tree type structured pattern which is due to several types of HTML tags or designing XML tag in any page of web. Here mining is focused to automatically extraction of document object model structures out of document.

The WSM is a type of web mining process for data which is type of tool applied to recognize links between web pages that are connected through information or direct connection. This type of data structure can be applied through supplying web architecture model by techniques of database for web pages. This type of connection enables any search engine in order to take data which is similar to query for search which directly connected with linking page of web site on which the content is placed. Generally this type of work may be performed by use of spiders scanning process over website and retrieve content from home pages , after that with the help of links of reference it brings some particular pages in front that have some desired details.

The main objective of the web structure mining process is to discover the unknown association in between webpages and web. This structure type of DM can enable to use in the business purpose to connect the details of websites to allow the navigation function and also make cluster information over site maps. This may enable of their websites to navigate function and group the information over the site maps. Also this may help users to have capability to use required information with the help of association of keywords along the content based mining process. Here web structure mining can also be define as in terms of graphs. The webpages are representing the nodes and hyperlinks represents the edges. Its shows the relationship between the user and web. The main motive of web structure mining is creating structured summaries about information on web pages.

Processes in Web-Structure-Mining:



Link-oriented process of Classification

This process is latest DM work in linking the domains. This main target of the process is to anticipate the categories of the web pages and also depends on the word which may found over the pages or links in between the pages over HTML tags, anchor text and may be present on the attributes over web pages.

Link-oriented Cluster-Analysis process:

The objective of this process is finding the sub classes which are naturally found. The data in this process is fragmented into clusters in which same type of object are grouped and different type of objects are also grouped separately within other groups. It is different process in which this analysis is considered as unsupervised and also helpful in finding out the hidden pattern in the data.

Type of Link:

There is a wider range of works in process of prediction of pattern of links between two different entities or can anticipate the objective of the links.

Strength of Links:

In this each link should have unique weight so they may be related along their weights.

Cardinality of Links:

Main objective of the process is the anticipation of the links present in between the objects.

3. OPTIMIZATION TECHNIQUES IN WSM

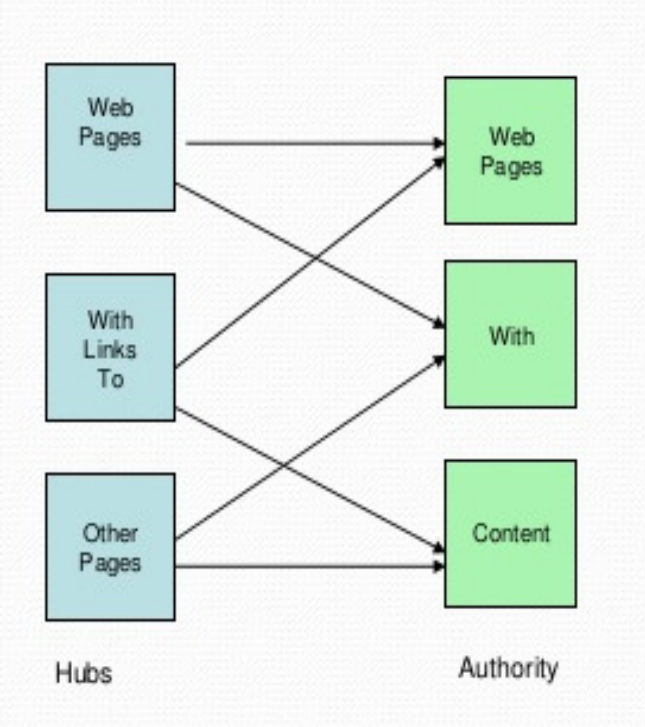
Clustering-technique:

This Clustering technique is the process of grouping of sets of objects in such manner that same type of object is in same group are termed as clusters. The primary task of explanatory DM process that basic type of approach is used for analysis of statistical data which is used for various domains like pattern , ML approach , Data retrieval , Bioinformatics and picture analysis.

Page-Ranking Technique:

This ranking technique is relationship between group of items like two given points either the first is ranked lesser to other or first is ranked greater than other or both can be equal ranked . It is inadequate to use total number of object as two separate object as they may have similar ranking, in this case page ranking is completely merged. This technique leads to better approach that can calculate the importance of a web page by calculating the number of pages linked to it, the greater the number of pages the greater will be the importance. The calculated links are called back-links. In any case back link generate from key page then this link gives higher weightage than those which are coming from non-important pages so link from first page to other is measured by VOTE.

Hyper-link induced-topic-search algorithm:



This approach is defined with two attributes like hubs and authorities. It uses process of link analysis that may rate any webpage and also developed an approach which enable use of linking structure for web so that it finds and rank the page related with specific concept. The hits algorithm uses sampling and iterative steps. This technique follows the concept of search engine called ASK.

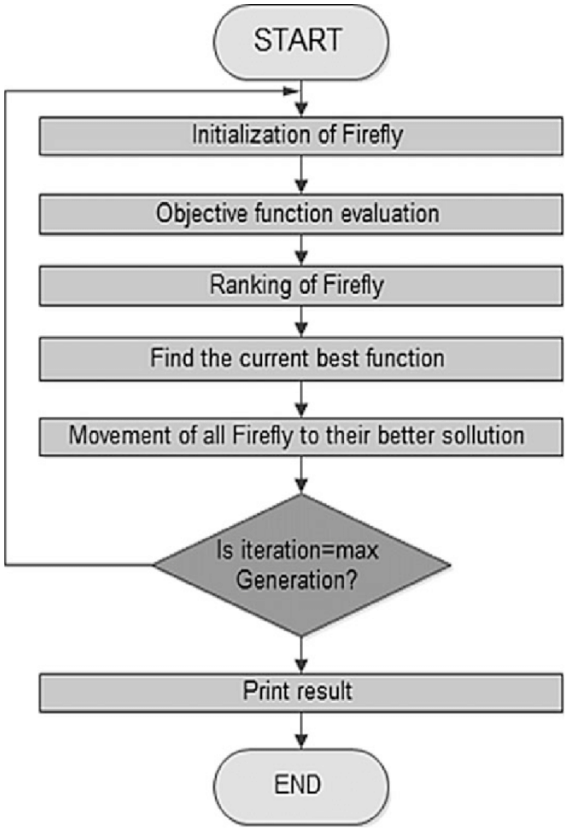
Ant-colony-optimization algorithm:

In this pheromone trails are used. According to the probabilities the artificial ants may carry one or more objects and drop them. These agents do not directly contact each other, but through configuration of objects on the floor they may influence themselves. Thus from this the artificial ants can able to create same type of objects and a problem known as data clustering. A traveling salesperson problem is perfect for ACO because this problem resembles to finding the closest path to the food source. Unless pheromone evaporation is implemented (a solution disappears after a period of time), ACO results in premature convergence to local optimal solution.

Genetic-Algorithm:

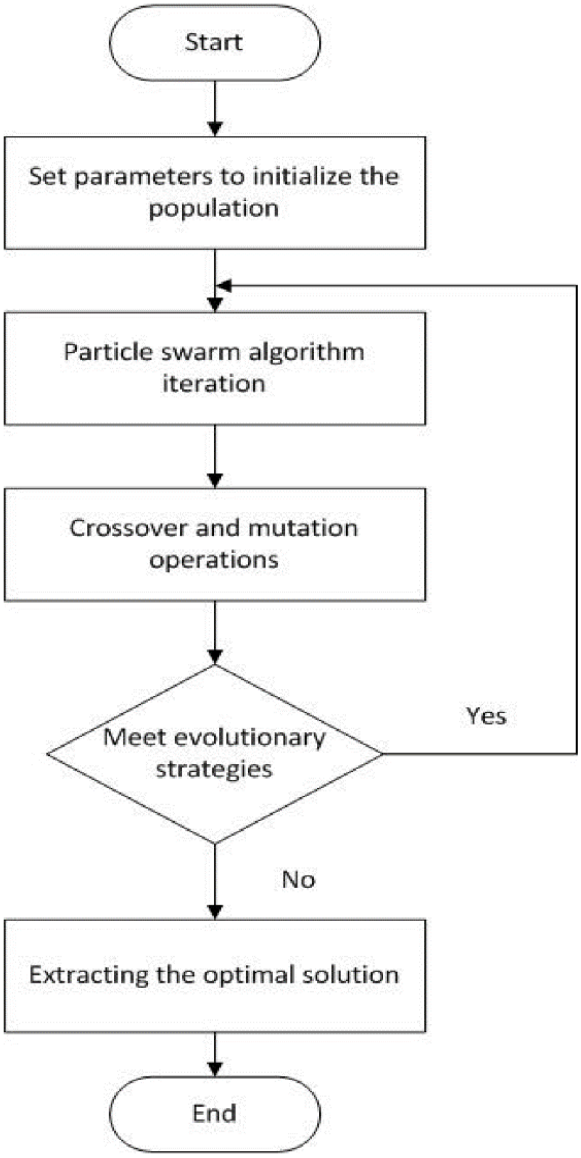
Due to the extra ordinary enhancement in the WWW have increased the subject of crawling of a web therefore this approach is used. The targeted process of crawling includes mechanical categorization of web page approach like WPC which is required to find whether the page is being taken for approach or not. For this type of approach of leaning process, the genetic algorithm is dependent over mechanical type of WPC approach that uses HTML and conditions both that are related with each tag as feature of classification and also it analyses best classifier obtained from webpages through just estimating within learned type of classifier and some latest web pages. By applying the hybrid type of GA or PSO approaches this may have various issues help to recover those complexities.

Firefly Algorithm:



It is Meta heuristic approach that is inspired through glowing nature of fire flies. The main aim of firefly's flash is to attract other fireflies. In this analysis of the algorithm it formulates through supposing situations like entire fireflies are unisex such that any one firefly may attract the entire type of fireflies. The brightness of fireflies flash make them attractive accordingly. Any two type of fireflies if one is less glowing then one will attract to brighter one. Here increase in distance makes decrease in brightness. If no firefly is glowing as compare to given then it may be randomly move.

PSO algorithm:



In PSO (particle swarm optimization) algorithm the birds in a flock are represented as particles in n dimension. For representing one solution of problem, best fitness value of particle at a location in the n-dimensional problem space. When a particle updates its position it generates another problem solution is generated and then new solution is evaluated by fitness function and process is repeated until the stopping criteria is met. PSO has random population matrix like GA, but rows in matrix are called particles instead of chromosomes. Particles are potential solution that move in particular direction on cost surface with certain velocity. They

update theirvelocities and position using formulas based on knowledge about best solution achieved by complete swarm of particles and each particles at its best.

4. CONCLUSION

Web mining is powerful technique used to extract the information from past behaviour of users. It plays an important role in this approach. Various algorithm are used in ranking process as described in this research paper . when distributing the rank score Page rank and HITS treat all links equally. In this we majorly focused on web structure mining, its types, approaches, techniques and algorithm and some related issues with them to have an idea about their application and effectiveness. Since this is a huge 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] G. D. Kumar and M. Gosul, "Web MiningR easearch and Future Directions", Advances in Network Security and Applications, 4th International Conference on Network Security and Application, pp. 489–496, Springer Berlin Heidelberg, (2011).

[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, (2000).

[3] Miguel Gomes da Costa J´unior and Zhiguo Gong, "Web Structure Mining: An Introduction", Proceeding of IEEE

[4] International Conference on Information Acquisition, Hong Kong and Macau, China, (2005).

[5] Core.ac.uk

[6] En.wikipedia.org

[7] www.researchgate.net

[8] www.ijser.in