Volume 4, Issue 8, August 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Web Usage Mining Based on Ant Colony Optimiztion

Er. Reena
Research Scholar,
Desh Bhagat University,
Gobindgarh Mandi,
Punjab, India

Er. Jyoti Arora
Associate Professor,
Desh Bhagat University,
Gobindgarh Mandi,
Punjab, India

Abstract: As the web is growing very rapidly, Search Engines play a chief role in retrieving data from web. When we search for a topic in a web, it presents hundreds of search results. It is highly impossibly to visit all the web pages to find relevant information. Extracting the required data for the user is a big challenge for the administrators or developers nowadays. And here is the role came for the web mining to play. Web mining is a way to extract the meaningful data for the users as per their need. Thus the following paper basically focuses on the extracting or retrieving the data for users as per their requirements based on the ants behavior. In this research we are basically dealing on the real ants behaviors.

Keywords: OLAP, ACO, JPEG, GIF, URL

I. INTRODUCTION

Data mining is commonly defined as the process of discovering useful patterns or knowledge from data sources (e.g., databases, texts, images, the Web, etc.). The patterns must be valid, potentially useful, and understandable. Generally, data mining uses structured data stored in relational tables, spread sheets, or flat files in the tabular form. With the growth of the Web and text documents, Web mining and text mining are becoming increasingly important and popular. Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data.

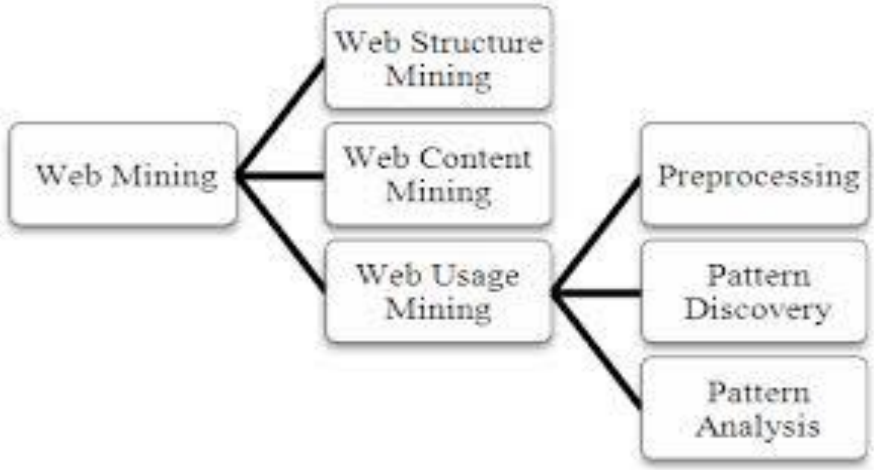

Fig1. Classification of Web Mining

It is divided into the three parts as defined: web content mining, web structure mining and web usage mining. Basically in web content mining is defined as the resource discovery from content of millions of sources across the World Wide Web. It deals the contents like text, Image, audio, video, metadata and hyperlinks. Content mining is used using the various technologies as classifications, clustering and associations. The second category of web mining is web structure mining, in this Generate structural summary about the Web site and Web page. The third category is web usage mining, in this Discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities.

II. WEB USAGE MINING

In this, the data is derived from the user activity i.e it derives the user activity, the data that is concerned by the user as a meaningful data is derived from the user log. There are various logs depending upon which the process is carried out. In web usage mining, the sources of retrieval of data is Server access logs, Server Referrer logs, Agent logs, Client-side cookies, User profiles, Search engine logs, Database logs. Using the various entries stored in the defined logs the web usage mining in processed. Depending upon the different logs and different methods, the various processes are designed in the manner to extract the meaningful data to the user.

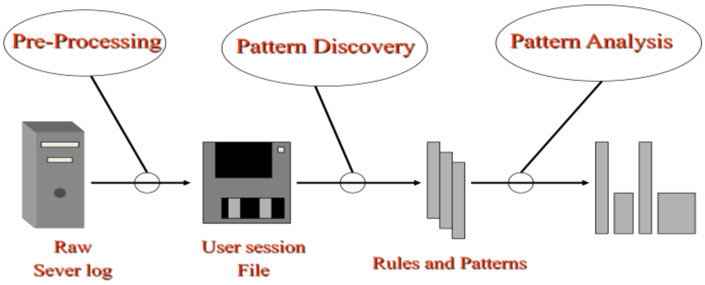

Fig2. Steps of Web Usage Mining

As shown in fig, web usage mining contains the three phases of life cycle as pre processing, pattern discovering and pattern analysis as described:

Pre-Processing: in this initial part of the cycle, the system deals with the different logs as per programmed in the system. In this basically the raw data of user log is derived and pre processed to extract the wanted data. This information needs to be integrated to form a complete data set for data mining. However, before the integration of the data, Web log files need to be cleaned/filtered, using techniques like filtering the raw data to eliminate outliers and/or irrelevant items, grouping individual page accesses into semantic units.

Filtering the raw data to eliminate irrelevant items is important for web traffic analysis. Elimination of irrelevant items can be accomplished by checking the suffix of the URL name, which tells you what format these kind of files are. For example, the embedded graphics can be filtered out from the Web log file, whose suffix is usually the form of “gif”, “jpeg”, “jpg”, “GIF”, “JPEG”, “JPG”, can be removed. It is used to convert the different data into the data abstractions needs for pattern discovery.

Pattern Discovery: It draws the methods and algorithms developed from the preprocessed data in web mining. There are number of algorithms that are used to discover the requested data as association rules, used to discover unordered correlation between items found in a database of transactions. the support is the percentage of the transactions that contain a given pattern. The Web designers can restructure their Web sites efficiently with the help of the presence or absence of the association rules. then loading a page from a remote site, association rules can be used as a trigger for prefetching documents to reduce user perceived latency. Similarly the algorithms like clustering; Clustering analysis is a technique to group together users or data items (pages) with the similar characteristics. Clustering of user information or pages can facilitate the development and execution of future marketing strategies. Clustering of users will help to discover the group of users, who have similar navigation pattern. It’s very useful for inferring user demographics to perform market segmentation in E-commerce applications or provide personalized Web content to the individual users. The Classification, the technique to map a data item into one of several predefined classes.

Pattern Analysis: Pattern Analysis is a final stage of the whole Web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output o his can be done with the help of some analysis methodologies and tools. There are two most common approaches for the patter analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations of the pattern discovery process.

III. ARCHITECTURE OF WEB USAGE MINING

Data cleaning is the first step performed in the Web usage mining process. Some low level data integration tasks may also be performed at this stage, such as combining multiple logs, incorporating referrer logs, etc. After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. The goal of transaction identification is to create meaningful clusters of references for each user. The task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. The input and output transaction formats match so that any number of modules to be combined in any order, as the data analyst sees fit.

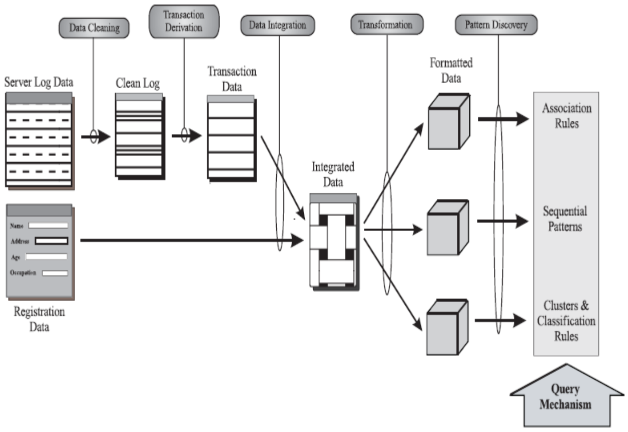

Fig 3. Architecture of Web Usage Mining Process

Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data mining task. For instance, the Format of the data for the association rule discovery task may be different than the format necessary for mining sequential patterns. Finally, a query mechanism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints.

IV. ARTIFICIAL COLONY OPTIMIZATION

Definition: *Ant Colony Optimization is a technique for optimization that was introduced in nearly 1990s. The inspiration of ACO algorithms are the behaviour of real ant colonies.*

One problem with the way in which ants find the shortest path naturally involves the addition of a new shorter path after the ants have converged to a longer path. Because pheromone has built up on the older longer path the ants will not take the newer shorter path. This would be a major problem if the algorithm is applied to dynamic problems in computing. However a simple solution was created. If the virtual pheromone evaporates then the longest path will eventually be abandoned for the shortest one. This is because some ants will still take the shorter path by chance (at least in a virtual probabilistic system) and so pheromone will build up and eventually overtake the longer one. This pheromone evaporation does occur in the wild but it is much slower than equivalent computer based implementations. The way in which ants find the shortest path has been used to create the ant colony paradigm, which is commonly used in problems with large search spaces.

Real Ant Behaviours

The social behaviour of ants is among the most complex in the insect world. They communicate by touching and smelling their odour. There are many behaviours of ants through which they communicate and perform different tasks. Some of those behaviors' are described below:

Ant Foraging: Ants form and maintain a line to their food source by laying a trail of pheromone, i.e. a chemical to which other members of the same species are very sensitive. They deposit a certain amount of pheromone while walking, and each ant prefers to follow a direction rich in pheromone. This enables the ant colony to quickly find the shortest route.

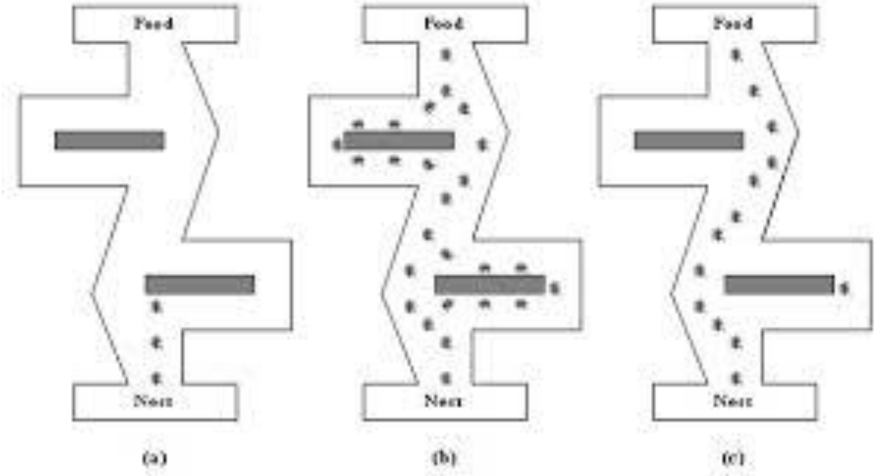

Fig 4. Food Chain

The first ants to return should normally be those on the shortest route, so this will be the first to be doubly marked by pheromone (once in each direction). Thus other ants will be more attracted to this route than to longer ones not yet doubly marked, which means it will become even more strongly marked with pheromone. Soon, therefore, nearly all the ants will choose this route. These algorithms form the heart of a new research field called "***Ant Optimization***".

Colonial Odor

Every day, real ants solve a crucial recognition problem when they meet. They have to decide to which nest they belong in order to guarantee the survival of the nest. This phenomenon is known as "**colonial closure**". It relies on continuous exchange and updating of chemical cues on their cuticle. Each ant has own view of its colony odor at given time, and updates it continuously. Ant in this way preserves its nest from being attacked by predators or parasite and reinforces its integration of the nest. These artificial ants are able to construct group of similar objects, a problem which is known as ***data clustering***.

V. ALGORITHM FOR WEB USAGE MINING BASED ON ANT BEHAVIOUR'S

The proposed work aims to develop an algorithm for retrieval of requested data from the users in more efficient ways. In this, the algorithm is develop on the behavior of the real ants. In this an artificial ant colony is developed and concerned over which the assumptions are formed and the implementation of the system is performed on various assumptions.

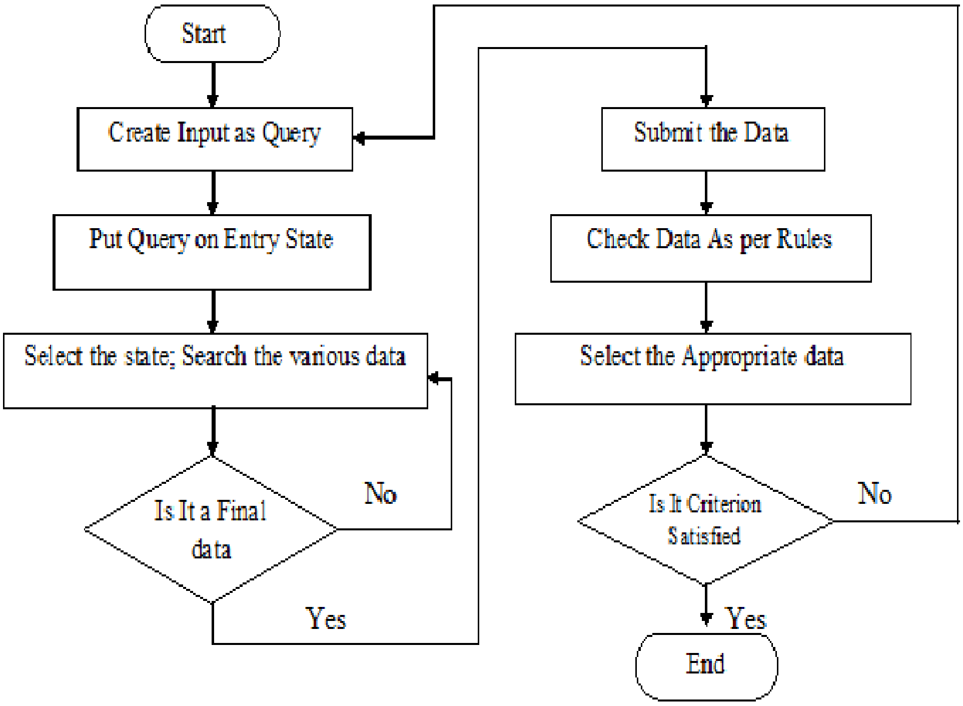

Fig.5 Flow Chart for Mining Using Ant Optimization Colony

Basically in this system, the query entered by the user is concerned as the requested query needs to be processed by the system to show the results as per the request. In this the system preprocesses the raw data, as per the requested query and initially selects the different links to display the various results as per the query.
The selected results are selected after the first phase of process of the system i.e. after pre-processing, now the next step is to to extract the relevant data from the database to the requested query. In this, the different results are examined and the different results are discovered as per the request. The whole process is completed under the second phase of the process.
Now, the final phase of the process is carried out. in this, from the results that are discovered from the pattern discovery phase, it discover the appropriate and required data as per the user request till up to date. In this, using the algorithm condition, the results are again filtered as per the accuracy and concurrency of the data.
These are the three phases based on which the flowchart is designed in the manner to develop the algorithm for the process to retrieve the accurate data.

VI. CONCLUSION

Web Usage mining is an active area of research in the web technology. In this, Ant-based clustering algorithms are an appropriate alternative to traditional clustering algorithms. Using the idea of behavior of the ants, how they function properly in their nest becomes an interesting and meaningful thought for the researchers to research on it and used the concept for the web.
The algorithm has a number of features that make it an interesting study of cluster (group of same data) analysis. It has the ability of automatically discovering the number of clusters. The nature of the algorithm makes it fairly robust to the effects of outliers within the data. This technique has overcome all the limitations up to some extent and yet research in this is continued to make the web more usable and consistent.

REFERENCES

[1] O.A. Mohamed Jafarand R. Sivakumar, "Ant-based Clustering Algorithms: A Brief Survey", International Journal of Computer Theory and Engineering, Vol. 2, No. 5, October, 2010 1793-8201.

[2] Richa Gupta, "Web Mining using Artificial Ant Colonies: A Survey", International Journal of Computer Trends and Technology (IJCTT) – volume 10 number 1 – Apr 2014.

[3] Ajith Abraham and Vitorino Ramos. "Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming", 2004**.**

[4] Nicholas Holden, "Web Page Classification with an Ant Colony Algorithm".

[5] Abhishek Mathur and Trapti Agrawal, "A Survey: Access Patterns Mining Techniques and ACO", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-5, June 2013.

[6] M. Krishna Kumar and Dr. S. Poonkuzhali, "Enhancing the Search Results through Web Structure Mining Using Frequent Pattern Analysis and Linear Correlation Method", International Journal of Electronics Communication and Computer Engineering Volume 4, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209.

[7] G. Anuradha , G. Lavanya Devi and M.S Prasad Babu, “ANTRANK: An ant colony algorithm for ranking WebPages”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 3, Issue 2, March – April 2014, ISSN 2278-6856.

[8] Yezheng Liu, Haifeng Ling and Shanlin Yang, “An artificial colony methodology for users navigation pattern mining”, http://www.paper.edu.cn.

[9] Ankita Kusmakar and Sadhna Mishra, “International Journal of Advanced Research in Computer Science and Software Engineering”, Volume 3, Issue 9, September 2013, ISSN: 2277 128X.

[10] Nicholas Holden and Alex A. Freitas, “Web Page Classification with an Ant Colony Algorithm”.

[11] Raymond Kosala and Hendrik Blockeel, “Web Mining Research: A Survey”, arXiv:cs/0011033v1 [cs.LG] 22 Nov 2000.

[12] Yan wang,” Web Mining and Knowledge Discovery of Usage Patterns”, 2000.