Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Hybrid approach for Clustering Weblog

Chandana S. Khatavkar, Prof. Mangesh Wanjari
Computer Science and Engineering Department,
Autonomous SRCOEM, Nagpur, India

Abstract— Web Usage mining is the application of data mining techniques used to extract useful patterns from the web data. Various web usage mining techniques are used to analyse the user's navigational patterns. Clustering is unsupervised classification of data items into groups called clusters. There are numerous clustering algorithms based on different techniques. This paper proposes a clustering method based on Ant Colony Optimization. For clustering ant-based algorithm is applied to pre-processed logs to extract frequent patterns for pattern discovery then first order Takagi-Sugeno rules are applied for analysing the clustered output.

Keywords— Web usage mining, ant-based clustering, Fuzzy inference rules

I. INTRODUCTION

Web Mining is technique in data mining to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. In Web Mining, data can be collected at the server side, client-side, proxy servers, or obtained from an organization's database (which contains business data or consolidated Web data). There are many kinds of data that can be used in Web Mining. According to data analysis objective, web mining can be divided into three different types, which are web usage mining, web content mining and web structure mining. Web content mining describes the automatic explore of information resources available online, and involves mining of web data content. Web structure mining is the process of analysing hyperlink and tree-like structure of a web site using graph theory. Web usage mining is the process of extracting effective information from web server logs.

Users show different interests when looking for internet. Some users might be looking at only documentary data, whereas some others might be engaged in multimedia data. Web usage mining (WUM) involves the automatic detection of user access patterns from one or more web servers. Organizations rely on internet for their business work which often generate and collect bulk size data in their daily practices. Most of this information is generated automatically by web servers and collected in server access logs. The companies can establish better customer manager relationship by giving them exactly what they require. Companies can understand the requirements and serve them accordingly. They can also increase profitability and productivity based on the profiles generated.

II. RELATED WORK

There are several methods for pattern extraction from the secondary data (web logs). Markov models have been extensively used to model Web users' navigation behaviours on Web sites. [5] Clustering has been widely used in Web Usage Mining to group together similar sessions among large amount of data based on a general idea of distance function which computes the similarity between groups. [10]

In [6], a cluster optimization technique is proposed to improve web usage mining using ant nestmate approach. As the size of the cluster increases, it will become an inevitable need to optimize the clusters. Cluster optimization methodology is based on ant nestmate recognition ability and is used to eliminate the data redundancies. For clustering ART1-nueral network based approach is used. The accuracy and completeness of the user profiles increases by cluster optimization.

In [7] Fuzzy c-means clustering incorporates fuzzy set theoretic concept of partial membership and may result in the formation of overlapping clusters. The algorithm calculates the cluster centers and assigns a membership value to each data item corresponding to every cluster within a range of 0 to 1. [5] adopted a CLIQUE (CLUstering in QUEst) algorithm for clustering web sessions for web personalization.

In [8], the hybrid framework uses an ant colony optimization algorithm to cluster Web usage patterns. This paper proposed an ant clustering algorithm (ACLUSTER) to segregate visitors or find the web usage patterns (data clusters) and a linear genetic programming approach to analyse the visitor trends. [9] Presents how to mine the secondary data (web logs) derived from the users' interaction with the web pages during certain period of Web sessions. At first Ant-based clustering algorithm is applied to pre-processed log files to extract frequent patterns, then it is displayed in an interpretable format and secondly decision tree method is used to find and predict user's navigation behaviour. Decision trees are used in classification and prediction.

Clustering analysis aims to group similar web usage sessions into identical clusters. The process cannot be performed unless WUM data is passed through sophisticated pre-processing steps. We clustered the pre-processed WUM data using a swarm intelligence based optimization, PSO based clustering algorithm. This paper, showed that the performance of the Particle Swarm Optimization (PSO) algorithm is better than K-means clustering.

III. Methodology

Web usage mining is the application of data mining techniques to extract usage patterns from web data. Web usage mining uses data mining techniques to discover useful access patterns from web logs. Web log data is a record of all URLs accessed by users on a Web site. Each log entry consists of access time, IP address, URL viewed, REFERRER (the Web page visited just prior to the current one), etc.

Web usage mining consist of three phase such as pre-processing, pattern discovery, pattern analysis.

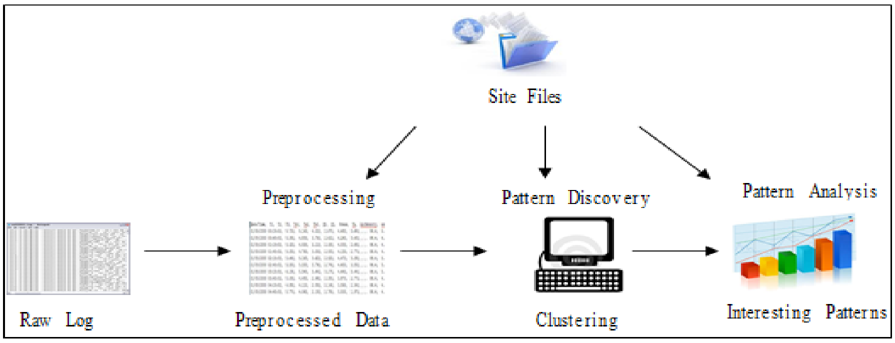

Fig. 1 Web usage mining process

Pre-processing consists of converting the usage, content, and structure information contained in the various available data sources into the data constructs necessary for pattern discovery. The unwanted and unrelated data are removed using the Pre-processing methodology.

Pattern Discovery is the next step after the pre-processing phase. The main objective of this step is to find out the user behaviour and the navigation patterns. For this purpose clustering algorithm is used. Clustering plays a significant role in data analysis and understanding the behaviour of users in the websites. It combines the data into classes or clusters with the intention that the data objects inside a cluster have huge similarity in relationship to one another, but are very dissimilar to those data objects in other clusters.

This paper proposes a method to extract patterns from web logs based on ant clustering algorithm. Many similar methods applied ant colony clustering to segregate visitors[4] but here we have applied ant based clustering for pattern discovery.

Pattern Analysis is the final stage of web usage mining. In this step interesting rules or patterns are extracted from the output of the pattern discovery process. To retrieve the relevant patterns, Takagi Sugeno Fuzzy Inference rules have been applied to the clustered data.

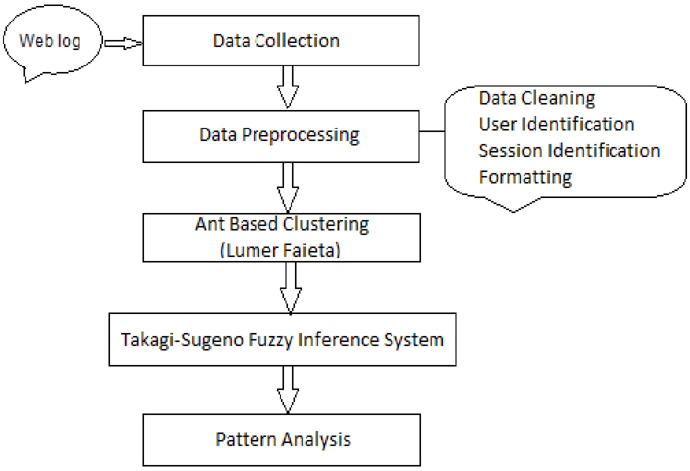

Fig. 2 Proposed Architecture

A. *Data Collection*

The input for the web usage mining process is collected from the web log file. During a user session, all navigation activity on the web site is recorded in a log file by the web server. It is a huge repository of web pages and links, accesses web sites are recorded in web logs file. Log file is available in two formats. The first is the common log format and extended log format.

```
151.48.123.70 - - [08/Dec/2007:00:00:43 -0800] "GET /img/abull.gif HTTP/1.1" 200 411
"http://www.smsync.com/order/?ref=002" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)"
"www.smsync.com"
151.48.123.70 - - [08/Dec/2007:00:00:43 -0800] "GET /img/dowld_btn.gif HTTP/1.1" 200 3083
"http://www.smsync.com/order/?ref=002" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)"
"www.smsync.com"
```

```
151.48.123.70 - - [08/Dec/2007:00:00:44 -0800] "GET /img/buynow_btn.gif HTTP/1.1" 200 2621
"http://www.smsync.com/order/?ref=002" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)"
"www.smsync.com"
200.88.101.168 - - [08/Dec/2007:00:04:37 -0800] "GET /order/main.html HTTP/1.1" 200 9690 "-" "Mozilla/4.0
(compatible; MSIE 6.0; Windows NT 5.1; SV1; InfoPath.1; .NET CLR 1.1.4322)" "www.smsync.com"
200.88.101.168 - - [08/Dec/2007:00:04:38 -0800] "GET /css/main.css HTTP/1.1" 200 7565 "-" "Mozilla/4.0
(compatible; MSIE 6.0; Windows NT 5.1; SV1; InfoPath.1; .NET CLR 1.1.4322)" www.smsync.com
```

B. Pre-processing of Weblog

Pre-processing is the process of preparing log data for further analysis by removing irrelevant data items. The first step in pre-processing is data cleaning. Data cleaning can be done by checking the suffix of URL name and deleting the entries which are of no support to the analysis, such as gif, jpeg, JPG and GIF.

The second step in pre-processing is the User Identification. The required fields are extracted from the cleaned log file and stored in the database for further processing. Here, IP addresses are considered to identify a particular user. After data cleaning and User Identification the user sessions are identified. A request from a particular user within a predefined time period is considered as a user session. Each user session has identified by the session ID.[3]

C. Ant – based Clustering.

The basic ant clustering algorithm was proposed by Deneubourg [1]. In this model, the ants would walk randomly on the workspace and could sense if there in any similarity in surrounding objects or not. Based on this information, they would pick the element or drop the element. The probability of picking and dropping an object depends on the objects lying in immediate environment.

Lumer Faieta (LF) model is an extension of the basic ant model to cluster complex datasets into clusters. The ants would not try to pick or drop anything in areas with low complexity (complexity of ants surrounding area is determined by the presence or absence of objects). The ants can take a deterministic or probabilistic approach. These ants spend less time in random movement in the area of low complexity and more time in careful processing at borders. Each ant-like agent can only sense the similarity of the objects in their immediate region. The probability of picking or dropping an object is a function of this measure of similarity.

For an unladen ant the probability of picking an object increases with low density and decreases with similarity of the objects.[2]

The probability of picking an object i is defined as :

$$P_{pick}(i) = \left(\frac{k^+}{k^+ + f(i)}\right)^2$$

and the probability of dropping an object i is defined as :

$$P_{drop}(i) = \begin{cases} 2f(i)\ if\ f(i) < k^- \\ 1\ \ otherwise \end{cases}$$

where, f is an estimation of the fraction of nearby points occupied by objects of the same type, and K+ is a constant.

D. Fuzzy Inference System

From the clusters, URL and the Fuzzy Inference System session to which the user belongs are determined. After the clustering is performed, the output will be a set of clusters

$n_{p'} = <n_{p1}, n_{p2},\ldots n_{pn}>$ where $n_{pi} = <P_1, P_2,\ldots, P_k>$ where k represents the set of web pages identified as user navigation patterns and $1 \leq i \leq n$.

For pattern analysis fuzzy if-then rules according to first order Sugeno model are considered:

$Rule\ 1: I(x\ is\ A1) and\ (y\ is\ B1) then\ (f1 = p1x + q1y + r1)$

$Rule\ 2: I(x\ is\ A2) and\ (y\ is\ B2) then\ (f2 = p2x + q2y + r2)$

where x and y are represents the inputs, Ai and Bi indicating the fuzzy sets, fi indicates the outputs within the fuzzy region indicated by the fuzzy rule, pi , qi and ri shows the design parameters that are determined while performing training procedure.

IV. CONCLUSIONS

In this paper cluster optimization technique using Ant Colony Optimization is proposed. Web log file is given as input and perform data cleaning to eliminate irrelevant data items. The cleaned web log is used for pattern discovery. The proposed model uses Ant Colony algorithm for clustering based on user sessions. The users with similar access patterns will come under the same cluster. The clusters obtained are feed into the Fuzzy Inference system for analysis of users navigational patterns. Analysis provides better understanding of user's interests.

REFERENCES

[1] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan ,"Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, 1(2), pp. 12-23, 2000.

[2] Saroj Bala, S. I. Ahson, R. P. Agarwal ,"An Improved Model for Ant based Clustering", International Journal of Computer Applications (0975 – 8887) Volume 59– No.20, December 2012.

[3] Nayana Mariya Varghese, Jomina John ,"Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic", IEEE 2012.

[4] Kobra Etminani Mohammad-R. Akbarzadeh-T. Noorali Raeeji Yanehsari ,"Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method", IFSA-EUSFLAT 2009.

[5] Abhishek Mathur, 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] Alphy, S.Prabakaran, "Cluster Optimization for Improved web Usage Mining using Ant- Nestmate Approach", IEEE-International Conference on Recent Trends in Information Technology, June 3-5, 2011.

[7] zahid ansari, a. vinaya babu, waseem ahmed, mohammad fazle azeem ,"a fuzzy set theoretic approach to discover user sessions from web navigational data".

[8] ajith abraham, vitorino ramos "web usage mining using artificial ant colony clustering and linear genetic programming",IEEE 2010.

[9] mrs. v. sujatha, dr. punithaval li ," an approach to user navigation pattern based on ant based clustering and classification using decision trees", 2010.