WEB USAGE MINING BASED ON ACAKHM ALGORITHM

Mr. Jay D Amin
{L. J. Institute of Engineering and Technology
Mentoring Masters in Computer Engineering
Gujarat Technological University
Ahmedabad-Gujarat-India.
Email: aminjay23@gmail.com
Telephone: (+91)9925287421}

Ms. Priyanka K Prajapati
{L. J. Institute of Engineering and Technology
Student of Masters in Computer Engineering
Gujarat Technological University
Ahmedabad-Gujarat-India.
Email:prajapatipriyanka278@gmail.com
Telephone: (+91)7383599539}

Abstract

The web usage mining uses data mining techniques to discover interesting usage patterns from web data. Web personalization uses web usage mining techniques for the process of customization. Customization involves knowledge acquisition done by analysis of user's navigational behavior. A user when goes online would like to get the links which suits his requirements or usage in the website he visits. Clustering is a popular technique of data mining for unsupervised learning in which labels are not defined previously. K-Mean is a well known partitioning technique for forming different clusters, but it has the drawback of initial sensitivity and local optima convergence. K-Harmonic algorithm solves the initial sensitivity problem, but it stuck in local optima problem. The Ant Clustering Algorithm (ACA) can avoid trapping in local optima solution. In this paper, we will propose a new clustering algorithm using Ant Clustering Algorithm with K-Harmonic mean clustering (ACAKHM).

Keywords- Web Mining, Web Usage Mining, Log Files, ACAKHM, Pattern Analysis

I.Introduction

Web usage mining is also called web log mining. Web Usage Mining (WUM) is the approach to extract the knowledge
from analysis of web usage data about a particular website. This usage data can be obtained from server logs and can analyze the behavioral patterns and profiles those interact with the web sites.

In data mining, a method often used is clustering. Object clustering is done based on its characteristics. Companies can use clustering methods to identify patterns of data so that companies can found a certain pattern of the data. Clustering is one of the important data mining technique to discover usage pattern. K-harmonic means clustering algorithm is used with web mining process to discover cluster from web usage data. K-harmonic means clustering algorithm (KHM) can easily runs into local optima. The main drawback of K-harmonic means algorithm is that it is tend to be trapped by local optima. The Ant clustering algorithm (ACA) can avoid trapping in local optimal solution.[4,5]

Web usage mining focuses on two different points: how the website administrators want their websites to be used by the users and how the users actually use these websites.[1]

WEB USAGE MINING OVERVIEW

Web usage mining is a web mining technique which is used for discovery and analysis of web usage patterns from web usage data (or web logs). The main aim of web usage mining is to extract interesting information from web logs and, therefore, helps the website administrators to make personalized and adaptive websites that will better serve the needs of the users visiting their websites.[1]The web usage mining process involves three main steps:

1) Preprocessing
2) Pattern Discovery
3) Pattern Analysis

A.Preprocessing

The preprocessing is the first step in web usage mining process in which firstly the web usage log file is cleaned and transformed so as to remove the useless or noisy data from it and to reduce its size. Then using this cleaned log file, user identification (identifying different users through IP address) Preprocessing and session identification (identifying different sessions) is done.[1]

B.Pattern Discovery

Pattern discovery is the second step in web usage mining process in which the cleaned log file generated in the preprocessing step is used to discover web usage patterns.[1]

C.Pattern Analysis

Pattern analysis is the final step in web usage mining process in which the patterns discovered in the second step are further analyzed to generate more interesting patterns and to find more useful information related to the users browsing patterns. [1]

K-Harmonic Means

K-Harmonic Means is a center-based clustering algorithm. In the K-Harmonic Means any k-th object (k = 1,2, .., n) is considered to be a member of all clusters to-i (i = 1,2, .. c) with membership function value between 0 to 1.In addition, the k- harmonic means has the dynamic weight function to each of the data when the data are far from any center will be given a great weight and the data near the center will be given a small weight. This causes the k-harmonic means is not sensitive to the early initialization so that K-Harmonic Means can work with various initialization (both good and bad). A good initialization is initialization which generate centers spread in all areas of its data.[9]

K-harmonic means (KHM) algorithm is as follows:

1. Determine the value of k as the number of clusters

2. Generate k initial centroids (cluster center) randomly

3. For each data point xi, calculate:

o Membership

$$m(c_j/x_i) = \frac{\|x_i - c_j\|^{-p-2}}{\sum_{j=1}^{k}\|x_i - c_j\|^{-p-2}}, \quad m(c_j/x_i) \in [0,1].$$

o Weight

$$w(x_i) = \frac{\sum_{j=1}^{k}\|x_i - c_j\|^{-p-2}}{\left(\sum_{j=1}^{k}\|x_i - c_j\|^{-p}\right)^2}.$$

4. Calculate the k-th cluster center:

$$c_j = \frac{\sum_{i=1}^{N} m(c_j/x_i)\cdot w(x_i)\cdot x_i}{\sum_{i=1}^{N} m(c_j/x_i)\cdot w(x_i)}.$$

Ant Clustering Algorithm

AC-Algorithm provides an appropriate partition of data. Grid size depends on number of objects and agent ants perform random walks on a grid on which the objects are dispersed randomly. Ants are allowed to move throughout the grid, selecting and dropping the objects inclined by the resemblance and density of the object. The likelihood of selecting up an object will be increased with low density neighborhoods, and decreased with high similarity among objects in the surrounding area.
The picking function and dropping is described as follows:

$$p_{pick}(i) = \left(\frac{k_p}{k_p + f(i)}\right)^2,$$
$$p_{drop}(i) = \begin{cases} 2f(i) & \text{if } f(i) < k_d, \\ 1 & \text{otherwise,} \end{cases}$$

II.RELATED WORK

In The Integrating Between Web Usage Mining And Data Mining Techniques Omer Adel Nassar et.al Clickstream data is one of the most important sources of information in websites usage and customers' behavior in Banks e-services.. While simple traffic analysis based on click stream data may easily be performed to improve the e-banks services. The banks need data mining techniques to substantially improve Banks e-services activities. The relationships between data mining techniques and the Web usage mining are studied. The integration between the Web usage mining and data mining techniques are presented for processes at different stages, including the pattern discovery phases, and introduces banks cases, that have analytical mining technique. A general framework for fully integrating domain Web usage mining and data mining techniques is represented for processes at different stages. Data Mining techniques can be very helpful to the banks for better performance, acquiring new customers, fraud detection in real time, providing segment based products, and analysis of the customers purchase patterns over time.

DATA MINING TECHNIQUES

A. Association

B. Classification (predictive)

C. Clustering (descriptive)

D. Prediction (regression)

E. Sequential Patterns

F. Classification Decision Trees

The Web Usages mining process is presented in Figure 2.1.

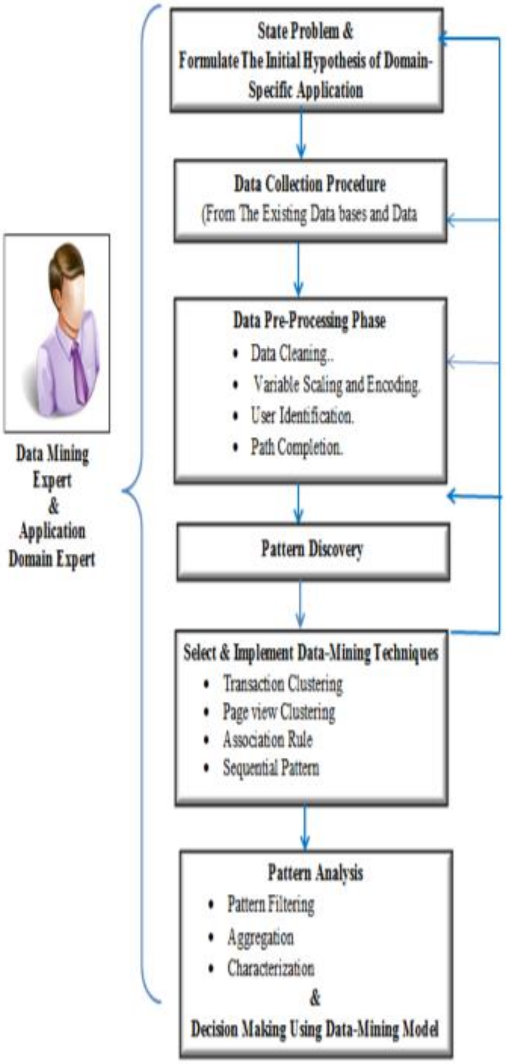
State Problem &
Formulate The Initial Hypothesis of Domain-Specific Application
Data Collection Procedure
(From The Existing Data bases and Data
Data Pre-Processing Phase
• Data Cleaning..
• Variable Scaling and Encoding.
• User Identification.
• Path Completion.
Pattern Discovery
Select & Implement Data-Mining Techniques
• Transaction Clustering
• Page view Clustering
• Association Rule
• Sequential Pattern
Pattern Analysis
• Pattern Filtering
• Aggregation
• Characterization
&
Decision Making Using Data-Mining Model
Data Mining Expert
&
Application Domain Expert

Fig: 2.1 Proposed Framework[1]

In an optimized k-harmonic mean based clustering user navigation patterns R.Gobinath et.al The web mining is an application process which plays an important role in analyzing the behavior of the website users. The web usage mining is a sub category of web mining has a major impact on web personalization. The main concept of the paper deals with the extraction of necessary information from web access log files and applying clustering for easy analyzing of navigational patterns for web personalization. The adapted k-harmonic algorithm is used for clustering the obtained navigational patterns from the various iterating process.

The proposed framework is based on the following process.

1) Data collection

2) Pre-processing

3) Feature extraction

4) Pattern Discovery

5) Pattern analysis

The methodology involved in this paper is shown in the following architecture.

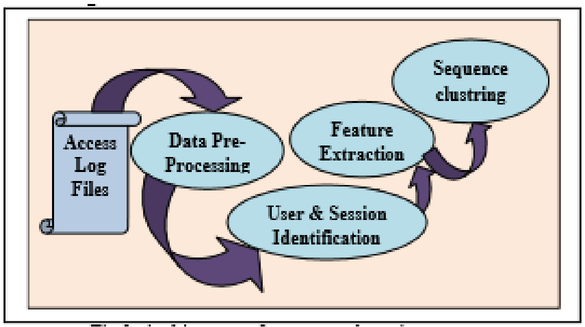

Fig: 2.2 Architecture of sequence clustering process

In Cluster Optimization For Improved Web Usage Mining Using Ant Nestmate Approach Anna Alphy et.al This paper proposes a cluster optimizing methodology based on ants nestmate recognition ability and is used for eliminating the data redundancies that may occur after the clustering done by the web usage mining methods. For clustering an ART1-neural network based approach is used. “AntNestmate approach for cluster optimization” is presented to personalize web page clusters of target users.

Pattern recognition based on web usage mining

Web usage mining deals with the automatic discovery of user access patterns from web log files. This includes the following step.

1. Preprocess Web log file to extract user sessions.

2. Pattern discovery based on data mining methods

This paper uses the ART1 neural network based clustering technique to group the user sessions. The clusters obtained are optimized using the proposed Ant Cluster Track Algorithm

3. Generate user profiles from clusters

4. Track Evolving User Profiles.

In An Efficient Hybrid Data Clustering Method Based On Candidate Group Search And Genetic Algorithm Suvarna P. Patil **et.al** K-Mean is a well known partitioning technique for forming different clusters, but it has the drawback of initial sensitivity and local optima convergence. K-Harmonic algorithm solves the initial sensitivity problem, but it stuck in local optima problem. Genetic algorithm is an efficient tool of the search and optimization problems, which offers the benefits like selective search. In this paper, presents a new scheme in which the initial centroids are calculated using the Candidate Group Search which results in reduction of time for genetic process. Genetic algorithm is used to assign the data elements to the suitable cluster.

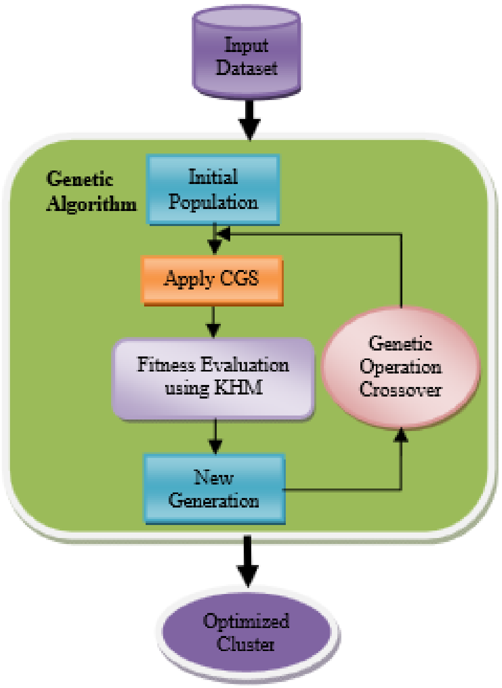

Fig. 2.3 Proposed Model

In An Efficient Prediction Based On Web User Simulation Approach Using Modified Ant Optimization Model And Hierarchical Clustering Trapti Agrawal et.al The proposed work and innovative research is on the basis of the improved next node election, through which we get an improvement in basis ant learning behavior algorithm. This modified ant behavior learning algorithm predicts a larger matching sequence size of real website user sessions along with increased learning rate of the software agents that means most of the ants reach to the uppermost threshold most of the time which directly turns into increased prediction correct rate.

Proposed algorithm modifies this next node selection as below:

Ants have to take decision about the next node after utility has been calculated on the current node.

$$\text{utility} = \frac{\langle \mu \cdot L_p \rangle}{||\mu||\,||L_p||}$$

Pheromone levels of each degree node of the current node are compared and the strongest one is selected. Further at the node with strongest pheromone link, ant compares its preference vector with the LP vector of the node. Only if the similarity is greater than a specified threshold, that node is selected as the next node. Otherwise node with next strongest pheromone value is tested with the same.

The modified algorithm performs better in terms of following parameters:

Matching Sequence Size: This algorithm predicts the session till four nodes with improved efficiency while in the base paper they showed efficiency only up to three nodes.

Learning Rate of the Ant: Here software agents learn faster i.e. from starting iterations they can predict the session sequence size four while in the base paper, after 200 iterations learning algorithm starts matching 3 nodes.

Improved Efficiency: Increased efficiency of the algorithm in terms of comparison of one node, two nodes, and three nodes of the artificial session with the respective nodes of representative real session.

III.Conclusion

This paper deals with a cluster optimization technique. The web log is accessed and performs data cleaning. The cleaned web log is used for pattern analysis. This paper uses the clustering technique for discovering interesting usage patterns. Clustering is done based on user identification and session identification. In this paper we present a new algorithm using the Ant clustering algorithm with K-haromonic means clustering (ACAKHM).It overcomes initialization sensitivity of KM and KHM, and reaches a global optimal effectively. The result of ACAKHM algorithm is better than the KHM algorithm. The accurateness of the results of proposed method is improved over the existing methods.

In the future, we intent to improve the Ant clustering algorithm so as to reduce the runtime of the ACAKHM algorithm. Moreover, we are planning to study the KHM algorithm with other combinatorial optimization technique.

References

[1] R. Gobinath, M. Hemalatha “An Optimized k-Harmonic Mean Based Clustering User Navigation Patterns ” International Conference on Computational Intelligence and Computing Research, 2013. 978-1-4799-1597-2/13/ 2013 IEEE

[2] Omer Adel Nassar , Dr. Nedhal A. Al Saiyd "The Integrating Between Web Usage Mining and Data Mining Techniques." International Conference on Computer Science and Information Technology (CSIT) *2014 International Conference on*, pp. 978-1-4673-5825-2013 IEEE

[3] Anna Alphy , S. Prabakaran " Cluster Optimization for Improved Web Usage Mining using Ant Nestmate Approach" International Conference on Recent Trends in Information Technology *Fourth International Conference on*, pp. 978-1-4577-0590-, 2011

[4] Suvarna P. Patil, Anuradha D. Thakare, C. A. Dhote. "An efficient hybrid data clustering method based on Candidate Group Search and Genetic Algorithm.". 978-1-4673-7231-2/15 IEEE, 2015.

[5] Trapti Agrawal, Shailendra Srivastava, Abhishek Mathur "An Efficient Prediction Based on Web User Simulation Approach Using Modified Ant Optimization Model and Hierarchical Clustering." International Conference on Machine Intelligence Research and Advancement,*2013 International Conference on*, pp. 978-0-7695-5013-. IEEE, 2013