Fuzzy Clustering with Artificial Bee Colony Algorithm using Web Usage Mining

K.Abirami  
Research Scholar, School of Computing Sciences,  
Vels University,  
Chennai, India  
abiramidharmarajan@gmail.com

Dr. P.Mayilvaganan  
Professor, Department of MCA,  
School of Computing Sciences,  
Vels University,  
Chennai, India  
mayilkadir@yahoo.com

Abstract— Now a days the internet is very useful to human being. The main difficult is user to discover the precise information or any items from internet is exhausting and time consuming extremely tedious task. There are many existing methods in statistical analysis, pattern matching, machine learning, and databases. It is very few attempt at bringing the hidden relationship between web pages and its users. The proposed work is applied the Personalized Artificial Bee Colony (PABC) fuzzy clustering algorithm to web usage mining. This algorithm find the web user pattern recognition and get the user navigational behavior. The results indicate that the performance of Personalized Artificial Bee Colony algorithm is successful in fuzzy clustering in web usage mining to calculate the web user pattern. Interests and habits, browsing sequences, classification and predict feature user behavior. The result can improve the navigation effectiveness of website user. Then the three different weblog datasets have been used to demonstrate the results of the algorithm. The experimental results are encouraging in terms of the quality of the high solutions. This work could be effectively used for web personalization.

Keywords—Web Mining, Web Usage Mining; Pattern Discovery, Web Personalization; PABC Algorithm

I. INTRODUCTION

Web data mining is defined as the process of discovering meaningful patterns/knowledge from data [1]. In this internet era, the World Wide Web has become a major source of communication and vast source of information in everyday life. The growth of the internet over the last two decades has resulted in a large amount of data that is available for user access. The user interactions with the Web are recorded and stored in web access logs. Web mining [1] [2] is the use of data mining techniques to automatically discover and extract information from web documents and services.

Web mining can be categorized into three areas of interest based on which part of the web to mine: 1) Web content mining: refers to discovery of useful information or knowledge from web page contents i.e. text, multimedia data like images, audio, video etc. 2) Web structure mining: aims at analyzing, discovering and modeling link structure of web pages and/or web site to generate structural summary. 3) Web usage mining deals with understanding user behavior while interacting with web site, by using various log files to extract knowledge from them [1].

Web mining are classified in four categorized. 1. Data collection 2. Preprocessing 3. Pattern discovery 4. Pattern analysis. In this paper, to discover the web user pattern for web page. In first phase, the data would collected to the web log files from web server. In second phase, preprocessing the web log files. That is web log files are cleaned. The Preprocessing are categorized in four stages. First, the data are cleaned in unwanted, noisy, irrelevant and meaningless.

Second, User Identification phase would identify the IP address of the web site user. Third, Session Identification phase would identify the user session time that is logged in and log out were calculated. Fourth, Path completion phase is when the web user visit the web page [14].

The web data mining functionalities and the variety of knowledge discover are briefly given below.

A. Characterization

Data characterization is a summarization of overall features of objects in a target class, and produces what is called characteristic rules [3]. The web log data relevant to a user-specified class are normally retrieved by a web server and run through a summarization module to extract the essence of the web data at different levels of abstractions [4].

B. Discrimination

Data discrimination produces what are called discriminant rules and they are basically a comparison of the general features of objects between two classes referred to as the target class and the contrasting class [5].
C. Association analysis

Association analysis is the finding of what are commonly called association rules [9]. It readings the frequency of items occurring together in transactional lists, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability that an item appears in a transaction when another item appears, was used to pinpoint association rules. Association analysis is commonly used for market basket analysis.

D. Classification

Classification analysis is the organization of web data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects [10].

E. Prediction

Prediction is an area of data mining that deals with extracting useful information from data and using it to predict trends and behavior patterns. There are two major types of predictions: one can either try to predict some unobtainable data values or incomplete styles, or predict a class label for some data [8] [12]. Once a classification model is built based on a training set, the class label of an object can be expected based on the attribute values of the object and the attribute values of the classes.

Prediction is however more often referred to as the calculation of missing numerical values, or increase / decrease trends in time related data [10]. The major idea is to use a huge number of past values to consider possible future values [11].

F. Clustering

Clustering is the group of web data in classes like web user. However, dissimilar classification, in clustering, class labels are unfamiliar and it is up to the clustering algorithm to discover suitable classes. Clustering is also called unsupervised classification [12]. There are many clustering approaches and all are based on the principle of maximizing the similarity between web users in the same web page and minimizing the similarity between web users of different web pages.

In this research paper, clustering concept only used to identify the similar user pattern. Although, the pattern discovery phase, we introduced the Personalized Artificial Bee Colony algorithm is used to discover the web user pattern [7].

II. CLUSTERING METHOD

Clustering is the process of grouping web data into a number of clusters. The goal clustering is to make the data in the same cluster share a high degree of similarity [6]. Distance measurement is generally used for evaluating similarities between web users [5]. The fuzzy clustering algorithms result in values between 0 and 1 that indicate the degree of membership for each object to each of the clusters. The clustering methods are classified in Fig 1:

![Cluster Methods Diagram]

- Hierarchical
- Partitioning
- Density Based
- Model based
- Grid based
- Soft computing
- Agglomerative
- Divisive
- Error Minimization
- Graph-Theoretic
- Decision Tree
- Neural network
- Fuzzy clustering
- Evolutionary Approaches
- Simulated Annealing
III. ARTIFICIAL BEE COLONY ALGORITHM

The Artificial Bee colony algorithm consists of three different ways of using the web user visit the web page for web usage mining[2][3].

(i) More visited webpages And More Spend Time (employed bees) – MPMT web user
(ii) Less visited webpages And More Spend Time (onlookers) – LPMT web user
(iii) Less visited webpage And Less Spend Time (scout) – LPLT web user

This research paper, we introduced the artificial bee colony algorithm using web usage mining. Now purposely used this algorithm, such that all bees like web user visit the web page[13]. The web user divided into three types using an artificial bee colony algorithm [4]. First employed bees consists more content and more spend time of viewing the web page of web user like MCMT, onlookers consists of less content and more spend time of viewing the web page of web users like LCMT and scouts consists of less content and less spend time of viewing the web page of web user like LCLT [14].

In the Artificial Bee colony algorithm MCMT phase, each visitor finds a new web page \( W_i \) in the neighborhood of its current web page \( C_i \). The new web page is calculated using the following expression:

\[
W_{ij} = C_{ij} + \phi(C_{ij} - C_i)
\]

where, \( C_{ij} \) denotes the \( j \text{-th element of } C_i \) and \( j \) is the random index, \( C_i \) denotes neighbor or current web page selected randomly from the web site. And \( \phi_j \) is the uniform random number \([-1,1] \). \( k \rightarrow \{1,2,\ldots \text{ Number of Web User}, \} \) \( j \rightarrow \{1,2,\ldots \text{D}\} \) are randomly chosen index, \( D \rightarrow \text{Number of path to optimize, } \) \( k \neq i \rightarrow \text{both are randomly chosen}. \)

If the web user to visit the new web page is higher than that of its currently associated a web page, then this MPMT web user(employed bees) moves to this new web page, otherwise it continues with the old one. After all MPMT web user complete the search process, they share the information about their web page with LPMT (onlooker bees) web user. An LPMT Web user evaluates the web page information taken from all MPMT web users and discover the patterns of a web user with a probability related to its web page size. This method, known as roulette wheel selection method. It provides better web user navigation pattern to discover.
Load the preprocessed web log file
Initialize fuzzy web page \( Y_i, i=1,2,\ldots, wp \)
Evaluate the web user
Set cycle to 1
Repeat
FOR each MPMT web user
Visit new web page \( W_i \) by (5)
Calculate the web user properties
Apply the fuzzy clustering method
FOR each LPMT
Choose a webpage \( Y_i \) depending on \( p_i \)
Visit new webpage \( W_i \)
Calculate the web user properties
End

The flowchart for the proposed system is shown in Fig 2. The formless data is prepared for processing by computing the fuzzy clustering method and discover maximum no. of web user patterns. The PABC fuzzy clustering method is originated to compute the user pattern of the web page[3][4]. This process has been repeated to attain the efficient classification precision.

The steps of PABC fuzzy clustering method using web log mining following below:

1. Load the preprocessed web log file
2. Initialize fuzzy web page \( Y_i, i=1,2,\ldots, wp \)
3. Evaluate the web user
4. Set cycle to 1
5. Repeat
   - FOR each MPMT web user
     - Visit new web page \( W_i \) by (5)
     - Calculate the web user properties
     - Apply the fuzzy clustering method
   - FOR each LPMT
     - Choose a webpage \( Y_i \) depending on \( p_i \)
     - Visit new webpage \( W_i \)
     - Calculate the web user properties
6. Until cycle = Max. No. of webpage viewer based on content and spend time
7. Obtain accuracy of web user

\[ P_i = \frac{C_{T_i}}{\sum_{j=1}^{wp} C_{T_j}} \]  

(6)

Where \( C_{T_i} \) is the content of the web page, \( i \) proportional to the memory size of the web page to the web in the position \( i \) and \( wp \) is the number of web page equal to the number of MPMT web users. Once all LPMT web users have selected their web pages, each of them determines a new neighboring web page of its selected web page and computes its web page size [8]. The web user memorizes the new position and forgets the old one whether its web page is higher than that of the previous one; otherwise it keeps that.

The MPMT web user becomes a LPLT (scout) web user when a web page is exhausted by the MPMT and LPMT web users [3][7]. Any position cannot be improved further through a predetermined number of cycles which is called limit parameter, the web page is assigned as abandoned and MPMT web user of that web page becomes LPLT web user. In that position, a new solution is randomly generated by the LPLT web user, given in following Equation, where unrestricted web page is represented by \( Y_i \) and \( j \in \{1,2,\ldots,D\} \). Where \( D \) is the path of the web site, the LPLT web user discovers a new web page which will be replaced with \( Y_i \)

\[ Y'_i = Y'_{\text{new}} + \text{rand}(0,1) (Y'_{\text{new}} - Y_i) \]  

(7)

Each search cycle has three steps after initialization: moving MPMT users to view web page and discover the user patterns; placing LPMT users onto view web page and calculating the webpage viewer; determining LPLT users and directing them to possible web pages. In this way, to discover the pattern of webpage viewer.

\[ P_i = \frac{C_{T_i}}{\sum_{j=1}^{wp} C_{T_j}} \]  

(6)

Where \( C_{T_i} \) is the content of the web page, \( i \) proportional to the memory size of the web page to the web in the position \( i \) and \( wp \) is the number of web page equal to the number of MPMT web users. Once all LPMT web users have selected their web pages, each of them determines a new neighboring web page of its selected web page and computes its web page size [8]. The web user memorizes the new position and forgets the old one whether its web page is higher than that of the previous one; otherwise it keeps that.

The MPMT web user becomes a LPLT (scout) web user when a web page is exhausted by the MPMT and LPMT web users [3][7]. Any position cannot be improved further through a predetermined number of cycles which is called limit parameter, the web page is assigned as abandoned and MPMT web user of that web page becomes LPLT web user. In that position, a new solution is randomly generated by the LPLT web user, given in following Equation, where unrestricted web page is represented by \( Y_i \) and \( j \in \{1,2,\ldots,D\} \). Where \( D \) is the path of the web site, the LPLT web user discovers a new web page which will be replaced with \( Y_i \)
IV. EXPERIMENTAL RESULT

We implement the algorithm Personalized Artificial Bee Colony (PABC). All the experiments are conducted on a Intel(R) Core(TM)2 Duo CPU T5470 @1.60GHz with RAM 2.0 GB and 500GB main memory, running Microsoft Windows 7 64 bit Operating System.

This research work used the data collected from the web server, www.shivaquapaddlaraetrs.com (Dataset1), www.collegenotes.org (Dataset 2) and www.veluniversity.ac.in (Dataset 3) website from 01 July 2016 to 31 July 2017. This collected data is analyzed by using Web Log Expert Lite 9.2 web mining tool. The complete experimental analysis was done on the basis of web log data of an educational institutions website. The design and execution of such work is restricted and time consuming.

In this research paper the performance of the proposed approach is evaluated based on recall, precision and F-measure values. The analysis is accompanied based on two measures. Initially, websites which are web user frequent pattern by the visitor are retrieved. Secondly, the likeness values of the user interested webpages are analyzed.

**Performance calculation Metrics:**

The performance of the proposed algorithm is calculated based on the following three rudimentary evaluation criteria which follow information retrieval context

A. **Recall:**
Recall measures the probability ratio of retrieving relevant users by the system. Recall is the ratio of the difference between the number of retrieved users and relevant users to the number of relevant users in a particular website [15].

$$\text{Recall} = \frac{W_{ret} \cap W_{rel}}{W_{ret}}$$

Where, Wret and Wrel are the number of retrieved and relevant website user correspondingly.

**TABLE I. PERFORMANCE ANALYSIS BASED ON RECALL**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
<th>Dataset 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ant colony optimization</td>
<td>0.68</td>
<td>0.71</td>
<td>0.62</td>
</tr>
<tr>
<td>K-harmonic means(KHM)</td>
<td>0.71</td>
<td>0.78</td>
<td>0.72</td>
</tr>
<tr>
<td>PABC</td>
<td>0.79</td>
<td>0.82</td>
<td>0.83</td>
</tr>
</tbody>
</table>

The above Table I and Fig 3 shows that the proposed work has highest recall rate 0.79, 0.82 and 0.83 for Dataset 1, Dataset 2 and Dataset 3 respectively. If the recall process, the PABC result is better than ACO and KHM.

B. **Precision:**

This technique is to identify the frequent user and top webpages. It measures the percentage of relevant users retrieved from the website. Precision takes all retrieved users into account and consider only the frequent users returned by the website. The ratio of the difference between the number of relevant users and retrieved users to the number of the retrieved users is given as the precision measure

$$\text{Precision} = \frac{W_{rel} \cap W_{ret}}{W_{ret}}$$

**TABLE II. PERFORMANCE ANALYSIS BASED ON PRECISION**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
<th>Dataset 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ant colony optimization</td>
<td>0.63</td>
<td>0.69</td>
<td>0.78</td>
</tr>
<tr>
<td>K-harmonic means(KHM)</td>
<td>0.81</td>
<td>0.72</td>
<td>0.8</td>
</tr>
<tr>
<td>PABC</td>
<td>0.89</td>
<td>0.91</td>
<td>0.92</td>
</tr>
</tbody>
</table>
The main reason for finding our proposed work is that, for a large web site, there is a small number of web users with similar behaviors and thus, we only use these kind of frequent patterns. The PABC fuzzy clustering algorithm that overcomes the limitations of existing clustering algorithms like Ant colony optimization and K-harmonic means. The proposed PABC algorithms are used in the process to enhance efficiency and accuracy. The PABC fuzzy clustering is used to reduce similarity between the web users, analyze the mining result and offer valuable recommendations on the enhancement of the website.

**References**


