International
Journal
Of Advanced Research in Engineering & Management (IJAREM)
ISSN: 2456-2033 || PP. 58-64

AN EFFICIENT APPROACH FOR ANALYSIS PERSONAL BEHAVIOUR IN WEB USAGE MINING

S. Revathi . S[1] M.SC, M.PHIL, S.Ranjitha[2] MCA, M.PHIL
[1]*Cms College Of Science And Commerce, Bharathiyar University, M.Phil Scholar, Coimbatore, Tamilnadu, India,*
[2]*Cms College Of Science And Commerce, Bharathiyar University, Associate Professor, Coimbatore, Tamilnadu, India,*

Abstract: Web usage mining refers to the automatic discovery and analysis of patterns in click stream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. An important task in any data mining application is the creation of a suitable target data set to which data mining and statistical algorithms can be applied. In this paper we proposed the usage of web pages using two different clustering algorithms such as k-means, and Fuzzy c means (FCM) clustering using MATLAB.
Keywords: Web mining, Web usage mining, MATLAB, Fuzzy C Means

Introduction:

The process of extracting interesting information from web log file is called as Web Usage Mining; and this is method consists of four task from start to reading web log file till increasing the virtualization on the web log file referred as the input phase, the preprocessing phase, and the pattern discovery phase, and pattern analysis phase. In the input phase, few types of untreated web server log files are retrieved and separated by their records as access, referrer and agent logs. . In preprocessing, the most common tasks are data cleaning, data filtering, session identification, user identification, and path completion. As per the researcher's taking part, not all of the models handled in the pattern discovery phase but fairly speaking it would be considered important it any of them is suitable for their research work. In last phase Pattern analysis, individual analysts the serve's details and more knowledgeable, useful and actionable models.
This paper presenting the work strongly, and this working towards web usage mining for predicting access behavior. For this, we are using web log file based on developed platform named MATLAB. Data mining can also be used to extract the knowledge from E-learning system such as Module [1].

Web mining:

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.). Commonly, data mining uses structured data stored in relational tables, spread sheets, or flat files in tabular form. In general, Web mining can be divided into three separate categories depending on the type of data to explore: Web Structure Mining, Web Content Mining and Web Usage Mining. Web usage mining is that the method of extracting helpful Information from server logs . The actual processing of this paper is presented as follows.

BLOCK DIAGRAM:

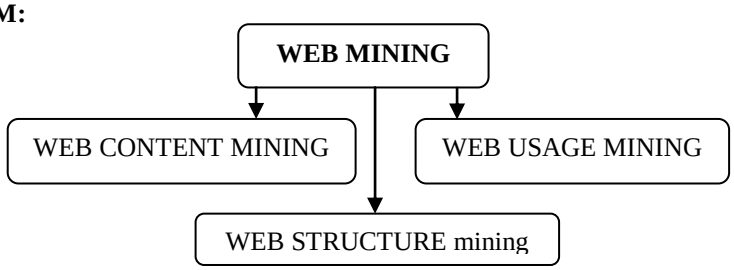

Figure 1: Data Mining

Web content mining:

It aims to extract information linking to the website page contents. It extracts or mines the useful information or knowledge from Web pages. For example, we can automatically categorize and cluster Web pages according to their topics. These tasks are similar to those in usual data mining. However, we can also realize patterns in Web pages to extract the useful data such as descriptions of products, postings of forums, etc, for several purposes. Furthermore, we can mine customer reviews and forum postings to discover consumer sentiments. These are not conventional data mining tasks [3].

Web usage mining:

It is also known as Web log mining, is used to analyze the behavior of website users. It refers to the discovery of user access patterns from Web usage logs, which record every click made by each user. Web usage mining applies many data mining algorithms. One of the key issues in Web usage mining is the pre- processing of click stream data in usage logs in order to produce the right data for mining. These data is used for generating histories about selected special events [4].

Two main approaches are used in Web Content Mining:

(1) Unstructured text mining approach

(2) Semi-Structured and Structured mining approach.

Unstructured Text Data Mining:

Web content data is a lot of unstructured text data. The research around applying data mining techniques is to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Hence, one could think text mining is an instance of the Web content mining. To provide effectively utilizable of the results, preprocessing steps for any structured data is done by way of information mining, and text categorization, or applying the NLP techniques. In OEM, data is in the form of atomic or compound objects: atomic objects may be integers or strings; compound objects refer to other objects through label edges. HTML is a special case of such intra-document structure [4].

DATA COLLECTION:

An important task in any data mining application is the creation of a suitable target data set to which data mining and statistical algorithms can be applied. This is particularly important in Web usage mining due to the characteristics of click stream data and its relationship to other related data collected from multiple sources and across multiple channels. And this process is critical to successful extraction of the useful patterns from the dataset. The process may involve to pre-processing the original data, and integrating the data from multiple sources, and also transforming the integrated data into the form suitable for the entire input into the specific data mining operations in the data [4].

SERVER LEVEL COLLECTION:

A Web server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the (possibly concurrent) access of a Web site by multiple users. However, the site usage data recorded by server logs may not be entirely reliable due to the presence of various levels of caching within the Web environment. Cached page views are not recorded in a server log. In Packet sniffing technology is an alternative method to collecting the web usage data are using through server logs. And also Packet sniffers monitoring the network traffic coming to a Web server and extract the web usage data directly from TCP/IP packets. The Web server can also store other kinds of usage information such as cookies and query data in separate logs. Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the site visitors. Once the CGI program has completed its execution, the Web server sends the output of the CGI application back to the browser.

CLIENT LEVEL COLLECTION:

Client-side data collection can be implement by using the remote agent (such as Java scripts and Java applets) or by modify the source code of the presented browser (such as Mosaic or Mozilla) to enhance the data collection of capabilities. The implementation of client-side data collection of methods requires to the user cooperation, either enable the functionality of the Java scripts and Java applets, or to freely use the adapted

browser. Client-side collection has advantage over the server-side collection because it is the ameliorates of the both caching and session identification of the problems [5].

PROXY LEVEL COLLECTION:

A Web proxy acts as an intermediate level of the caching between the client browsers and Web servers. Proxy caching (can be used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides. The performance of proxy caches depends on their ability to predict future page requests correctly.

PROPOSED METHOD:

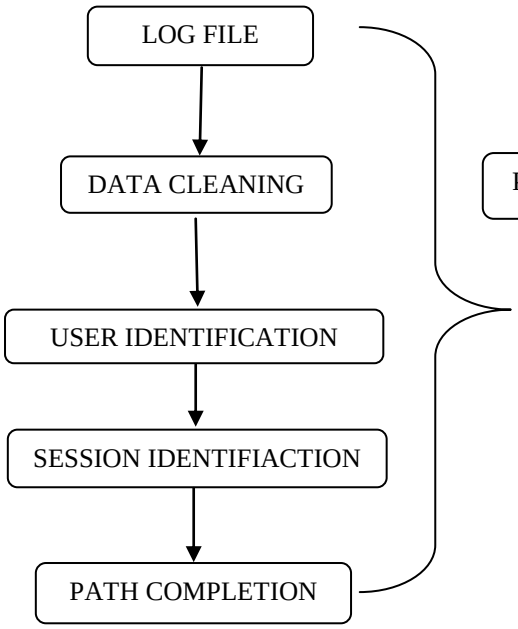

Figure 2: Proposed work design

LOG FILE:

This log is used for get the information about the user's details. Automatically user details will be stored in log file. Log file is a plain text file and which records the information about each and every user which includes name, and also user IP address, and also date, user login time, and how many bytes the user transferred, when they have access the request. The Web server also can writes the information about the log file each and every time user requests a resource from that particular site. When user submits their request to a web server that activity is recorded in to the web log file. Log file size ranges from 1KB to 100MB[4].

Log files give us information about:

1. Pages requested in website.
2. Bytes sent from server to user.
3. Type of error occurred

CLIENT BROWSER:

These log files reside in client's browser window itself. This information can be recorded only if cookies are enabled. Cookies are pieces of information generated by a web server and stored in user's computer C. Log File formats Data in log files exists in three different formats

- W3C Extended log file format.

This format is default log file format on IIS server. Fields are separated by space. Time is recorded as GMT (Greenwich Mean Time).Details store like Year is recorded as YYYY-MMDD.

It contains fields:

- Software used

- Version of Software
- Date and Time of access
- IP Address of user
- Method URI stream

PREPROCESSING:

- This Data preprocessing is transforms the data into a format that will be extra easily and also efficiently processed for the intention to feature user. The main mission of this data preprocessing is to be select the equal and also same data from the original log files, prepared for user steering pattern discovery algorithm. The main point of data preprocessing is includes the data cleaning, and second think user identification and session identification.

DATA CLEANING:

- We can do this data cleaning after get the log file. This is using for performed to clean the unnecessary data in log file. Content which is having the requests for images, styles and script or other files. This stage is using for filtering the unwanted data from the log file. Through this stage all URLS with jpeg, gif and .css extensions are remove the data with using algorithm. After data cleaning we reduced the unnecessary data or user details from 441 to 500 records.

USER IDENTIFICATION:

- The new user can identify by the using their IP address assign to the particular user. User's identification is, used identify who access which web site and which pages are they have accessed. A session is a series of web pages user browse in a single access. This user identification is difficulties to accomplish this step are introduced by using this proxy servers, e.g. different users may have same IP address in the log. And we can use this user identification and we solve the problem.

SESSION IDENTIFICATION:

- Session identification is used to identify the user session from Web access log file. To identify the sessions from the untreated data is a There are Web server logs that do not hold the enough information to reconstruct the user sessions, If all of the IP address, have browsers and also operating systems are same, the user details information should be taken from the used account[6].

PATH COMPLETION:

- This path completion process of identifying the user references web pages. Using this path completion step is conceded out to find the missing pages. Path Set is the incomplete accessed pages in a user session. It is information extracted from the every user session set [6].

K-Means Clustering Algorithm:

This section describes the original k means clustering algorithm. One of the most popular clustering algorithms is k-means clustering algorithm [5]. The clustering is containing with the group of objects together the similar to each other and other dissimilarity of clusters. There is having many different clustering algorithms like: K-Means, K-Medians, clustering is a method of cluster analysis. This K-Mean is a partition clustering algorithm and this is very useful in smaller datasets to access the datasets. Mainly this algorithm mostly Euclidean distance is used to find distance between data points and centroids.

1. The Euclidean distance between two multidimensional data points X=(x1, x2, x3….xm) and Y= (y1, y2, y3…ym) is described as follows: $d(X, Y) = \text{Sqrt}(x1-y1)^2+(x2-y2)^2+\ldots+(Xm-Ym)^2$
   Mean= Sum of elements / Number of element =a1+a2+a3+.....+an/n **Algorithm:**
   The k-means clustering algorithm
   Input: D:{d1,d2….dn}\\set of n items K //Number of desired clusters
   Output: A set of k-clusters.
   1. Arbitrarily choose k-data items from D as initial centroids;
   2. Repeat assigns each item di to the cluster which has the closest centroid, Calculate new mean for each cluster; until convergence criteria are met.

Also, the k-means procedure is easily programmed and is computationally economical, so that it is feasible to process very large samples on a digital computer. Euclidean distances generally considered to determine the distance between data points and the centroids.

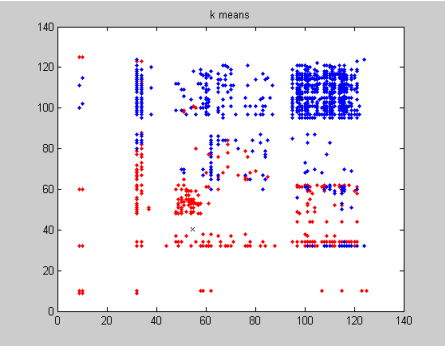

Figure 3: K Means Diagram

Here the K-means are used to carried out the user similarity based on the user what they have used.

Fuzzy C Means Clustering:

Process of this Fuzzy C Means Clustering is to find the same similar object one another. This method is developed by dunn in 1973 and also it has improved by Bezdek in 1981. In this paper, we can use object of user session as time generated by preprocessing stage [5]. Given a set of data objects S = {X1,X2…Xn}, where

$$X = (Xi1, Xi2, \ldots Xil)^P \in R^1$$

This Fuzzy C-means (FCM) clustering algorithm is overlapped clustering which allows one data.

It is based on minimization of the following objective function:

$$j_m = \sum_{i=1}^{N}\sum_{k=1}^{c} U_{ij}^{m}\left\|X_{ki}-X_{kj}\right\|^2, \quad 1<=m< infinity$$

Where, m- is a real number greater than 1,
Uij- is the degree of membership of Xi in the cluster j,
C is the total number of clusters,
N- is the total number of user sessions,
X i is the feature vector,
C j is the center of the cluster, and

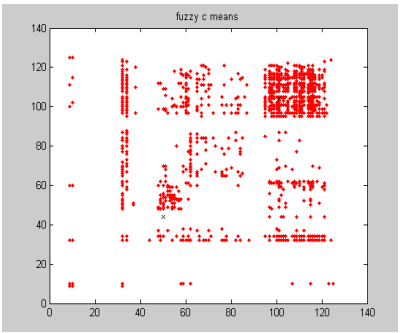

Figure 4: Fuzzy C Means Diagram

The following steps explain the working of FCM:

Input: The feature vector Xi that signify the navigational sequence of the each every user and the number of clusters.
Output: The clusters having users with similar access sequence.
Step 1: Start
Step 2: Initialize or update the fuzzy partition matrix Uij with equation (2)
Step3: Calculate the center vectors Cj using equation (3)
Step 4: Repeat step (2) and (3) until the termination criterion is satisfied.
Step 5: Stop The fuzzy c-means procedure continues until the close standard is satisfied [4-5].

Image Processing:

IMAGES IN MATLAB

When we handling the images in Matlab, there are many things to keep it in mind such as loading the image, use correct format, should store the data as different data types from previous data types. How to show the image, translation between two different image formats, etc.
This the formats for supported by Matlab:

BMP,HDF,JPEG,PCX,TIFF,XWB

we can see the images on the Internet are JPEG-images which is the name for one of the most widely used solidity values for images.

Digital Image Processing

FUNDAMENTALS

The various basic steps are as follows

1. Image Acquisition
2. Image preprocessing
3. Image segmentation
4. Image Representation and Description
5. Image Recognition and Interpretation and
6. Knowledge base.

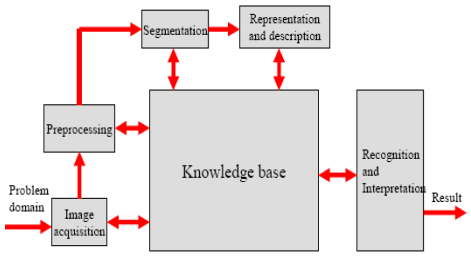

Figure 5: Fundamental steps in digital image processing

i) Image Enhancement:
- More relevant than the original image for particular application.

ii) Image Restoration:
- A process that attempts to build or recuperate an particular image.

MATLAB

MATLAB = Matrix Laboratory

MATLAB is a high-level language. And this fourth generation programming language. MATLAB is an interactive surroundings that enables to perform the tasks faster than the traditional programming languages like: C, C++ and Fortran.

And very easy-to-use the environment where problems and solutions are displaying the familiar mathematical notation.

Typical uses include:
•Math and computation

•Algorithm development

- **MATLAB System:**
  The MATLAB system consists of five main parts:
- **Development Environment**. Using this development environment we can get the help to use MATLAB functions and files. More and more tools are using by the graphical user interfaces. and also includes the MATLAB desktop and also Command Window, to know the a command history, and get get the information for viewing help, the workspace(to work out), files, and the search path.

- **The MATLAB Mathematical Function Library**. This is a vast collection of computational algorithms, commencing elementary functions such as sum, sine, cosine, and also complex arithmetic, functions like matrix inverse, matrix Eigen values, Bessel functions, and fast Fourier transforms. In MATLAB mathematical function have special functions like Bessel are available.
- **The MATLAB Language**. This is a high-level matrix/array language. And scientific numeric analysis programming language.

Accuracy Graph

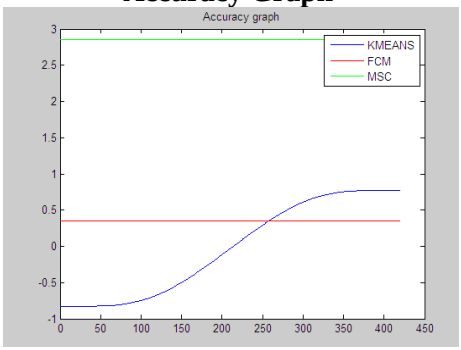

Figure 6: Accuracy Chart

Conclusion:

The main objective of this paper is to clustering the web usage of person and analysis the usage of web pages and extract the content and clustering by efficient K means and Fuzzy C Means are used. To perform the analysis, web access log data has been collected through Internet. In this work, the important step is multi spectral clustering, are used to improve the time complexity. In feature can extended by access multiple websites at a time.

Refrence:

[1]. Sunita B Aher , Lobo L.M.R.J " **Best Combination of Machine Learning Algorithms for Course Recommendation System in E-learning** ", International Journal of Computer Applications , Volume 41– No.6, March 2012.

[2]. P. Sengottuvelan, T. Gopalakrishnan "**Efficient Web Usage Mining Based on K-Medoids Clustering Technique**" International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:4, 2015.

[3]. Mahendra Pratap Singh Dohare1, Premnarayan Arya2, Aruna Bajpai3" **Novel Web Usage Mining for Web Mining Techniques"** International Journal of Emerging Technology and Advanced Engineering , Volume 2, Issue 1, January 2012.

[4]. Vijayashri Losarwar, Dr.Madhuri Joshi **"Data Preprocessing in web usage Mining",** international conference on Artificial intelligence and embedded System,(ICAIES 2012)july 15,2012.

[5]. K.SudheerReddy,G.Partha Sarathi Varma,and M.Kantha Reddy "**An Effective Preprocess Method for Web Usage mining",** International journal of computer theory and Engineering,vol.6,No.5,October 2014.

[6]. P Nithya et al ,Int.J.Computer Technology &Applications,Vol 3 (4), July-August 2012."A Survey on Web Usage Mining: Theory and Applications ".