Asian Journal of Information Technology 15 (20): 3899-3904, 2016
ISSN: 1682-3915

Optimizing Web Log Data to Perceive User Behavior

B. Prasanna Kumar Reddy and Duvvada Rajeswara Rao
Department of Computer Science Engineering, KL University, Guntur, India

Abstract: Day to day the information in world wide web is increasing tremendously along with number of users. So it was difficult for the web application/website admin to maintain huge amount of data about the user and his needs. With the help of web usage mining techniques, the user's behavior can be extracted from log data. It helps in analyzing errors of a website so that website administrator or designer can improve their system. Clustering has a key role in analyzing web log data. In this study we propose aclustering method to help mining log data for understanding user behavior.

Key words: Fuzzy C-means clustering, web log data, SVM, web usage mining, web usage analysis

INTRODUCTION

In this internet world, there is huge amount of information is available which is used for variety of purposes. Server log files is a type of information that holds key data to understand the user. It contains data regarding user includes IP address, request date/time, HTTP code, page requested bytes served etc. (Eltahir and Dafa, 2013). The difficulty lies in finding right information in it. Not all the information available is not useful, there is also lot of irrelevant data included. The main objective here is to find the required information. Web mining is the area that is active presently in extracting information from this kind of massive data.

Web Usage Mining (WUM) is one of the classification of web mining. It is a technique for exploring the usage patterns from web log data or web data. These patterns reveal identity, objective origin and the behavior of users at a particular web site. For example, search engine retrieves the results for the queries asked by the user. The efficient search engines will do the search fast, precise and even they suggest search queries while user typing a query. This is done by knowing the user's behavior/searching patterns. All this information is collected form the server log files, data is cleaned and only the knowledge regarding the user needs is observed from that data. And that knowledge is used to suggest user. WUM has preprocessing, pattern discovery and analysis phases (Raju and Satyanarayana, 2008) as shown in Fig 1. In the first phase, data cleaning, user identification, segregating sessions etc., will be done. In second phase, navigation patterns are discovered and clustered into groups. In final phase the patterns will be analyzed to find user similarities.

In this study, we focused on selecting features, clustering by fuzzy C-means and classification using Support Vector Machines (SVM). The proposed system is shown in Fig. 2.

Literature review: Presently, WUM is utilized broadly for finding user navigation patterns from web server log data. WUM helps todecide on system errors and broken links by analyzing website errors (Suneetha and Krishnamoorti, 2009). A new method to classify user navigation patterns and foresee the future user requests is proposed by Araya *et al.* (2004). Grouping of navigation patterns using clustering through evaluating similarities and dissimilarities in a data set and wrap similar data into a group. Anyhow, there lies some vital contrasts between clustering in general applications and in web mining. Since, fuzzy clustering allows creating overlapping clusters and presents memberships for data objects in every cluster and resolves data ambiguity, it is noted as suitable approach. This model is presented in (Shi, 2009).

In the last three decades, a wide variety of classification techniques are available. Among them, support vector machines are the best machine learning algorithm for several complicated binary classification problems (Cristianini and Shawe, 2000).

Web usage mining datasource: There are lot of data sources some of them are:

Web server logs: It has information about the request history of a user i.e., IP address, requested page, date/time bytes served etc. These log files are not generally accessible to users, only for the administrators (Pani *et al.*, 2011).

Corresponding Author: B. Prasanna Kumar Reddy, Department of Computer Science Engineering, KL University, Guntur, India

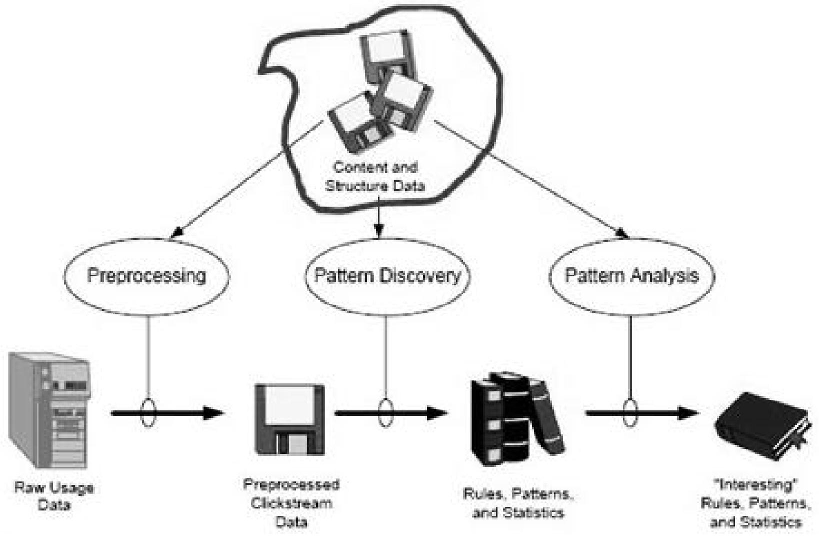

Content and Structure Data
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Usage Data
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Fig. 1: Web usage mining phases

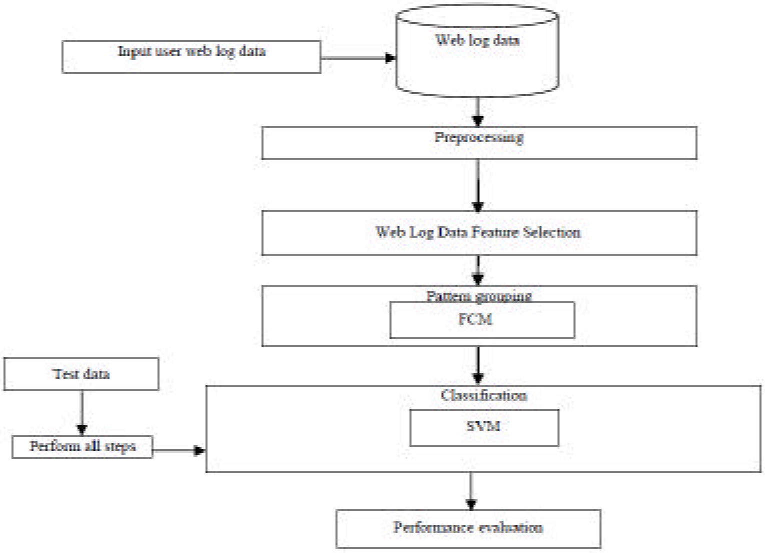


Fig. 2: Proposed system achitecture

Proxy server logs: It reduces webpages loading time and network traffic with the help of caching mechanism. It contains all HTTP requests between clients and servers. It can be utilized as a tool for discovering usage patterns.

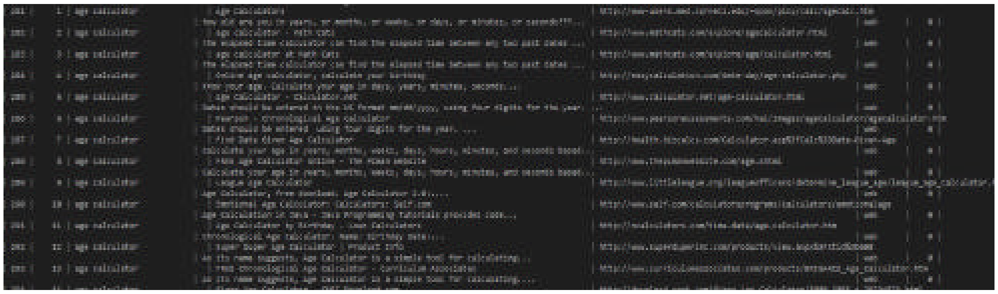
Browser logs: It collects data from client-side using applets and java script. This functionality is optional to use he may either enable or disable in browser settings (Pani *et al*., 2008).

Web usage mining process

Data collection: In this stage, usage information from different sources are gathered from web servers, clients related to a server or from middle sources corresponding to packet sniffers and proxy servers.

Data preprocessing: It represents whichever kind of processing carried out on raw data to make it ready for additional processing technique. Usually utilized as an experimental data mining application, data preprocessing alters the data in such a way that it can be easily processed for the user benefit. The approaches that can be used here can specify client data elaboration. The different data preprocessing tasks are:

Data cleaning: Cleaning raw web data is first step of data preprocessing. Throughout this step the feasible data



```
[illegible] |   1 | age calculator   |  Age Calculator                                                        | [illegible]
            |     |                  | How old are you in years, or months, or weeks, or days, or minutes, or seconds!?...
[illegible] |   2 | age calculator   |  Age Calculator - Math Cats                                            | http://www.mathcats.com/explore/agecalculator.html
            |     |                  | The elapsed time calculator can find the elapsed time between any two past dates ...
[illegible] |   3 | age calculator   |  Age Calculator at Math Cats                                           | [illegible]
            |     |                  | The elapsed time calculator can find the elapsed time between any two past dates ...
[illegible] |   4 | age calculator   |  Online age calculator, calculate your birthday                        | [illegible]
            |     |                  | Know your age. Calculate your age in days, years, minutes, seconds...
[illegible] |   5 | age calculator   |  Age Calculator - Calculator.net                                       | http://www.calculator.net/age-calculator.html
            |     |                  | Dates should be entered in the US format mm/dd/yyyy, using four digits for the year. ...
[illegible] |   6 | age calculator   |  [illegible] - Chronological Age Calculator                           | [illegible]
            |     |                  | Dates should be entered using four digits for the year. ...
[illegible] |   7 | age calculator   |  Find Date Given Age Calculator                                        | [illegible]
            |     |                  | Calculate your age in years, months, weeks, days, hours, minutes, and seconds based...
[illegible] |   8 | age calculator   |  Free Age Calculator Online - [illegible]                              | [illegible]
            |     |                  | Calculate your age in years, months, weeks, days, hours, minutes, and seconds based...
[illegible] |   9 | age calculator   |  League Age Calculator                                                 | http://www.littleleague.org/[illegible]
            |     |                  | Age Calculator, free download. Age Calculator 2.0.....
[illegible] |  10 | age calculator   |  Emotional Age Calculator: Calculators: Self.com                       | http://www.self.com/[illegible]
            |     |                  | Age Calculation in Java - Java Programming tutorials provides code...
[illegible] |  11 | age calculator   |  Age Calculator by Birthday - [illegible] Calculators                  | [illegible]
            |     |                  | Chronological Age Calculator: Name: Birthday Date:...
[illegible] |  12 | age calculator   |  Super Duper Age Calculator | Product Info                             | http://www.superduperinc.com/products/[illegible]
            |     |                  | As its name suggests, Age Calculator is a simple tool for calculating...
[illegible] |  13 | age calculator   |  Free Chronological Age Calculator - Curriculum Associates             | [illegible]
            |     |                  | As its name suggests, Age Calculator is a simple tool for calculating...
```

Fig. 3: An example of raw log file

is analyzed and immaterial or redundant items are deleted from the data set. Immaterial records are removed in data cleaning. The goal of WUM is to acquire traversal pattern.

User identification: A personalized website's success depends on identifying users who access it. The easiest method is to allocate discrete user to each discrete IP recognized within the log file. Cookies can also be used to identify user of a website by keeping an ID which is produced by the web server for every user who visits website (Babu *et al*., 2011).

Session identification: Identifying user sessions acquired a remarkable recognition in WUM as they encode use's navigational behavior which is a key aspect for discovering patterns. The user sessions are the delimited set of pages explored by the same user within the period of one specific visit to a website (Babu *et al*., 2011).

Pattern discovery: Here, knowledge is explored by classifying the user with their navigational behavior. The main objective of classification is to recognize the distinctive features of predefined classes. This needs extracting and selecting features that perfectly represents the properties of given class.

Pattern analysis: It is the final step of WUM process. After the pattern discovery, the acquired patterns are studied to clean information and obtain the useful knowledge. This step allows us to detect patterns automatically in data and predict new data from the same source (Ivancsy and Vajk, 2006) (Fig. 3).

MATERIALS AND METHODS

Web log data is gathered from web server and clients are identified by cleaning the data. The sessions of every client are recognized by navigation oriented technique. Sessions are recreated as a matrix. For the most part the vicinity of excess and immaterial properties could deceive the examination. The capacity and processing of information increases the difficulty of the evaluation and debases the exactness of the outcome. Thus feature selection is a critical stride in pre-processing which improves further evaluation. Clustering is important for grouping similar users. To determine a new classification of user is a promising area in analyzing web log data.

Feature selection: The vicinity of repetitive and unimportant properties could deceive the analysis. Thus features are decreased by Independent Component Analysis (ICA) and it is the more intense procedure to discover the patterns. New patterns are chosen to decrease the matrix dimensions to improve efficiency. In the proposed strategy the session matrix is diminished. In the proposed technique a session matrix which comprises of navigation patterns is studied. The class "c" in this system is the distinctive web pages of the site perused over a time frame. Another matrix is made subsequent to normalizing every pattern with standard deviation and mean. ICA is usedon new patterns and deletes least weighted values (Catlegde and Pitkow, 1995).

The motivation behind ICA is to straightly change the original matrix into parts which are about analytically independent (Hyvarinen and Oja, 2000). The job of ICA is to discover Independent matrix W to make y = Wx where y = (y1, y2,… .yN,… .) T is called yield variable and x = (x1,x2,x3,… xN.) T is called watched arbitrary variable. In the event that yi is autonomous then yi is the assessed value of an independent arbitrary variable s = (s1,s2,s3…sn).

Step 1: Take user matrix as input . Columns; web pages.

Rows: Users along with sessions. Every row is considered as navigation pattern.

Step 2: Normalize each user navigation pattern using $(np_i\text{-}m_i)/2\partial_i$ where mi and ∂ are the mean and standard deviation of np_i, respectively.

Step 3: Calculate mean.

$$x_i = \frac{1}{V+1}\sum_{j=1}^{V+1} | u_{ij} |$$

Step 4: For all $|w_{ij}|$ in w if $|w_{ij}| < x | x_i,$ then shrink $|w_{ij}|$ to zero.

Step 5: Multiply new weight matrix W' and original user navigation pattern.

Step 6: Delete columns with values zero.

Fuzzy C-means clustering: After feature selection process, the user navigation patterns are to be grouped basing on similarities. Clustering is used for this purpose. In this study for grouping user navigation patterns Personalized Posterior Probability based Fuzzy C-Means (PPPFCM) clustering algorithm is used. In fuzzy c-means, the patterns are the centroid of the cluster. The set of co-efficients of user patterns gives the degree of being in kthclusterw$_k$(x):

$$C_k = \frac{\sum_a w_k(a)^m x}{\sum_a w_k(a)^m}$$

Set the cluster centroids vi, fuzzification parameter q, the cluster index l and number of clusters c.

Step 1: Calculate membership values using Equation:

$$\mu_{ik} = \frac{1}{\sum_1^c \left(\frac{d^2(x_k, v_i) + \Upsilon \sum_{J=1}^{N} (1-\mu_{ij})^q Wkj}{d^2(x_k, vl) + \Upsilon \sum_{J=1}^{N} (1-\mu_{ij})^q Wkj} \right)} \frac{1}{(q-1)}$$

Step 2: Evaluate the highest membership value result μ_{ik}based on the probability function .The posterior probability is the probability of the parameters μ^*_{ik}given the evidence ($\mu^*_{ik}|x_k$). P($\mu^*_{ik}|x_k$)is the probability of the evidence given by the parameters and it is different from thelikelihood function in conventional FCM. The posterior probability is defined as:

$$P(\mu^*_{ik} | x_k) = \frac{P(x_k | \mu^*_{ik}) P(\mu^*_{ik})}{p(x_k)}$$

Step 3: Compute the cluster centroids for user navigation patternsnpi:

$$v_i^* = \frac{\sum_{K=1}^{N} \left(P(\mu^*_{ik} | x_k)\right)^q x_k}{\sum_{K=1}^{N} \left(P(\mu^*_{ik} | x_k)\right)}$$

until convergence criteria of the similar user navigation patterns are clustered. After clustering a defuzzification procedure happens, so as to change over the patterns matrix to a fresh segment. Among numerous systems accessible for defuzzification the most extreme participation technique is the best strategy for navigation patterns. The procedureassigns the user navigation patterns k to the class C with the highest membership.

$$c_k = \arg_i \left(\max \left(p(\mu^*_{ik} | x_k) \right) \right), i = 1, \ldots c$$

Classification using support vector machine: After patterns are clustered then next step is to classify patterns. In this work SVM is used for classification. The hyperplane that separates two groups of data isexpressed in the objective function:

$$wx+b = 0$$

Where:

x = The set of training vectors

w = The vectors perpendicular to the separating hyper plane and b represents the offset parameters which allows increase of the margin

For n-dimensional data n-1 hyperplanes are to be introduced. A positive slack variable is introduced in the objective function in order to add some flexibility in separating the categories.The improvised objective function is:

$$yi(wxi+b)> = 1-i$$

Where, > = 0. There are several kernel functions like Gaussian, laplace, radial basis function are available in which laplacian is observed as suitable for the log data classification:

$$k(x,y) = \exp(-||x-y||/\sigma)$$

Given a training set of instance-label pairs (xi; yi); I = 1… l where x ∈ i Rn and y ∈ {1,-1}1, then the support vector machines requires the solution for the objective function:

$$yi(w\varphi\ (xi)+b)> = 1-i$$

The φ maps training vectors xi into a higher dimensional space. Linear separating hyper plane with maximal margin in higher dimensional space is found by SVM.

RESULTS AND DISCUSSION

Experimental evaluation: The proposed web log preprocessing is assessed in this segment. The consequences of diverse stages are given as takes after. The system is executed by utilizing JAVA 1.8 with net beans 8.0 (Fig. 4). There are 358 novel clients distinguished in the wake of applying the calculation. Session recognizable proof procedure is did. Sessions are

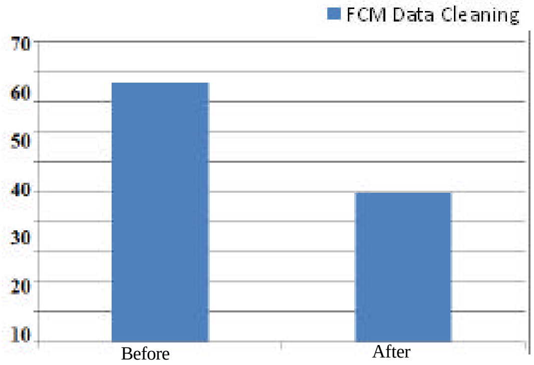


Fig. 4: FCM data cleaning procedure in data extraction

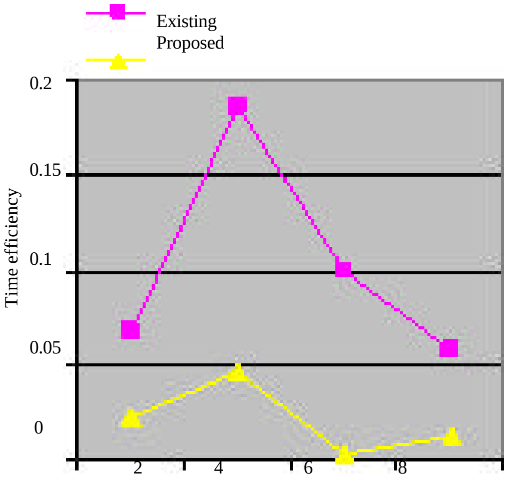


Fig. 5: Varying query results in both existing and proposed approaches

remade in a network configuration which gives insights about the recurrence and route example of distinctive clients. The rand file is a measure used to think about instigated clustering structure (C) and a Classification structure (C). Let 'a' be the quantity of examples that are doled out to the cluster in C and C. 'b' be the quantity of occasions that are in the same cluster C, however not in the group in C."c" be the number of cases that are in the same cluster in C , however not in the group in C and "d" be the quantity of cases that are doled out to distinctive groups in C and C. The amounts a and d can be deciphered as understandings and b and c as disagreements.

Experimental setup: We study the behavior and performance of our algorithms on partitioning a user's query history into one or more groups of related queries. For example, for the sequence of queries "Caribbean cruise"; "Bank of America"; "expedient"; "financial statement" we would expect two output partitions: first, {"Caribbean cruise", "expedia"} pertaining to travel-related queries and, second, {"bank of America", "financial statement"} pertaining to money-related queries.

Using search logs: Our query grouping algorithm relies heavily on the use of search logs in two ways: first, to construct the query fusion graph used in computing query relevance and second to expand the set of queries considered when computing query relevance. We start our experimental evaluation by investigating how we can make the most out of the search logs. In our first experiment we study how we should combine the query graphs coming from the query reformulations and the clicks within our query log (Fig. 5).

Above graph describes the horizontal axis represents (i.e., how much weight we give to the query edges coming from the query reformulation graph) while the vertical axis shows the performance of our algorithm in terms of the RandIndex metric.

CONCLUSION

To increase the number of user visits to web site the administrators should improve the efficiency of the site by understanding the user needs. WUM is the best technique for extracting the knowledge from web log data to understand the user behavior. It is a technique for exploring the usage patterns from web log data or web data. This data is a type of information that is vital to understand the user. It has both the relevant and irrelevant data. In order to extract the exact information clustering and classification techniques must be used. In this study, fuzzy c-means clustering and support vector machines classification is proposed for extracting knowledge. The entire process is done in three phases in the first phase data is cleaned and only certain features are selected in the second phase user navigation patterns are grouped and in third phase patterns are classified. Our approach here is to present and hypothesize the proposed method.

REFERENCES

Araya, S., M. Silva and R. Weber, 2004. A methodology for web usage mining and its application to target group identification. Fuzzy Sets Syst., 148: 139-152.

Babu, D.S., S.A. Nabi, M.A. Ali and Y. Raju, 2011. Web usage mining: A research concept of web mining. Int. J. Comput. Sci. Inf. Technol., 2: 2390-2393.

Catlegde, L. and J. Pitkow, 1995. Characterising browsing behaviours in the world wide web. Comput. Netw. ISDN. Syst., 27: 1065-1073.

Cristianini, N. and T.J. Shawe, 2000. An Introduction to Support Vector Machines and other Kernel-Based Learning Methods. Cambridge University Press, Cambridge, UK., ISBN: 0-521-78019-5, Pages: 187.

Eltahir, M.A. and A.A.F. Dafa, 2013. Extracting knowledge from web server logs using web usage mining. Proceedings of the 2013 International Conference on Computing, Electrical and Electronics Engineering (ICCEEE), August 26-28, 2013, IEEE, Khartoum, Sudan, ISBN: 978-1-4673-6231-3, pp: 413-417.

Hyvarinen, A. and E. Oja, 2000. Independent component analysis: Algorithms and applications. Neural Networks, 13: 411-430.

Ivancsy, R. and I. Vajk, 2006. Frequent pattern mining in web log data. Acta Polytech. Hungarica, 3: 77-90.

Pani, S.K., L. Panigrahy, V.H. Sankar, B.K. Ratha and A.K. Mandal et al., 2011. Web usage mining: A survey on pattern extraction from web logs. Int. J. Instrumentation Control Autom., 1: 15-23.

Raju, G.T. and P.S. Satyanarayana, 2008. Knowledge discovery from web usage data: Complete prepropcessing methodology. Int. J. Comput. Sci. Netw. Secur., 8: 179-186.

Shi, P., 2009. An efficient approach for clustering web access patterns from web logs. Int. J. Adv. Sci. Technol., 5: 1-14.

Suneetha, K.R. and R. Krishnamoorti, 2009. Identifying user behavior by analyzing web server access log file. Int. J. Comput. Sci. Network Security, 9: 327-332.