

# CLASSIFICATION OF WEB LOG DATA TO IDENTIFY INTERESTED USERS USING DECISION TREES

**K. R. Suneetha, R. Krishnamoorthi**

Bharathidasan Institute of Technology, Anna University  
krs\_mangalore@hotmail.com  
rkrish\_26@hotmail.com

## ABSTRACT

Focus of many industries change towards customer orientation to retain regular frequent accessors for the improvement of customer relationship management. Study of interested web users, provides valuable information for web designer to quickly respond to their individual needs. The aim of this paper is, instead of tracking the behavior of overall users to redesign the web site, our model extracts only focused group of interested users. The proposed model consists of two phases. In first phase, the web server log data is preprocessed. The purpose of data preprocessing is to extract useful data from raw web log. In the second phase data is classified using enhanced version of Decision tree algorithm C4.5. NASA web server data is used for experimental purpose, which results in less execution time and reduced memory utilization with high accuracy.

**Keywords:** Web Usage Mining, Preprocessing, Decision Tree, Frequent accessors.

## 1 INTRODUCTION

The World Wide Web (WWW) is rapidly emerging as an important communication means of information related to a wide range of topics (e.g., education, business, government). It has created an environment of abundant consumer choices, where organizations must give importance to improve customer loyalty. To reorganize a website in terms of structuring links and attractive design of web pages, organizations must understand their customers behavior, preferences, and future needs. This imperative leads many vendors to develop more e-service systems for data collection and analysis. The navigation patterns of users generally gathered by the web servers and stored in server access logs. Analysis of server access log data provide information to restructure a web site to increase effectiveness, better management of work group communication, and to target ads to specific users. Web mining [1] is a popular technique for analyzing

users activities in e-service systems. It includes i) Web Content Mining- extracting knowledge from the content of documents, ii) Web Structure Mining- inferring knowledge from Internet links iii) Web Usage Mining-extracting interesting patterns from web access logs. Our work focuses on Web Usage Mining.

Web Usage Mining (WUM) [2],[3] discovers usage patterns from web data, to better serve the needs of web-based applications. Three different tasks of usage mining are Data Preprocessing, Pattern Discovery and Pattern Analysis. Data preprocessing consists of data cleaning, data integration, data transformation and path completion. Pattern discovery techniques used to extract knowledge from preprocessed data. Some of the techniques used in pattern discovery are Association rules, Classification, Clustering, Sequential Patterns, etc. Pattern Analysis filters out uninteresting rules or patterns from the set found in the pattern discovery phase.

This paper focuses on extracting interested user

group from web log data. Due to large amount of irrelevant information in the web log, the original log cannot be directly used in the web log mining procedure. Thus in the first phase the web server log data is preprocessed, to extract useful data and then to map these data in to the abstract data necessary for pattern discovery. In this step, the original size of the database will be reduced. In second phase, data is categorized using decision tree depending on splitting attributes.

The rest of the paper is organized as follows. In section 2 we discuss the related work. The proposed approach and its details are presented in section 3. Section 4 discusses decision tree algorithm C4.5 and generation of decision rules. Section 5 presents the results. Finally, conclusions are drawn in section 6.

## 2 RELATED WORK

In recent years abundant work has been carried out in the area of web mining, specifically on analysis of web log data of Web Usage Mining.

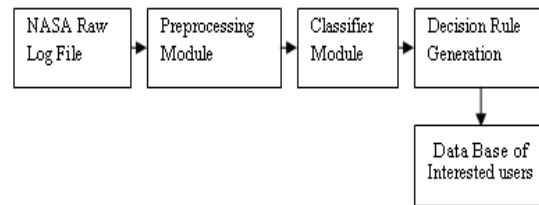
The paper [4] describes improved statistical-based time oriented heuristics for the reconstruction of user sessions from a server log. The authors Bamshad Mobasher et.al., [5] presents transaction data models for various web mining tasks such as the discovery of association rules and sequential patterns from the Web data. In paper [6] Cooley et.al presented methods for user identification, sessionizing (i.e. constructing or reconstructing sessions), and page view identification. In another work [7], the authors highlights the importance of analyzing users web log data and extracting their interests of web-watching behaviors and describes a method for clarifying users interests based on an analysis of the site-keyword graph. The papers [8],[9] discusses importance of decision trees in web log mining, the authors suggested several hypotheses to improve web sites retention and showed that it is possible to predict future user actions with reasonable misclassification error as well as to find combinations of sequential pages resulting in browsing termination. Swamy Kutti et.al., [10] discuss the importance of navigation patterns of users. It provides information about the major aspects and problems related to the task of modeling the user behavior. It also gives information about recent developments related to automatic web navigation, implicit capturing of user interests and future directions on web log analysis.

A number of research work concentrated on applying data mining techniques on to preprocessed web log data to identify frequent patterns. But our model tries to form well focused data of interested users and then frequent pattern mining algorithm is

applied on this group instead of considering overall entries, which intern improves the performance.

## 3 PROPOSED MODEL

Finding information relevant to user needs has become increasingly important now a days, and how efficiently and effectively retrieve the `required web pages are becoming more challenge. This issue is addressed in our approach and is focused on discovering frequent accessors behavior which helps in efficient organization of the web site. The Fig. 1 shows the building blocks of the entire process.



**Figure 1: User Grouping Model**

The raw data should be cleaned to eliminate irrelevant information from the original log file and to form the Web log database convenient for user identification, session identification. Then the Classification model classifies preprocessed data by considering number of factors like time stamp, request type, number of pages referred and depth wise analysis as the attributes and predicates.

### 3.1 Preprocessing Module

The raw data should be cleaned to eliminate irrelevant information from the original log file and to form the Web log database convenient for user identification, session identification This module removes useless requests from the log files. Since all the log entries are not valid, we need to eliminate the irrelevant entries. Usually, this process removes requests concerning non-analyzed resources such as images, multimedia files, and page style files. Data cleaning includes the elimination of irrelevant entries like:

- Entries with unsuccessful HTTP status codes;

HTTP status codes are used to indicate the success or failure of a requested event, and we only consider successful entries with codes between 200 and 299.

- Entries with request methods except GET and POST.

Once the data is preprocessed, user identification is done based on the attributes of IP addresses, time spent on that particular session and also the agent or operating system which is used as one of the parameter

to identify unique users. These users are classified as interested and not interested users using Decision tree algorithm.

### 3.2 Decision Trees

Decision trees are used in classification and prediction. It is simple yet a powerful way of knowledge representation. The models produced by decision trees are represented in the form of tree structure. A leaf node indicates the class of the examples. The instances are classified by sorting them down the tree from the root node to leaf node.

In our work we have used enhanced version of decision tree C4.5 algorithm. The reason for selecting C4.5 decision tree algorithm is ability to handle data with missing values. It also avoids over fitting the data and reduce error pruning. The C4.5 algorithm is described in the following section.

## 4 ENHANCED C 4.5 ALGORITHM

Input: training samples, represented by discrete attributes; the set of candidate Attributes, attribute-list.

Output: set of classes

Method:

1. Create a node N;
2. If samples are all of the same class C, then Return N as a leaf node labeled with the class C;
3. If attribute list is empty then Return N as a leaf node labeled with the most common class in samples (majority voting)
4. Select test attribute, the attribute among attribute-list with the highest information gain ratio;
5. Label node N with test-attribute;
6. For each known value  $a_i$  of test-attribute
7. Grow a branch from node N for the condition test-attribute =  $a_i$ ;
8. Let  $s_i$  be the set of samples in samples for which test-attribute =  $a_i$ ;
9. If  $s_i$  is empty then
10. Attach a leaf labeled with the most common class in samples;
11. Else attach the node returned by generate-decision-tree

The information gain measure C4.5 algorithm is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure of the goodness of split. The attribute with the highest gain is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found. The information gain measure and gain ratio criterion is provided further.

### 4.1 Information Gain

Let S be a set of training set samples with their corresponding labels. If there are m classes and the training set contains  $s_i$  samples of class I and s is the total number of samples in the training set. Expected information needed to classify a given samples is calculated by:

$$I(S_1, S_2, \dots, S_m) = \sum_{i=1}^m (S_i/S) \log_2 (S_i/S) \quad (1)$$

A feature F with values  $\{f_1, f_2, \dots, f_v\}$  can divide the training set into v subsets  $\{s_1, s_2, \dots, s_v\}$  where  $s_j$  is the subset which has the value  $f_j$  for feature F.

Furthermore let  $S_j$  contain  $S_{ij}$  samples of class i. Entropy of the feature F is

$$E(F) = \sum_{j=1}^v (S_{1j} + \dots + S_{mj}) / S \times I(S_{1j}, \dots, S_{mj}) \quad (2)$$

Information gain for F can be calculated as:

$$\text{Gain}(F) = I(S_1, S_2, \dots, S_m) - E(F) \quad (3)$$

In our experiment, information gain is calculated for class labels by employing a binary discrimination for each class. That is, for each class, a dataset instance is considered in-class, if it has the same label: out-class, if it has a different label. Consequently, as opposed to calculating one information gain as a general measure on the relevance of the feature for all classes, we calculate an information gain for each class. Thus this signifies how well the feature can discriminate the given class from other classes.

### 4.2 Gain ratio Criterion

The notion of information gain introduced earlier tends to favor attributes that have a large number of values. For example if we have an attribute D that has a distinct value for each record, then Info (D, T) is 0,

thus Gain (D, T) is maximal. To compensate for this, it was suggested to use the following ratio instead of gain.

SplitInfo is the information due to the split of T on the basis of the value of the categorical attributes D, which is defined by

$$\text{SplitInfo}(X) = -\sum_{i=1}^n \frac{|T_i|}{|T|} \times \log_2 \frac{|T_i|}{|T|} \quad (4)$$

and Gain Ratio is then calculated by

$$\text{GainRatio}(D, T) = \frac{\text{Gain}(D, T)}{\text{SplitInfo}(D, T)} \quad (5)$$

The gain ratio, expresses the proportion of useful information generation split, i.e., that appears helpful for classification. If the split is near trivial, split information will be small and this ratio will be unusable. To avoid this, the gain ratio criterion selects a test to maximize the ratio above, subject to the constraint that the information gain must be large, at least as great as the average gain over all tests examined.

Using the decision tree algorithm of C4.5 and measures used to select test attribute, the generation of decision rules generation is discussed in the next section

### 4.3 Decision Tree Generation

The set of attributes (A1...A5) used to recognize users interest, their behavior (Table 1).The basic idea of our approach is to identify a small training data set in the beginning itself, so that we can achieve high accuracy for the most important class and user with purchase interest.

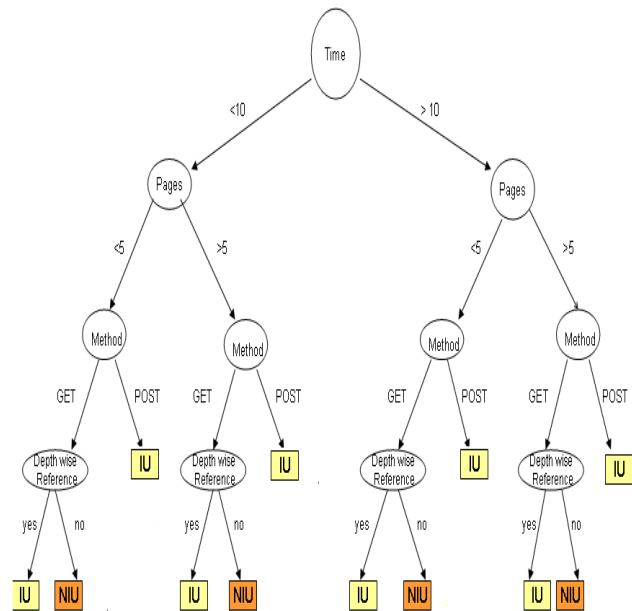
**Table 1:** Attributes Used For Classification.

Attributes	Description
A1	total session time: 15-30 min.
A2	total time a user stays at the site: > 30secs
A3	total number of accessed pages during the whole session: >5pages
A4	access methods used to interact with the site: GET, POST
A5	Depth Wise Access from particular page(DWA)

We subjectively identify users who have some

purchase interest and those who simply visit the site. Generally interested users exhibit certain access patterns; they access certain web pages for a rather long time because they need time to spend on its contents. The ratio between the time spent on content reading and the amount of time they navigate is large. The interested users often use the HTTP POST mode, because they are interested in registering with web sites and are willing to fill out forms with their own information. The user who does not have interest simply accesses many pages quickly to browse contents. These users do not often use POST method because they are not interested in registering at web sites. The web log files are classified as Interested users (IU) and Not Interested Users(NIU) based on the parameters Time Stamp (30 min), method used(GET/ POST ), number of pages referred(min 5pages.) and depth wise reference by the user.

Fig.2 shows the construction of decision tree by considering the above mentioned attributes and decision rules are generated according to the conditions.



**Figure 2:** Decision Tree Generation

IU: Interested Users  
NIU: Not Interested Users

Decision Rules:

Rule1: If (Time spent <30 and No. of pages referred <5 and Method used GET and DepthwiseReference='NO') = "NIU"

Rule2: If (Time spent <30 and No. of pages referred <5 and Method used GET and DepthwiseReference='YES') = "IU"

Rule3: If (Time spent <30 and No. of pages referred <5 and Method used POST) = "IU"

Rule4: If (Time spent <30 and No. of pages referred >5 and Method used GET and DepthwiseReference='NO') = "NIU"

Rule5: If (Time spent <30 and No. of pages referred >5 and Method used GET and DepthwiseReference='YES') = "IU"

Rule6: If (Time spent <30 and No. of pages referred >5 and Method used POST) = "IU"

Rule7: If (Time spent >30 and No. of pages referred <5 and Method used GET and DepthwiseReference='NO') = "NIU"

Rule8: If (Time spent >30 and No. of pages referred <5 and Method used GET and DepthwiseReference='YES') = "IU"

Rule9: If (Time spent >30 and No. of pages referred <5 and Method used POST) = "IU"

Rule10: If (Time spent >30 and No. of pages referred >5 and Method used GET and DepthwiseReference='NO') = "NIU"

Rule11: If (Time spent >30 and No. of pages referred >5 and Method used GET and DepthwiseReference='YES') = "IU"

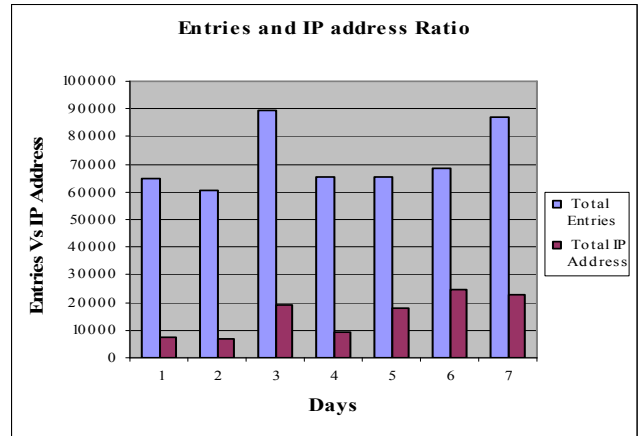
Rule12: If (Time spent >30 and No. of pages referred >5 and Method used POST) = "IU"

We have grouped all these classes as "Interested Users" and "Not Interested Users". This grouping helps to choose only focused set of data for identification of user behavior, user interest etc. Thus instead of analyzing whole web log file content, which in terms results in consumption of more memory and unnecessary processing of unwanted data, our method gives preference to the group of interested users for further behavior analysis which reflects in faster execution and less space utilization. The details of experimentations are presented in the next section.

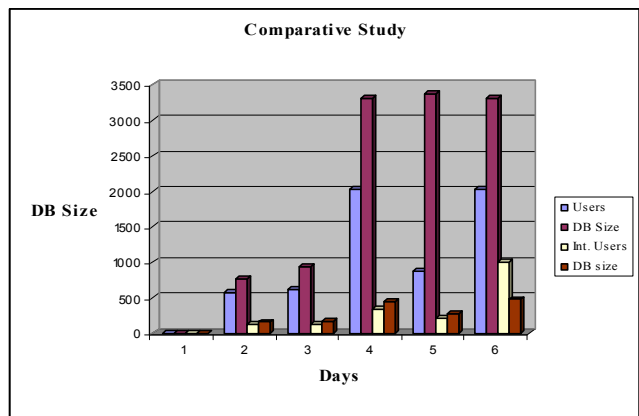
## 5 EXPERIMENTAL RESULTS

NASA server web log file of 695MB data is considered in our work. For the purpose of testing 10 days of data is collected (68MB). After preprocessing, the data base size is reduced to 18 MB that is 60 to 70% by eliminating unnecessary data sets (see Fig. 3). From the preprocessed data 1,34,937 records were considered and from which, we obtained 10,119 unique users of 12,857 sessions as a training data set. C4.5 algorithm is applied on its data set and 2618 entries are classified as entries of interested users. The graph shown in Fig. 4 gives comparative study of unique

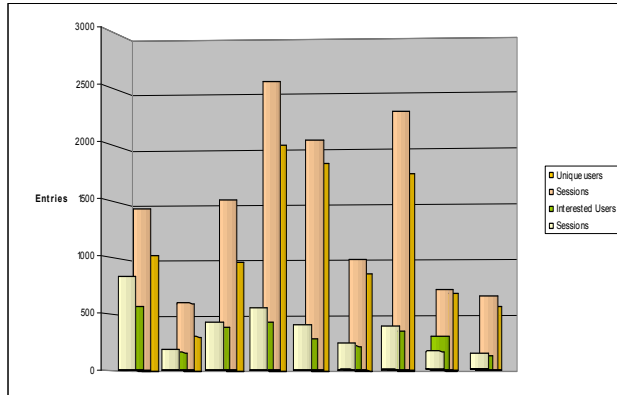
users and interested users with Data Base Size. Fig. 5 indicates variation between the entries of unique and interested users with sessions. Percentage of reduction in Data Base size of interested users with unique users is shown in Fig. 6. This shows a small training dataset (interested user group) achieve high accuracy for the most important class of users with purchase interest.



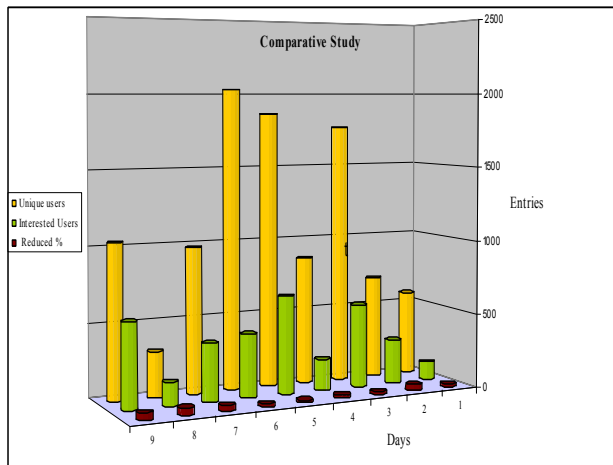
**Figure 3:** Total number of unique users after preprocessing



**Figure 4:** Comparative Study of users with interested users with Data Base Size



**Figure 5:** Observation of variation between sessions of unique and interested users



**Figure 6:** Percentage of reduction in data base of interested with unique users.

Our approach is well suited for application like e-Commerce and e-Business in identifying actionable information to acquire new customers and to retain old customers.

## 6 CONCLUSIONS

The work focused on group of the frequently accessed patterns of interested users. It assists the web site designers to improve the performance of the web by giving preference to the patterns navigated by the regular interested users. Step1 removes unnecessary and junk requests from the log file. The enhanced version of decision tree C4.5 algorithm is applied for identification of interested users from web server log file. Instead of tracking the behavior of overall users (interested or not interested) to redesign the web site, our approach generates the well focused set of interested users. The results showed the improvement

in time and memory utilization. More effective association rules can be generated from the well focused group formed by our work to find frequently accessed patterns, which results in less time consumption and less memory utilization with high accuracy.

## 7 REFERENCES

- [1] H. Blockeel and R. Kosala: Web Mining Research-a survey, SIGKDD Explorations., Volume 2, Issue 1, pages 1-15 (2000).
- [2] Hui-You Chang and Feng Zhang: Research and Development in Web Usage Mining System – key issues and proposed solutions, Proceedings of the First International Conference on Machine Learning and Cybernetics., pp.986-990 ( 2002).
- [3] Jaideep Srivastava, Prasanna Desikan, and Vipin Kumar: Web Mining – Accomplishments and Future Directions, Technical Report Computer Science Department, University of Minnesota, Minneapolis, USA. pp. 51-6 (2002).
- [4] Jie Zhang Ali and A. Ghorbani: The Reconstruction of User Sessions from a Server Log Using Improved Time-oriented Heuristics, Proceedings of the Second Annual Conference on Communication Networks and Services Research, IEEE (2004).
- [5] Namit Jain, Bamshad Mobasher, Eui\_Hong\_Sam\_Han and Jaideep Srivastava: Web Mining Pattern Discovery from World Wide Web Transactions, Technical Report, Department of Computer Science, University of Minnesota (1996).
- [6] Robert Cooley, B. Mobasher, and Srivastava: Data Preparation for Mining World Wide Web Browsing Patterns, Knowledge and Information System., Volume 1, No 1, pp. 5-32 (1997).
- [7] Tsuyoshi Murata and Kota Saito: Extracting Users Interests from Web Log Data, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, China. pp.343-346 (2006).
- [8] V Singh, Alsabti K and Ranka S: CLOUDS: A Decision Tree Classifier for Large Datasets, In Proceedings of KDD 98, 4th International Conference on Knowledge discovery and Data Mining, New York City, pp. 2-8 (1998).
- [9] Zidrina, Pabarskait: Decision trees for web log mining, Intelligent Data Analysis., Volume 7, Issue2, pp. 141-154 (2003).
- [10] Zuhoor Al-Khajri, Swamy Kutti, Muna Hatem and Haider AL Khajri: A Classification Technique for Web Usage Analysis, Journal of Computer Science, ISSN Publication, pp. 413-418, (2005).