Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification

A. K. Santra[1], S. Jayasudha[2]

[1]**Dean, CARE School of Computer Applications , Trichy – 620 009, India**.

[2]**Research Scholar, Bharathiar University, Coimbatore – 638401, India.**

Abstract

Web Usage Mining (WUM) is the process of extracting knowledge from Web user's access data by exploiting Data Mining technologies. It can be used for different purposes such as personalization, system improvement and site modification. Study of interested web users, provides valuable information for web designer to quickly respond to their individual needs.

The main objective of this paper is to study the behavior of the interested users instead of spending time in overall behavior. The existing model used enhanced version of decision tree algorithm C4.5. In this paper, we propose to use the Naive Bayesian Classification algorithm for classifying the interested users and also we present a comparison study of using enhanced version of decision tree algorithm C4.5 and Naive Bayesian Classification algorithm for identifying interested users. The performance of this algorithm is measured for web log data with session based timing, page visits, repeated user profiling, and page depth to the site length. Experimental results conducted shows that the performance metric i.e., time taken and memory to classify the web log files are more efficient when compared to existing C4.5 algorithm.

Keywords: Web Usage Mining, Web Mining, Web Log Files, Classification

1. Introduction

The ways in which users interact with a World Wide Web (Web) site provide enormous data processing on the usefulness and effectiveness of Web design elements and content built in it. Yet the informative log files recorded by Web servers, and client logs, offer potentially useful data about users Web site interactions. These data may be segregated and studied to generate inferences about Web site design, to test prototypes of Web sites or their modifications over time, and to test theoretical hypotheses about the effects of different design variables on Web user behavior.

Web usage mining involves with the application of data mining methods to discover user access patterns from web data. The main task of web usage data is to capture web-browsing behavior of users from a specified web site. Web usage mining can be classified according to kinds of usage data examined. In our context, the usage data is web log data, which maintains the information regarding the user navigation. Our work concentrates on web usage mining.

This paper proposes classification of web log data and studying the interested users from them. Due to the involvement of uninterested users in the web log, the original log cannot be used as a process in the web usage mining procedure. Thus in the first phase the web log data is preprocessed, to extract the interested data and then to proceed with the extracted data. During this phase, the actual size of the database will be minimized to certain extent. The second phase consists of segregating the data using Naive Bayesian Classification. The remainder of the paper is organized as follows. In section 2, we discuss the related work. In section 3, Naive Bayesian Classification approach is discussed in detail. Algorithm and mathematical evaluation is explained in section 4. Results on the experiments conducted are discussed in section 5. Finally conclusions is discussed in section 6.

2. Related Work

Classification of web log data using naïve Bayesian method is one of the well-known approaches that improve the overall performance of the web server. In this section, we provide taxonomy regarding web mining, classification rule mining methodology based on decision trees, the algorithm C4.5 that have been used in the existing work to identify the interested users.

Jie Zhang and Ali., A. Ghorbani [1] proposed Web usage mining plays an important role in the personalization of Web services. Users' access to pages of the Website should be separated into user sessions. The required user sessions are extracted from the Web server log. Several approaches have been proposed. In this paper we consider two different approaches in initially defining Web mining. First was a 'process-centric view', that defined Web mining as a sequence of tasks. Second was a 'data-centric view', which defined Web mining in proportion to the types of Web data that was being used in the mining process. Mahesh Thylore Ramakrishna1, Latha Kolal Gowdar, Malatesh Somashekar Havanur, Banur Puttappa Mallikarjuna Swamy [2] these authors follow the data-centric view, and refine the definition of Web mining.

Alka Gangrade□, Durgesh Kumar Mishra, Ravindra Patel [3] focused on the review of the techniques for privacy preserving classification under multi-party environment. Further, the two approaches, the classification model and secure multi-party computation algorithms have also been reviewed. The performance analysis of the algorithms has been concentrated in connection with the classification. Classification Rule Mining algorithms are based on centralized data model that is all data is gathered into a single site. Hidenao Abe [4] described a classification rule mining framework by combining the two models i.e., temporal pattern extraction and rule mining. This framework has been developed for mining if-then rules consisting of temporal patterns in left hand side of the rules. The right hand side helps us to predict both of important events and temporal patterns of important index.

Classification algorithms discussed by Hanady Abdulsalam, David B. Skillicorn [5], consist of three phases; a training phase that consists of labeled records, a test phase using previously unseen labeled records, and a deployment phase that classifies unlabeled records. In traditional decision tree classification, a feature (an attribute) of a tuple is either categorical or numerical. Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau Dan Lee [6], presented the problem of constructing decision tree classifiers on data with uncertain numerical attributes.

Quinlan J R [7], described a decision trees for classification tasks. These trees are constructed beginning with the root of the tree and proceeding down to its leaves.

Rules can also be extracted from decision trees easily. Many algorithms, such as ID3 and C4.5, have been devised for decision tree construction. These algorithms are widely adopted and used in a wide range of applications as discussed in [6]. Veronica S. Moertini [8] discussed an overview of data classification and its techniques, the basic methods of C4.5 algorithm, the process and the result analysis of the experiment in utilizing C4.5 for varied dataset.

The Decision Tree's can deal with one attribute per test node or with more than one. The former approach is called Univariate Decision Tree, and the second is the Multivariate method. Thales Sehn Korting [9] explains the construction of Univariate DT's and the C4.5 algorithm, used to build such trees. After this, we discuss the Multivariate approach, and how to construct such trees. Based on the analytical evaluation, Salvatore Ruggieri [10] has implemented a more efficient version of the algorithm called C4.5 (Enhanced C4.5 algorithm). It improves on C4.5 by adopting the best among the strategies for computing the information gain of continuous attributes. All the strategies adopt a binary search of the threshold in the whole training set starting from the local threshold computed at a node.

Mahdi Khosravi and Mohammad and J. Tarokh [11] proposed a dynamic mining approach to modeling and predicting users' navigation patterns. Naïve Bayesian algorithm was implemented and shown that this method is effective. This motivate us to present a Naïve Bayesian Classification model for the classification of web log data in quicker time with minimal memory utilization to identify the user preferences more accurately.

3. Naive Bayesian Classification Model

The need and requirements of the admin user's of the websites to analyze the user preference become essential, due to massive internet usage. Retrieving the decisive information about the user preferences is achieved, using Naive Bayesian Classification algorithm with quicker time and lesser memory, by means of constructive naïve bayes function. The Naive Bayesian Classification technique as shown in Fig 2, is applied on the web log data to evolve the classification of user page preferences and time spent on the pages of the respective web site (URL).

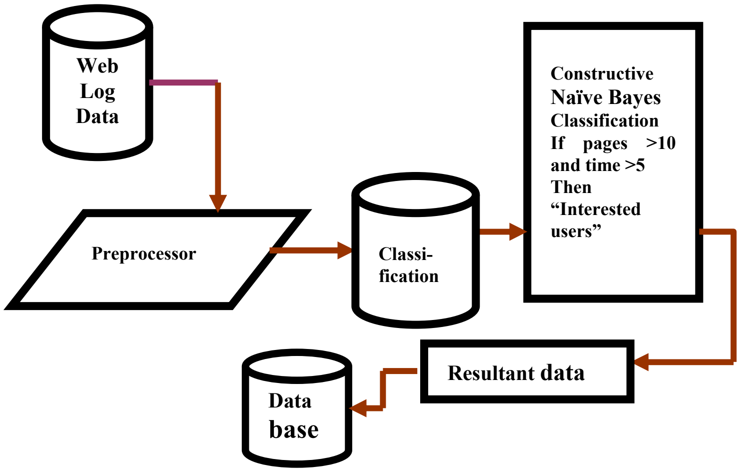

Fig 2 Framework of Naive Bayesian Classification

The web log data (training set) comprises of labels which indicates the class of the observations. New data is classified based on the training set. It preprocesses data in order to remove the irrelevant or redundant attributes and form the normalize data. The decision tree is drawn in a top-down manner. The training sets are at the root. They are partitioned on selected attributes. Partitioning of the training set is processed till there is no more leaf for classifying or there are no samples left and the resultant data is fed to the database.

3.1. Data Cleaning Model

All the data obtained from web log are not informative. There may be redundant data or data that are irrelevant for our work. Hence the web log data are fed to the preprocessor which does the work of cleansing to remove the redundant and irrelevant data from the original log file and produces a new file. Then the classification model performs the classification based on the decision tree. Some of the irrelevant information are the images, advertisements and screen savers etc. Based on these data cleaning model the users are identified as “interested users” or “not interested users” using classification by Naive Bayesian Classification model. As a result, data cleaning model has applied the following criteria:

Failure and aborted requests: The failure or failed requests are determined using the HTTP code. If the value of the code is 500 the request is said to be a success one, if it does not contain the value as 500 it is derived as a failure request. A page exited due to hardware or software failure is considered as aborted requests.

No of pages viewed: Depending on the number of pages viewed by the log file, we can reduce the number of unwanted files. In our work we have considered the minimum pages as 6.

Time taken: Depending on the duration of the time spent by the user, we can eliminate the log files which are not used for future references.

3.2. Classification Model

Given a training data set, the classification model is used to categorize the given training data set into attributes and the attributes are referred to as class. In our web log data time stamp, users, etc. are considered as attributes or class. Classification can be performed using different techniques. Our work concentrates on decision tree. Our goal is to predict the target class based on our source data (web log data). Our model takes into consideration the binary type of classification in which the target attribute has only two possible variations: for example, interested users or not interested users.

In our work we have used Classification by Naive Bayesian model. It results in less time consumption and less memory utilization. The description about the Naive Bayesian Classification model is described in the forthcoming section.

4. Naive Bayesian Classification theorem

Bayesian method is used in decision making that involves probability inferences. This method is more useful when the dimensionality of input is enormous. It uses the prior events to predict the future events. The theorem is explained as given below:

Let Q = $\{x_1,x_2,...,x_n\}$ be a sample training data set whose attributes represent values made on a set of n attributes. Here ‘x’ is considered as "evidence". Let H represent the hypothesis, in such a way that the data belongs to a specified class C. Our work is to determine P(H | Q). It represents the probability that the hypothesis H holds given the "evidence". For example, our training data set have attributes : session id, time taken, number of pages viewed and that session id is 127.0.0.1 , time taken is 6 minutes and number of pages viewed is 7. Then P(H | Q) is the probability that the session id may be an interested user or not interested user given the time taken and number of pages viewed. Again P(H) is called as priori probability of H. For our example we can represent it as that any session id can be considered as interested or not interested regardless of time taken and number of pages viewed.

4.1. Algorithm

Initialization

1. Let T be a training set of samples with k attributes as $A_1,A_2,\ldots..A_k$ given by n dimensional vector $Q = \{x1,x_2,\ldots x_n\}$

2. Let P denotes the probability

3. Let G be the Gaussian distribution value Process

4. Given a sample Q, the classifier performs the prediction to determine the attributes having the highest posteriori probability such that

$$P(A_i \mid Q) > P(A_j \mid Q) \text{ where } i,j = 1,2,\ldots k$$

5. Maximum posteriori hypothesis is calculated using

$$P(A_i \mid Q) = \frac{P(Q \mid A_i)\, P(A_i)}{P(Q)}$$

6. Maximize $P(Q \mid A_i)\, P(A_i)$ if both $P(Q \mid A_i)\, P(A_i)$ are known or $P(Q \mid A_i)$ if only $P(Q \mid A_i)$ is known.

7. If the web log data set contain many attributes it results in maximum of computation time which can be reduced using the following equation

$$P(Q \mid A_i) \equiv \lambda\, P(x_n \mid A_i)$$

8. Calculation of Gaussian distribution with mean μ and standard deviation σ is calculated by

$$G(x, \mu, \sigma) = \frac{1}{\sqrt{3 \prod}} \quad \frac{\exp \quad (x - \mu)^3}{(3\sigma)^3}$$

9. The above equation can be simplified as

$$P(xn \mid A_i) = G(xn, \mu A_i, \sigma A_i)$$

Where μA_i refers to the mean and σA_i refers to the standard deviation value of attribute S_k.

Maximum Likelihood Estimation (MLE) is used for estimating the parameters for a given training data set. If a clear result cannot be achieved due to time or cost constraint, by using the mean and standard deviation the maximum likelihood estimation can be accomplished. The MLE is discussed in the section given below.

4.2. Maximum likelihood Evaluation (MLE)

Let S be the training set with attributes (s1,s2,…,sn) with a vector as Q. To evaluate the maximum likelihood we have to form the density function that is given as :

$$\mathbf{F(s_1,s_2,\ldots s_n \mid Q) = f(s_1 \mid Q) * f(s_2 \mid Q)*\ldots..f(s_n \mid Q)}$$

$$\mathbf{= \Omega\, f(S_i \mid Q) \text{ where } i = 1,2,\ldots n}$$

Where s1,s2,…sn specifies the parameters and Q being the vector is a random variable. Maximum likelihood for S can be evaluated as

$$\mathbf{Max(\mu A_i, \sigma A_i) \mid \Omega\, f(S_i \mid Q)}$$

In our work Maximum likelihood Evaluation is calculated for a given training data set – web log data which produces a distribution function with the observed data having the greatest probability.

4.3. Decision Tree model

Decision tree model is a method most frequently used in data mining. The purpose is to create a model that predicts the resultant of a target variable based on several input variables given by the user as training data set. An example is shown in fig 3. The interior node corresponds to one of the input variables. The input variable consists of the children as edges. Each leaf shows a value of the target variable given with the values of the input variables which are shown by the path from the root to the leaf. The process is repeated in a recursive manner till a resultant value is derived. The parameters used in web log data to classify the user as interested or not interested is given in table 1.

Parameters used	**Explanation**
P_1	No of pages view >10
P_2	Time taken >5
P_3	Hyperlink >5
P_4	Personal Information given by user = “yes”

Table 1 Parameter consideration

Using the parameters given in the table 1 a decision tree is formed as in fig 3 using the naïve Bayesian classifier algorithm which helps to determine whether a user who logs into the system is an “interested user” or “not interested user”. By means of naïve Bayesian algorithm the memory utilized and time taken can be reduced and maximum likelihood of the parameter is also increased.

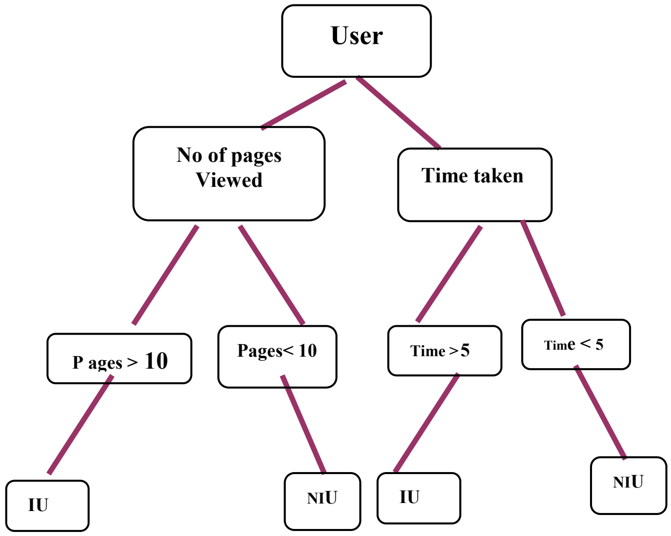

Fig 3 A Decision Tree generation for Interested User and Not Interested User

5. Experimental Results

Web log data was tested on log files stored by the server. We took into account some portions of log files during the same time interval of five hours of five different days; the sample data set is given in the figure 4.

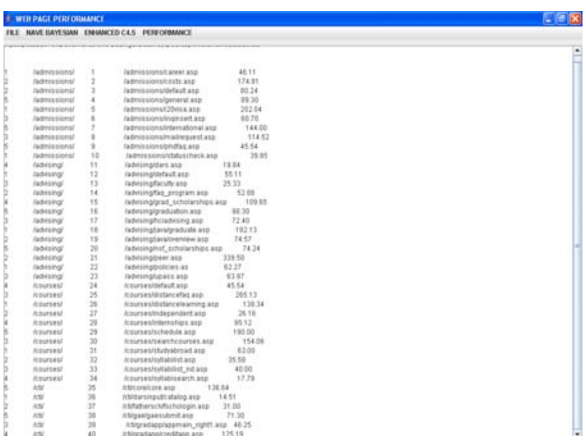

Fig 4. Sample Log Files

The data cleaning model evaluate the web log file to determine the log file that are redundant or irrelevant. As described before, the results of data cleaning is derived in 3.1 which removes all group of irrelevant requests.

The C4.5 algorithm is applied on this datasets. The experimental result shows that the time taken by this algorithm is 14.04 Secs and the memory utilization of this algorithm for the same data set is 6.33 KB.

The Naïve Bayesian classification algorithm is applied with the same data sets and experimental results shows that the time taken by this algorithm is 8.68 Secs and memory utilization of this algorithm is 4.35 KB. Which is comparatively efficient than C4.5 algorithm.

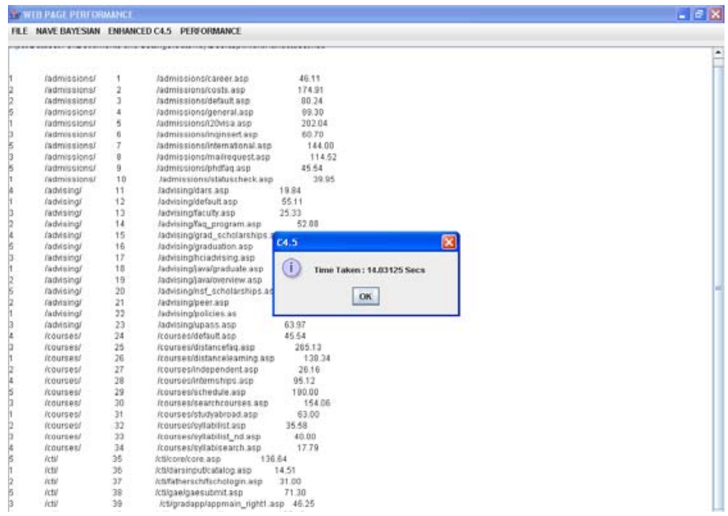

Fig. 5. Experimental Result that the time taken by C4.5 Algorithm.

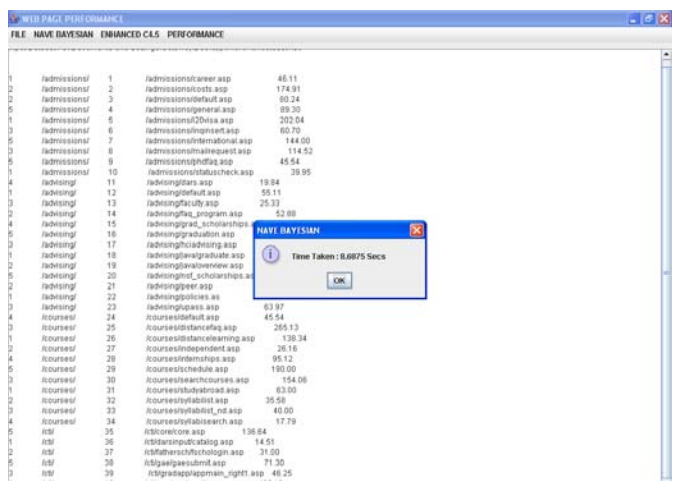

Fig. 6. Experimental Result that the time taken by Naive Bayesian Classification Algorithm.

IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012
ISSN (Online): 1694-0814
www.IJCSI.org

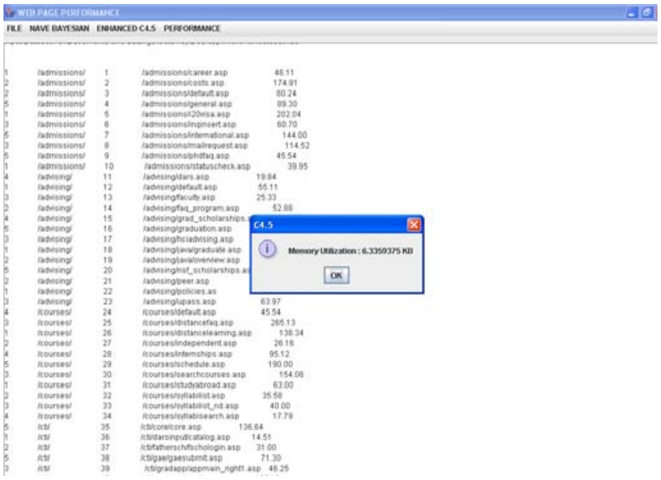

Fig. 7 Experiment Result for Memory Utilization by C4.5 Algorithm

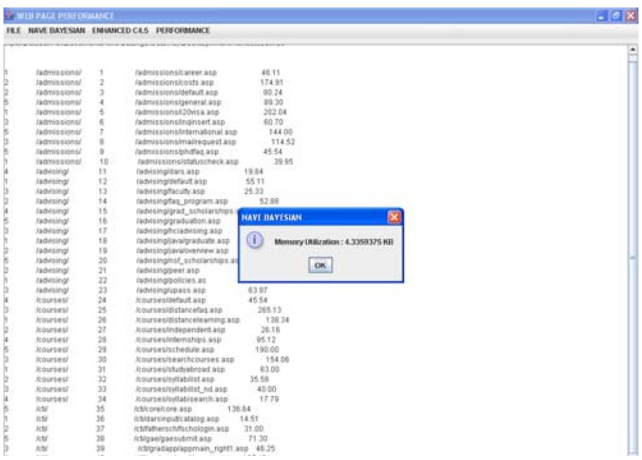

Fig. 7 Experiment Result for Memory Utilization by Naïve Bayesian Classification Algorithm

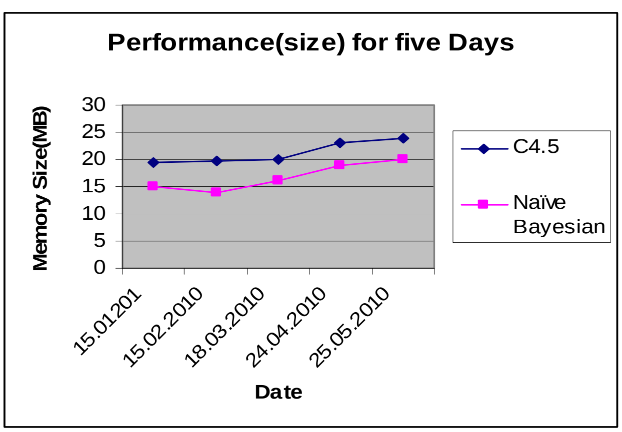

Fig 9.Comparitive analysis of average performance (size) for C4.5 and Naïve Bayesian Classification

The figure shows the performance of memory size utilized by the two models. From the graph it is clear that our proposed model Naïve Bayesian Classification outperforms the existing model C4.5 algorithm in terms of minimal memory consumption.

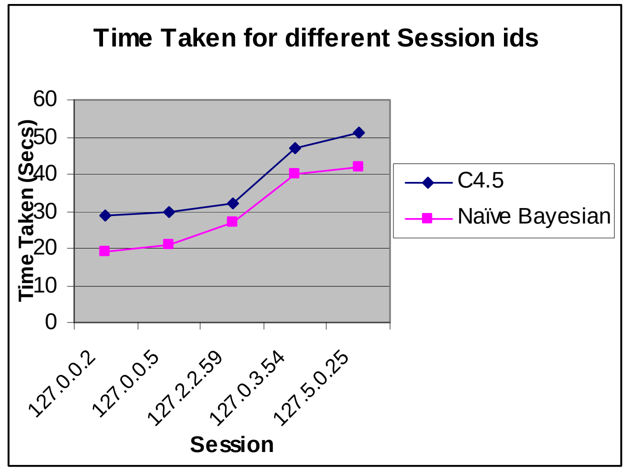

Fig 10. Comparative study for different session ids

Figure 10 shows on the x axis the session id and time taken measured on seconds on y axis for web log data file observed on five different days for different session ids. It shows that the average time taken to compute the maximum likelihood of user preference evaluation in Naïve Bayesian Classification for different session ids is better, when compared with the existing model of enhanced C4.5 decision tree algorithm.

6. Conclusion

Naive Bayesian Classification shows good result in the improvement in time and memory utilization, it can be applied to any web log files. From the experiments conducted, many attributes are not used for classifying as they are irrelevant. With the help of classification the number of irrelevant attributes can be reduced so that the performance can also be proved efficient. We have given a mathematical evaluation of maximum likelihood for the Naive Bayesian Classification which provided a more efficient implementation with a performance increase compared to enhanced C4.5 decision tree. This method can be used in e-commerce applications, such as Web Caching, Web page recommendation, and Web personalization.

References

[1] Jie Zhang and Ali., 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, IEEE, May 2004, pp. 315-322.

[2] Mahesh Thylore Ramakrishna1, Latha Kolal Gowdar, Malatesh Somashekar Havanur and Banur Puttappa Mallikarjuna Swamy, "Web Mining: Key Accomplishments, Applications and Future Directions", International Conference on Data Storage and Data Engineering, February 2010.

[3] Alka Gangrade□, Durgesh Kumar Mishra and Ravindra Patel ,"Classification Rule Mining through SMC for Preserving Privacy Data Mining: A Review", International Conference on Machine Learning and Computing IPCSIT, Singapore, 2009, vol.3 (2011), IACSIT Press, pp. 431 to 434.

[4] Hidenao Abe , "Development of a Classification Rule Mining Framework by Using Temporal Pattern Extraction", New fundamental technologies in data mining, January 2011, pp. 493-504.

[5] Hanady Abdulsalam, David B. Skillicorn, and Patrick Martin, "Classification Using Streaming Random Forests",IEEE Transactions on Knowledge And Data Engineering, January 2011, Vol. 23, No.1., pp.22-36.

[6] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau Dan Lee, "Decision Trees for Uncertain Data", IEEE Transactions On Knowledge And Data Engineering, January 2011, Vol. 23, No. 1, pp 63-78.

[7] J.R, Quinlan , "Induction of Decision Trees", Machine Learning , 1985, Academic Publishers, pp. 81-106.

[8] Veronica S. Moertini, "Towards The Use Of C4.5 Algorithm For Classifying Banking Dataset", Integral , 2003, Vol. 8 No. 2.

[9] Thales Sehn Korting, "C4.5 algorithm and Multivariate Decision Trees", Image Processing Division, National Institute for Space Research – INPE Sao Jose dos Campos – SP, Brazil, 2006.

[10] Salvatore Ruggieri, "Efficient C4.5", IEEE Transactions on Knowledge and Data Engineering, March/ April 2002, Vol. 14, No. 2, pp. 434 -444.

[11] Mahdi Khosravi, Mohammad and J. Tarokh, "Dynamic Mining of Users Interest Navigation Patterns Using Naive Bayesian Method", IEEE 6th International Conference on Intelligent Computer Communication and Processing Transaction, 2010, pp. 119-122.

A. K. Santra received the P. G. degree and Doctorate degree from I.I.T., Kharagpur in the year 1975 and 1981 respectively. He has got 20 years of Teaching Experience and 19 years of Industrial (Research) Experience. His area of interest includes Artificial Intelligence, Neural Networks, Process Modeling, Optimization and Control. He has got to his credit (i) 42 Technical Research Papers which are published in National / International Journals and Seminars of repute, (ii) 20 Research Projects have been completed in varied application areas, (iii) 2 Copy Rights for Software Development have been obtained in the area of Artificial Neural Networks (ANN) and (iv) he is the contributor of the book entitled **"Mathematics and its Applications in Industry and Business"**, Narosa Publishing House, **New Delhi.** He is the recognized Supervisor for guiding Ph. D. / M. S. (By Research) Scholars of Anna University-Chennai, Anna University-Coimbatore, Bharathiyar University, Coimbatore and Mother Teresa University, Kodaikanal. Currently he is guiding 12 Ph. D. Research Scholars in the Department. He is a Life member of CSI and a Life member of ISTE.

S. Jayasudha received her M. C. A., from Periyar University, Salem, M.Phil., from Bharathidasan University , Trichy. Currently she is working as Asst.Professor in Bannari Amman Institute of Technology, Sathyamangalam. Her area of interest includes Web Mining, Text Mining. She is a Life member of Computer Society of India and a Life member of Indian Society for Technical Education.