ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

Implementing Adaboost and Enhanced Adaboost Algorithm in Web Mining

Er.Ramanpreet Kaur[1], Dr. Vinay Chopra[2]
DAVIET, Jalandhar[1,2]

Abstract: The recent explosive growth of the measure of content on the Internet has made it progressively troublesome for clients to discover and use information and content providers find it difficult to classify and catalog documents. It gets highly tedious for users to browse with traditional web search engines as they often return hundreds or a great many results for a search. On-line libraries, web search engines, and other large document repositories (e.g. customer support databases, product specification databases, press release archives, news story archives, etc.) are growing so quickly that it is troublesome and exorbitant to classify each record physically. Keeping in mind the end goal to manage these issues, analysts look toward automated methods of working with web documents so they can be all the more effectively browsed, sorted out, and indexed with negligible human intervention. This paper throws light on the web mining concept and its techniques, explains the data mining process in addition to introducing the classification Adaboost Algorithm. The paper takes a step ahead in this direction and proposes an enhanced Adaboost Algorithm. Both the algorithms have been simulated and their results have been compared in terms of accuracy rate. The results show that Enhanced version shows better performance and accuracy than the AdaBoost Algorithm.

Keywords: Web data mining, information retrieval, Web usage mining, Pre-processing, Pattern Analysis, Content Mining; Structure Mining, Classification.

I. INTRODUCTION

Web information mining grab its significance with the expanding measure of Web data that is ending up being much bigger than any customary data sources. Web data mining includes applying data mining techniques to Web information. It concentrates on the Web pages link structure, their possessed content and their utilization. Web data mining is a procedure that finds the inherent connections among Web information, generally communicated in the forms of textual, linkage or usage information, by means of investigating the elements of the Web and online information utilizing data mining procedures. Web pages are Hypertext documents, which contain both text and hyperlinks to different records. Furthermore, web data are heterogeneous and dynamic. Thus, design and implementation of a web data mining research support system has turned into a challenge for researchers in order to utilize useful information from the web.

The web mining task can be categorized to following sub tasks [4]:

- Resource Finding: the retrieval of indented Web documents.
- Information Selection and Pre-processing: automatically selecting and pre-processing specific information from retrieved web resources.
- Generalization: discovering general patterns at individual web sites as across multiple sites.
- Analysis: validation and /or interpretation of the mined patterns.

Web Mining joins three stages: pre-processing, pattern discovery and pattern analysis. The pattern discovery has been an imperative study in the improvement of the efficiency of various Web based applications. Nowadays, web applications provide a more personalized experience for their users which make it extremely important to form some kind of interaction with Web users and always be one step ahead of them when it comes to predicting next accessed pages. For instance, thoroughly understanding all about the user's browsing history on the site serves to know which one among the most frequently accessed pages will be accessed next along with some extra information like the type of users and the user's preferences. Such objectives can be achieved by extracting useful knowledge and patterns applying different tools and techniques. Each of these pattern discovery techniques has its own particular strengths and weaknesses [3].

II. WEB MINING TECHNIQUES

The World Wide Web has developed in the past couple of years from a little research group to the greatest and most well-known method for communication and information dissemination. It has lot of information and keeps on expanding in size and complexity. It is extremely colossal task to seek relevant information from an enormous amount of data.

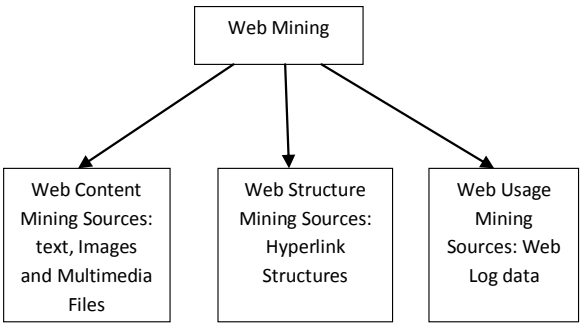

Fig.1 Web Mining Techniques

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

A. Web Content Mining:

Web Content Mining is the procedure of extracting relevant useful data from the contents of Web records. Content information relates to the collection of facts a Web page was intended to pass on to the clients. It may comprise of text, images, audio, video, or structured records such as lists and tables. Research exercises in this field additionally include utilizing systems from different disciplines, for example, Information Retrieval (IR) and natural language processing (NLP) [2].

B. Web Structure Mining

A typical Web graph consists of components including Web pages as nodes, and hyperlinks as edges associating between two related pages. Moreover, Web page content can likewise be sorted out in a tree structured format, based on the various HTML and XML tags within the page. Hence, Web Structure Mining can be viewed as the phenomenon of discovering structure information from the Web. Such kind of mining can be performed at two levels either at the (intra-page) document level or at the (inter-page) hyperlink level [2].

The main hurdle in Web structure mining is to manage the structure of the hyperlinks within the Web itself. With the escalating interest in Web mining, the research of structure analysis has increased and these efforts has resulted in a quickly emerging research area called Link Mining, which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a conceivably extensive variety of application areas for this new area of research, including Internet [6].

C. Web Usage Mining

Web usage mining is the third category in web mining. This type of web mining allows discovering interesting usage patterns from Web data to understand and better serve the needs of Web based applications. This usage data provides the paths leading to accessed Web pages.
This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access. It involves the automatic discovery and analysis of patterns in data as a result of the user's interactions with one or more Web sites. It focuses on tools and techniques used to study and understand the users' navigation preferences and behaviour by discovering their Web access patterns. The usage data that is gathered provides the companies with the ability to produce results more effective to their businesses and increasing of sales. Usage data captures the identity or origin of web users along with their browsing behaviour at a web site.

III. WEB PERSONALIZATION

Web users' content personalization and recommendation of appropriate Web pages implies being able to supply users with Web contents according to their specific tastes or preferences based on their previous interaction history within the same Web site without expecting from them to ask for it explicitly". Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in content delivery framework. A personalization mechanism is based on explicit preference declarations by the user and on an iterative process of monitoring the user navigation, collecting its requests of ontological objects and storing them in its profile in order to deliver personalized content [3]. Personalization can also be valuable to an individual or any organization, because it drives desired business results such as increasing visitor response or promoting customer retention [2].

IV. DATA MINING

The advances for producing and collecting data have been advancing rapidly. At the current stage, lack of data is no longer a problem; rather the inability to generate useful information from data is. The explosive growth in data and database brings about the need to develop new technologies and tools to process data into useful information and knowledge intelligently and automatically. Data mining (DM), therefore, has become a research area with increasing importance [9]. It is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems [8].

Data mining, popularly known as Knowledge Discovery in Databases (KDD), is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. It is actually the process of finding the hidden information/pattern of the repositories [10].

The KDD process includes an iterative sequence method [11], [12]:

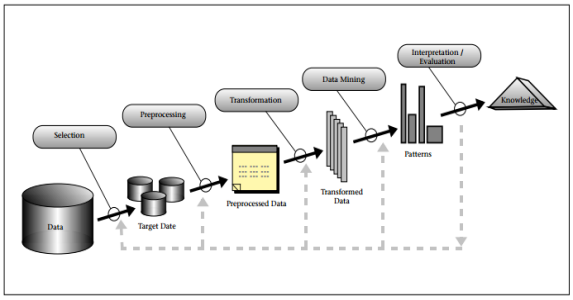

Fig.2 KDD steps

- Selection: The first step starts with collecting the necessary and relevant knowledge about the domain and setting the goals to be achieved. This information is then used in the preparation of a dataset which includes selecting an appropriate (sub) set of data samples and/or variables.
- Cleaning and Pre-processing: It includes finding incorrect or missing data. It also includes removal of noise or outliers, collecting necessary information to model or account for noise, accounting for time sequence information and known changes.
- Transformation: It is converting the data into a common format for processing. Some data may be encoded or transformed into more usable format. Data reduction, dimensionality reduction & data transformation method may be used to reduce the number of possible data values being considered.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

- Data Mining: This is the most elaborate step as it consists of choosing the function of data mining, choosing the right data mining algorithm and its application. Choosing the function includes deciding the purpose of the resulting data mining model, such as classification, regression, clustering and summarization. The selection of the data mining algorithm encompasses the decision which models and parameters are appropriate and matching with the criteria of the process.
- Interpretation/Evaluation: In the last step the discovered patterns are evaluated and their validity and relevance are assessed. Redundant and irrelevant patterns are removed while the remaining, relevant patterns are studied and interpreted, typically with the aid of computer-assisted visualization techniques. Application of the discovered knowledge includes resolving potential conflicts with existing knowledge, taking actions based on the obtained knowledge, such as aiding, modifying and improving existing processes and procedures, especially those involving human experts, and storing, documenting and reporting to interested parties.

V. FILTERING IN DATA

Classical methodologies from information retrieval/filtering (IR/IF) and data mining have been applied separately successfully in dealing with Web information overload and mismatch issues. Information filtering is an information access activity similar to information retrieval the only difference being IF systems are commonly personalized to support long-term, relatively stable, or periodic goals.
The objective of filtering is to quickly filter out the likely irrelevant data and it is expected that unmatched data can be greatly reduced after filtering. More sophisticated data processing can then be carried out on the obtained cleaned data effectively. As a result, Web mining system could be more efficient to deliver the users with more relevant results.

A data set usually may contain noisy or redundant data items and large number of features, large portions of them may not be relevant for the objective function at hand. Thus noise data may degrade the accuracy and performance of the classification models. Therefore, dealing with missing values in data pre-processing is an important step in building an effective and efficient classifier. It is a procedure by which missing quantities are replaced by suitable quantities with respect to an objective function or the noisy data may be filtered. It results in better performance of the classification models in terms of their predictive or descriptive accuracy, diminishing of computing time needed to build models as they learn faster and better understanding of the models [22].

In our approach, Replace Missing Values unsupervised Filter has been used to replace all missing values using means and modes. Data cleaning particularly, dealing with missing values is an important step in the knowledge discovery in database (KDD) process. In general, it improves the predictive capability of the classifiers [22].

VI. CLASSIFICATION ALGORITHM

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. The data classification process involves learning and classification. In Learning the training data are analysed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples.

The Classification process involves following steps [23]:
a. Create training data set.
b. Identify class attribute and classes.
c. Identify useful attributes for classification (Relevance analysis).
d. Learn a model using training examples in Training set.
e. Use the model to classify the unknown data samples.

Adaboost Algorithm

The Adaboost classifier can be built using fewer features and is considered more appropriate for real time applications. The selected AdaBoost classifiers can improve the classification accuracy as well as reduce the processing time and perform reliably better for defect classification. The benefits of AdaBoost include less memory and computation necessities. The real Adaboost algorithm gives minor error rates than the diverse AdaBoost.

AdaBoost is an algorithm for constructing a strong classifier as linear combination of weak classifiers ht(x). AdaBoost is adaptive only for this reason that subsequent classifiers built are weaken in favour of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers.

$$f(x)= \sum_{t=1}^{T} \alpha t\, ht(x)$$

where ht:X-> {1,-1} and αt ϵ R

A weak classifier is a very simple model that has just slightly better accuracy than a random classifier, which has 50% accuracy on the training data set. The set of weak classifiers is built iteratively from the training data over hundreds or thousands of iterations. At each iteration or round, the examples in the training data are reweighted according to how well they are classified. Weights are computed for the weak classifiers based on their classification accuracy.

The assigned weight is used to vote for each classifier. If there is less error rate of classifier then more weight assigned to its vote. This training process is repeated. The weight of classifiers which voted for an object of a class is added. The class which gains higher total weight is the final class and it will introduced as the predictive class for that object.

AdaBoost is a powerful classification algorithm that has enjoyed practical success with applications in a wide variety of fields, such as biology, computer vision, and speech processing. Unlike other powerful classifiers, such as SVM, AdaBoost can achieve similar classification

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

results with much less tweaking of parameters or settings. The user only needs to choose:

- which weak classifier might work best to solve their given classification problem;
- the number of boosting rounds that should be used during the training phase.

The AdaBoost algorithm will select the weak classifier that works best at that round of boosting.

Enhanced AdaBoost Algorithm

The flowchart for enhanced Adaboost algorithm is as:

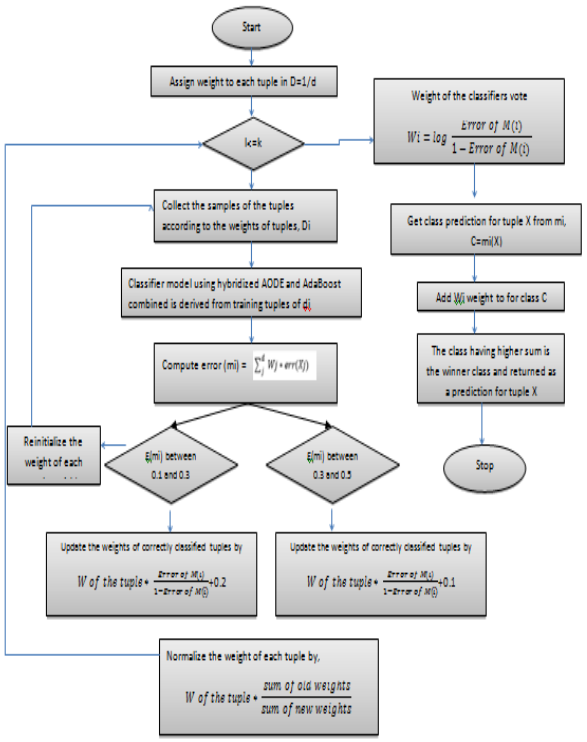

Fig.3 Enhanced Adaboost Algorithm

VII. MOTIVATION

As per the base paper, no algorithm has been used for classification. Every categorization has been performed on the basis of searching technique and extracting results from the dataset accordingly. But the proposed approach includes the use of AdaBoost algorithm for the same and also improves the performance of AdaBoost. Boosting is the machine learning method for improving the performance of any learning algorithm on the idea of creating a high accurate prediction rule by combining various weak classifiers and non-appropriate rules. AdaBoost is the first practical boosting algorithm with good results in many areas, but mostly in text classification. Research in this area has uncovered links to other algorithms such as Support Vector Machines. It was first presented by Schapire and Freud.

Further research on boosting techniques introduced a new boosting algorithm called AdaBoost Algorithm.

1. Previously no technique has been used with AdaBoost for classifying large training set in such an efficient manner to give quick results.
2. Using the proposed approach, large number of web datasets can be handled.
3. In this proposed approach weak Learners is not too simple as it is hybrid with another classification algorithm.
4. The proposed approach is used to improve the performance of a learning algorithm.
5. Weak classifier has been combined with non-appropriate rules to generate high prediction rule

VIII. PROPOSED SCHEME

The Research Methodology includes the following steps:

1) Collection of raw data and then apply filtering techniques to make that raw data into structured format: Filtering techniques like Replace Missing Value filter
2) Applying the AdaBoost algorithm on the collected data and classify the data according to the class attribute.
3) Apply the enhanced AdaBoost algorithm for classification.
   1. Replace the weak learner of AdaBoost with hybrid classifier that contains AdaBoost algorithm and AODE which are being hybridized on the basis of average of their probabilities.
   2. Add more decision making conditions while calculating the class for model prediction i.e. on the basis of error rate adds more weight to class that will give better class for the prediction.
4) Analyze the performance parameters like FP rate, TP rate, Recall, Precision of AdaBoost and new proposed enhanced algorithm and Compare the results of both.

The flowchart explains the methodology of the proposed approach:

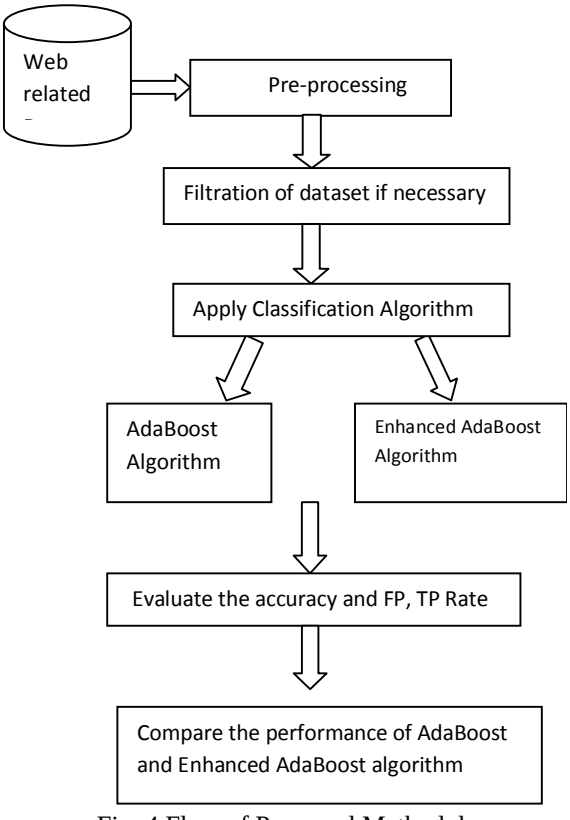

Fig. 4 Flow of Proposed Methodology

DOI 10.17148/IJARCCE.2015.4771

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

IX. RESULTS

This section presents the simulation results. The proposed approach has been simulated in Java NetBeans.

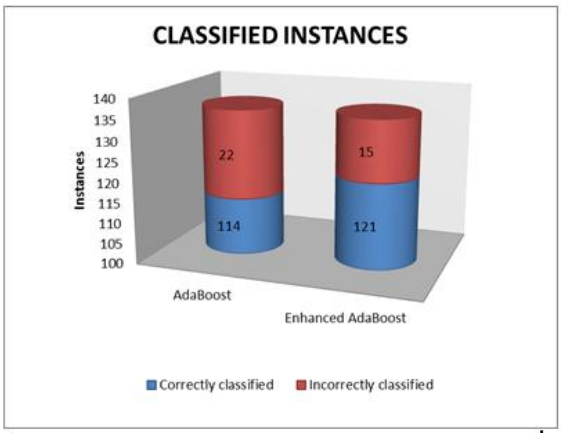

Fig. 5 shows that In Web mining dataset, out of 136 instances, 114 instances are correctly classified for AdaBoost, and 121 instances are correctly classified for Enhanced Ada Boost.

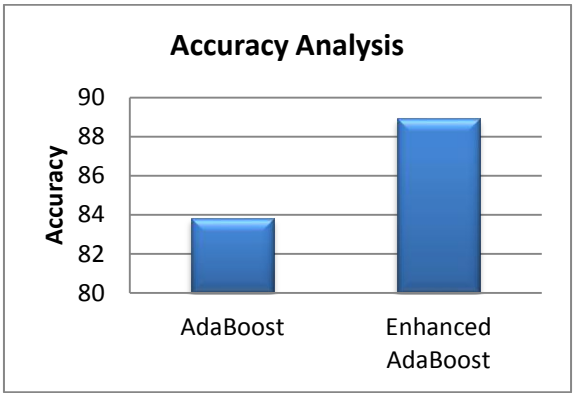

Fig.6 shows that Accuracy for Enhanced AdaBoost is around 88.97% and for AdaBoost is around 83.82%

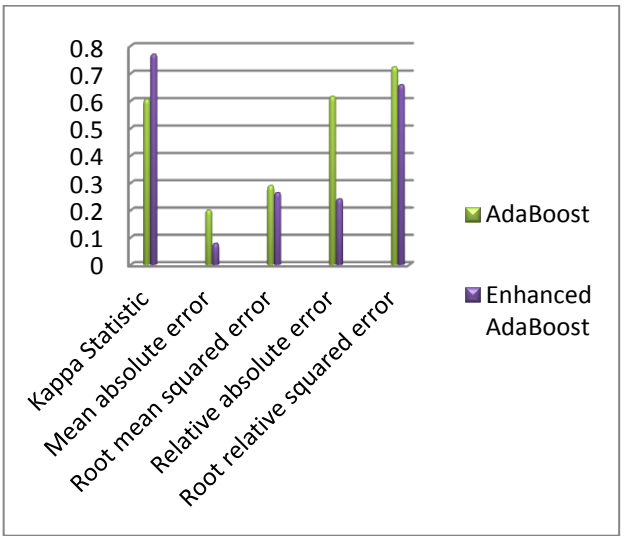

Fig.7 Evaluation of Parameter

X. CONCLUSION

In this paper, three different modes of web mining, namely web content mining, web structure mining and web usage mining have been elaborated. Needless to say, these three approaches cannot be independent, and any efficient mining of the web would require a judicious combination of information from all the three sources. Research in text mining is at the moment very general in nature hence to deal with data an algorithm named AdaBoost has been used for text classification. The main motive is to analyze the relevant information of different websites and find patterns and make predictions. In addition to this, in this paper, an enhanced version of AdaBoost has also been proposed and implemented. The simulation results using AdaBoost and Enhanced Adaboost, the results have been presented. It has been observed that Accuracy for Enhanced AdaBoost is around 88.97% and for AdaBoost is around 83.82% which clearly shows that the proposed approach is better in performance.

ACKNOWLEDGMENT

The paper has been written with the kind assistance, guidance and active support of my department who have helped me in this work. I would like to thank all the individuals whose encouragement and support has made the completion of this work possible.

REFERENCES

[1] Jin Xu Yingping Huang Gregory Madey, "A Research Support System Framework for Web Data Mining",
[2] A.Jebaraj Ratnakumar, "An Implementation Of Web Personalization Using Web Mining Techniques", Journal of Theoretical and Applied Information Technology © 2005 - 2010 JATIT
[3] Pooja Mehtaa, Brinda Parekh, Kirit Modi, and Paresh Solanki, "Web Personalization Using Web Mining: Concept and Research Issue", International Journal of Information and Education Technology, Vol. 2, No. 5, October 2012
[4] Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", Volume 3, Issue 11, November 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering
[5] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining—Concepts, applications, and Research Directions"
[6] T.Nithya, "Link Analysis Algorithm for Web Structure Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013
[7] Pradnyesh Bhisikar 1,Prof. Amit Sahu, "Overview on Web Mining and Different Technique for Web Personalisation", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 3, Issue 2, March -April 2013, pp.543-545
[8] Jayanthi Ranjan, "Applications of Data Mining Techniques in Pharmaceutical Industry", Journal of Theoretical and Applied Information Technology © 2005 - 2007 JATIT
[9] Sang Jun Lee, Keng Siau, "A review of data mining techniques", Industrial Management & Data Systems 101/1 [2001] 41-46
[10] Neelamadhab Padhy1, Dr. Pragnyaban Mishra 2, and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012
[11] Pramod Kumar Joshi1 and Sadhana Rana, "Era of Cloud Computing", High Performance Architecture and Grid Computing Communications in Computer and Information Science Volume 169, 2011, pp 1-8 ISSN 1865-0929, DOI 10.1007/978-3-642-22577-2_1, Springer-Verlag Berlin Heidelberg 2011
[12] Eman Elghoniemy, Othmane Bouhali, Hussein Alnuweiri, "Resource Allocation and Scheduling in loud Computing", 978-1-4673-0009-4/12, IEEE 2012

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

[13] Mrs. Bharati M. Ramageri, "Data Mining Techniques And Applications", Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305
[14] Kalyani M Raval, "Data Mining Techniques", Volume 2, Issue 10, October 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering
[15] Periyasamy. N Thamilselvan. P Dr. J. G. R. Sathiaseelan, "A Comparative Study of ANN, K- Means and Adaboost Algorithms for image classification", ResearchGate, IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15
[16] XindongWu • Vipin Kumar • J. Ross Quinlan • Joydeep Ghosh • Qiang Yang Hiroshi Motoda • Geoffrey J. McLachlan • Angus Ng • Bing Liu • Philip S. Yu Zhi-Hua Zhou • Michael Steinbach • David J. Hand • Dan Steinberg, "Top 10 algorithms in data mining", Published online: 4 December 2007 © Springer-Verlag London Limited 2007
[17] Yoav Freund Robert E. Schapire, "A Short Introduction to Boosting", Journal of Japanese Society for Artificial Intelligence, 14(5):771-780, September, 1999
[18] A. K. Santra1, S. Jayasudha, "Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012 ISSN (Online): 1694-0814
[19] Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations. Copyright 2000 ACM SIGKDD, July 2000.
[20] K. Sudheer Reddy, G. Partha Saradhi Varma, and M. Kantha Reddy, "An Effective Preprocessing Method for Web Usage Mining", International Journal of Computer Theory and Engineering, Vol. 6, No. 5, October 2014
[21] D. Jayalatchumy, Dr. P.Thambidurai, "Web Mining Research Issues and Future Directions – A Survey", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 3 (Sep. - Oct. 2013), PP 20-27
[22] Tapas Ranjan Baitharu and Subhendu Kumar Pani, "Effect of Missing Values on Data Classification", Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS) 4(2): 311-316 © Scholarlink Research Institute Journals, 2013 (ISSN: 2141-7016)
[23] Trilok Chand Sharma1, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013
[24] Raj Kumar, "Classification Algorithms for Data Mining: A Survey", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 1 Issue 2 August 2012, ISSN: 2319 – 1058
[25] Prabhjot Kaur, "Web Content Classification: A Survey", International Journal of Computer Trends and Technology (IJCTT) – volume 10 number 2 – Apr 2014
[26] Ahmad Mahir R. and Al-khazaleh A. M. H, "Estimation of missing data by using the filtering process in a time series modeling", rXiv:0811.0659v1 5 Nov 2008