



# Identifying Web Users from Weblogs Using Classification Algorithms

V.Vidyapriya<sup>1</sup>, V. Pushpa<sup>2</sup>

Associate Professor, Dept. of Comp. Sci., Quaid-E-Millath Govt College for women (Autonomous), Anna Salai,  
Chennai, India

M.Phil Research Scholar, Dept. of Comp. Sci., Quaid-E-Millath Govt College for women (Autonomous), Anna Salai,  
Chennai, India

**ABSTRACT:** The collection of abundant information sources accessible on the web today provides ideal opportunities and challenges to web mining for a varied range of applications. Web mining is the integration of much information gathered by data mining techniques and methodologies which is used to extract information from web data. It has three general categories, namely web content, web structure, and web usage mining. The web usage mining is used to determine valuable information from navigation of web users. It has three phases such as data pre-processing, pattern discovery and pattern analysis. This paper intends pattern discovery using various classification techniques to determine which classification technique such as Naïve Bayesian, CART, k-nearest neighbour which has the maximum accuracy and minimum error rate. The primary objectives of this paper are to identify the interest of user access pattern from the weblogs defining specific website.

**KEYWORDS:** web usage mining; data preprocessing; pattern discovery; classification algorithms; weka; rapid miner.

## I. INTRODUCTION

Now days, the World Wide Web contains the enormous information which is rapidly developing till the billions of users day by day increase and their needs also grows. It is very tough task to retrieve the exact information from web pages so one of the application techniques of data mining can be used to retrieve the data such as web mining. It used to discover or extract the knowledge from web files which has three general categories, namely web content, web structure, web usage mining. Web content mining is used to mine the content of a web page such as image, video, text etc. It is focused on the structure of the inner document level of web pages while web structures mining focused link structure of the inter document level with the related web pages. Web structure mining is used to generate the structural summary of web sites. Web usage mining is the final phase of web mining, which is used to mine the interesting patterns from web logs. It helps to know the browser activities of websites which has three stages such as data preprocessing, pattern discovery, pattern analysis[1].

This paper deals with data preprocessing and pattern discovery of web usage mining. Data preprocessing is used to remove the irrelevant, noisy data from weblogs. Pattern discovery is the next phase of preprocessing which takes the input as preprocessed weblogs. In pattern discovery, the data mining techniques of classification, clustering, association and machine learning can be applied, which is used to discover the interesting patterns. In this paper, classification techniques like KNN, CART, and Naïve Bayesian are used to identify the web users based on the accuracy and error rate using data mining tools weka and rapid miner.

## II. RELATED WORK

In [2] the author has done an experiment on the college data set to find the user access patterns with use of naïve bayesian algorithm in weka tool. The main goal of this paper to identify the browser behaviour but the naïve bayesian is not classified in the undefined class. In [3] the author proposed efficient technique cart algorithm which is to identify the interested user. This is not only reducing the irrelative attributes but also reduces the error rate. The proposed cart

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

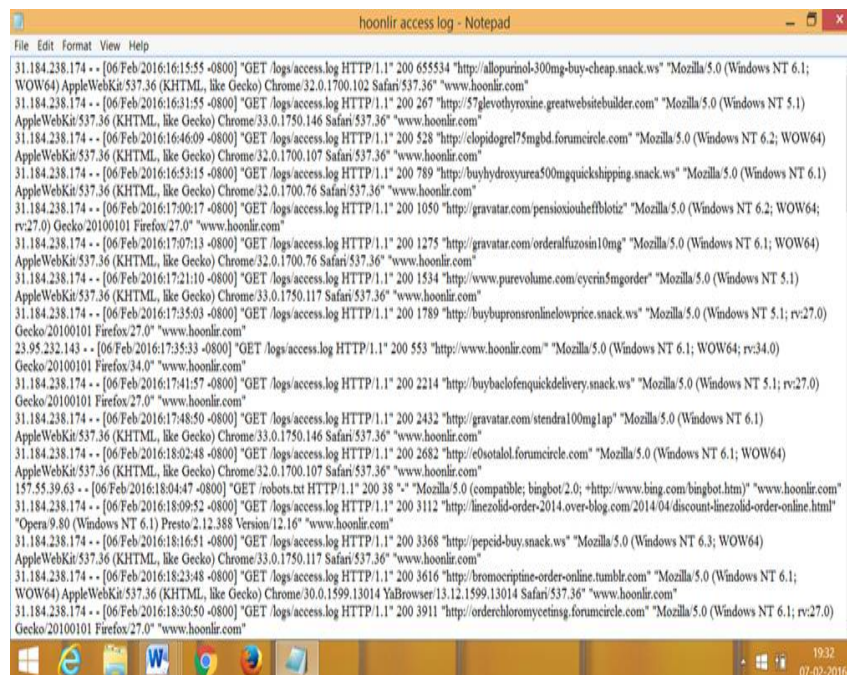
algorithm is better than existing algorithm of Naïve bayesian. In [4] the author has a study of automatic web usage data of recommendation system using k-nearest neighbour on RSS web site. The result of K-NN classifier is transparent, consistent, straightforward and simple.

## III. DATA PREPROCESSING

Data preprocessing is the fundamental steps of web usage mining. In this step the log file will be cleaned which means to remove the noisy, incomplete data from weblog file and identify the user session identification.[6]

### A. Log file:

The web server log file collected from” <http://www.hoonlir.com/log/> access.log” which is dated from 9.1.2016 to 15.1.2016. There are totally 4193 raw log entries are presented.



```
31.184.238.174 - - [06/Feb/2016:16:15:55 -0800] "GET /logs/access.log HTTP/1.1" 200 655534 "http://allopurnol-300mg-buy-cheap.snack.ws" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.102 Safari/537.36" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:16:31:55 -0800] "GET /logs/access.log HTTP/1.1" 200 267 "http://57glevothyroxine.greatwebsitebuilder.com" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.146 Safari/537.36" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:16:46:09 -0800] "GET /logs/access.log HTTP/1.1" 200 428 "http://clopidogrel75mgbd.forumcircle.com" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safari/537.36" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:16:53:15 -0800] "GET /logs/access.log HTTP/1.1" 200 789 "http://buyhydroxyurea500mgquickshipping.snack.ws" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.76 Safari/537.36" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:17:00:17 -0800] "GET /logs/access.log HTTP/1.1" 200 1050 "http://gravatar.com/pensioiouhefflotiz" "Mozilla/5.0 (Windows NT 6.2; WOW64; rv:27.0) Gecko/20100101 Firefox/27.0" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:17:07:13 -0800] "GET /logs/access.log HTTP/1.1" 200 1275 "http://gravatar.com/orderalfuzosini0mg" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.76 Safari/537.36" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:17:21:10 -0800] "GET /logs/access.log HTTP/1.1" 200 1534 "http://www.purevolume.com/eycirin5mgorder" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.117 Safari/537.36" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:17:35:03 -0800] "GET /logs/access.log HTTP/1.1" 200 1789 "http://buybuprenonlineowprice.snack.ws" "Mozilla/5.0 (Windows NT 5.1; rv:27.0) Gecko/20100101 Firefox/27.0" "www.hoonlir.com"
23.95.232.143 - - [06/Feb/2016:17:35:33 -0800] "GET /logs/access.log HTTP/1.1" 200 553 "http://www.hoonlir.com" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:34.0) Gecko/20100101 Firefox/34.0" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:17:41:57 -0800] "GET /logs/access.log HTTP/1.1" 200 2214 "http://buybaclafenquickdelivery.snack.ws" "Mozilla/5.0 (Windows NT 5.1; rv:27.0) Gecko/20100101 Firefox/27.0" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:17:48:50 -0800] "GET /logs/access.log HTTP/1.1" 200 2432 "http://gravatar.com/stendra100mg1ap" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.146 Safari/537.36" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:18:02:48 -0800] "GET /logs/access.log HTTP/1.1" 200 2682 "http://e0total.forumcircle.com" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safari/537.36" "www.hoonlir.com"
157.55.29.63 - - [06/Feb/2016:18:04:47 -0800] "GET /robots.txt HTTP/1.1" 200 38 "" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:18:09:52 -0800] "GET /logs/access.log HTTP/1.1" 200 3112 "http://linezolid-order-2014.over-blog.com/2014/04/discount-linezolid-order-online.html" "Opera/9.80 (Windows NT 6.1) Presto/2.12.388 Version/12.16" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:18:16:51 -0800] "GET /logs/access.log HTTP/1.1" 200 3368 "http://pepcid-buy.snack.ws" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.117 Safari/537.36" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:18:23:48 -0800] "GET /logs/access.log HTTP/1.1" 200 3616 "http://bromocriptine-order-online.tumblr.com" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.13014 YaBrowser/13.12.1599.13014 Safari/537.36" "www.hoonlir.com"
31.184.238.174 - - [06/Feb/2016:18:30:50 -0800] "GET /logs/access.log HTTP/1.1" 200 3911 "http://orderchloromycesinsg.forumcircle.com" "Mozilla/5.0 (Windows NT 6.1; rv:27.0) Gecko/20100101 Firefox/27.0" "www.hoonlir.com"
```

Fig 1. Sample raw logs

### B. Data cleaning:

In data cleaning the unwanted logs files will be eliminated such as the image file (.jpeg, .jpg), failed http code (error code), spider log (robots.txt), and other file like style sheet (.css, .log) which type of log files will be cleaned in data cleaning step[7].

### C. User/session identification:

The next step of data cleaning is user identification and session identification. User identification is identifying the web user with their unique IP address which is used to know about who accessed the web page. Session identification is used to identify how long the users spend on the web page. The default session timing is 30 minutes if the time exceeds more than 30 minutes the next session will be started even the same user spend on the same web page [8]. There are 253 unique users are identified in the preprocessing step which the each unique user is spending 30 minutes on each session.

## IV. PATTERN DISCOVERY



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Pattern discovery is the one of the phases of web usage mining. In this stage, the various data mining techniques available such as classification, clustering, association rules, sequential pattern. It can be applied on the preprocessed weblogs and discover the useful patterns which is used to easily know the browser behaviour[8]. This paper proposed to classify the weblogs using classification algorithms based on the accuracy and error rate.

## A. Existing system:

In data mining, classification is the supervised learning function that assigns objects in a collection to target class or categories. The main purpose of the classification is to precisely predict the target class or objects whose class label is indefinite. There are various techniques available in the classification that is naïve Bayesian, decision tree, cart, k- nearest neighbour, rule based, support vector machine all of these algorithms help to predict the class name which the item's label is unknown. In the existing system, the Naïve Bayesian (NB) classification technique using weka tool for identifying the constant access pattern and to categorize the browser behaviour of the user.[2]

The naïve bayesian (NB) is the one of the familiar classification technique which is used in the existing system using weka tool for identifying constantly the access pattern and to categorize the browser behaviour of the user. The naïve Bayesian shows better result in time and memory utilization it can be applied to any weblog files [5], but the fact is many attributes are not used for classifying as they are irrelevant. The naïve Bayesian has low level time complexity, but the efficiency as per the accuracy and error rate is not adequate. This paper is induced compared to other two classification algorithms, namely k-nearest neighbour, cart using weka and rapid miner tool for identifying the frequent web user from preprocessed weblogs.

## B. Proposed system:

In the proposed system, the other two classification algorithms compared to Naïve Bayesian such k-nearest neighbour and simple cart algorithm. The K- nearest neighbour (K-NN) is one of classification technique that is very simple and easy to understand but works extremely well in practice. The K-NN is a non- parametric lazy learning algorithm, which needs to establish a consistent, flexible [4]. The K-NN uses the simple Euclidean distance to measure the closeness between the test tuples and training tuples. The simple CART such as a decision tree algorithm abbreviates classification and regression tree. It is a classification technique which is used for data analysis and prediction, mainly this algorithm best for the training tuples. The simple cart used the best splitting attribute is entropy, which generates only two children[3].

Formula for Accuracy:

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}} \times 100$$

Formula for Error Rate:

$$\text{Error Rate} = \frac{\text{Number of incorrectly classified instances}}{\text{Total number of instances}} \times 100$$

The preprocessed weblogs are loaded into the weka and rapid miner tool. The classification algorithms of K-nearest neighbour, Naïve Bayesian, cart are applied to the web logs.

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

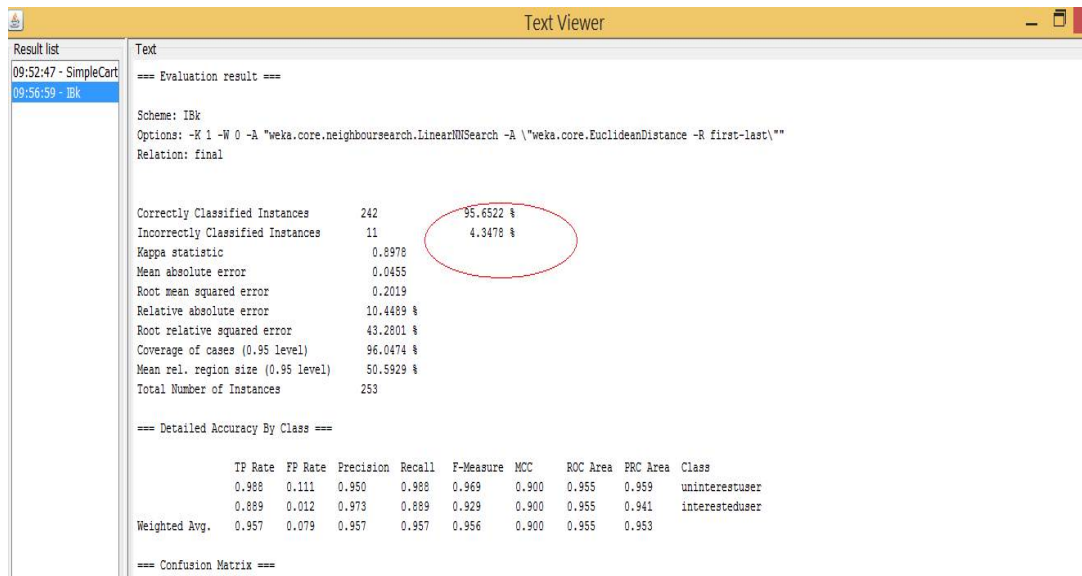


Fig 2. Accuracy and Error rate of KNN in Weka

In Fig 2 shows the k -nearest neighbour classification algorithm applied to weblogs using weka tools. It shows the accuracy and the error rate of k- nearest neighbour which is the total number of input instances is 253. The number of correctly classified instances 242 and the number of incorrectly classified instances is 11. So the accuracy of k- nearest neighbour is 95.622 and the error rate is 4.3478. The k- nearest neighbour displays the highest accuracy and lowest error rate in the weka tool.

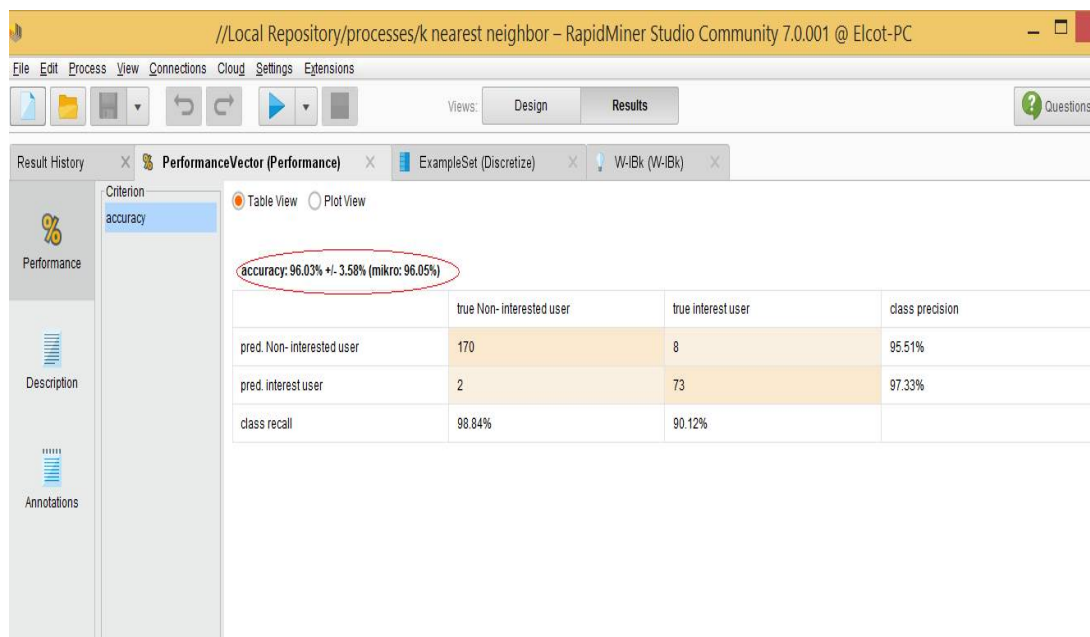


Fig 3. Accuracy and Error rate of KNN in Rapid Miner

In Fig 3 shows the classification technique of k-nearest neighbour are applied on the preprocessed weblogs using the rapid miner tool. It is familiar and user friendly tool of data mining. The weblogs are well classified using

# International Journal of Innovative Research in Computer and Communication Engineering

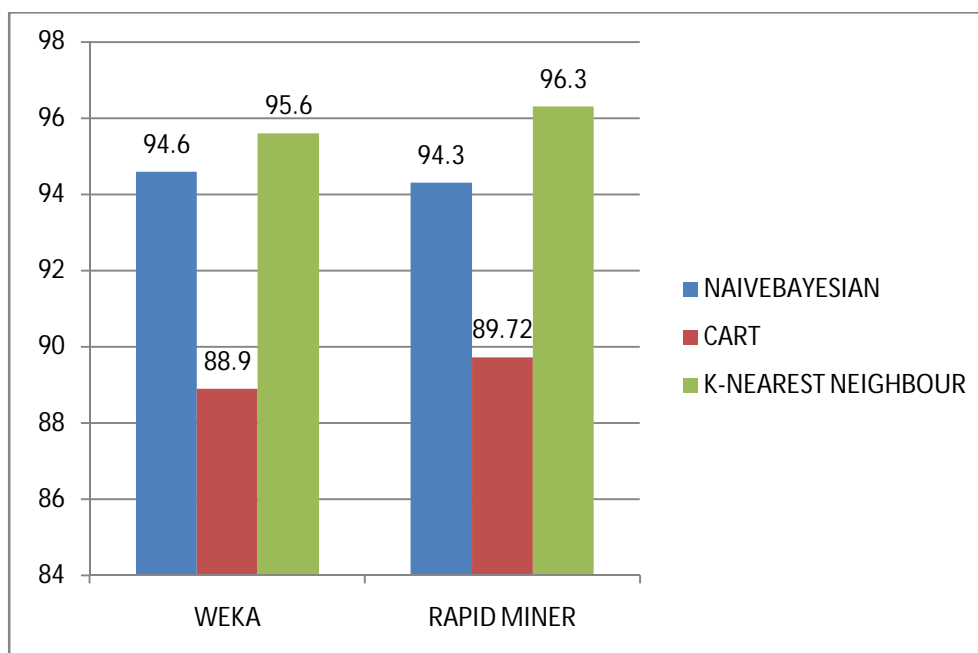
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

classification algorithms in the rapid miner tool. In rapid miner tool, there are total numbers of input instance is 253. The preprocessed weblogs are classified in that there are 172 non interested users and 81 interested users are presents which means there are 81 users only to view and place the order of product in that websites other 172 users are not placing any order in that website but they just visit the websites. The k- nearest neighbour shows the accuracy is 96.03 and the error rate is 3.05. Another two algorithms like naïve bayesian and cart not shows this much level of high accuracy and low error rate.

## V. RESULT AND FINDINGS

This paper conferred about identifying web user access pattern from weblogs using various classification algorithms, namely K- nearest neighbour, naïve bayesian, cart which is based on the accuracy and error rate.



**Fig 4. Accuracy of classification techniques**

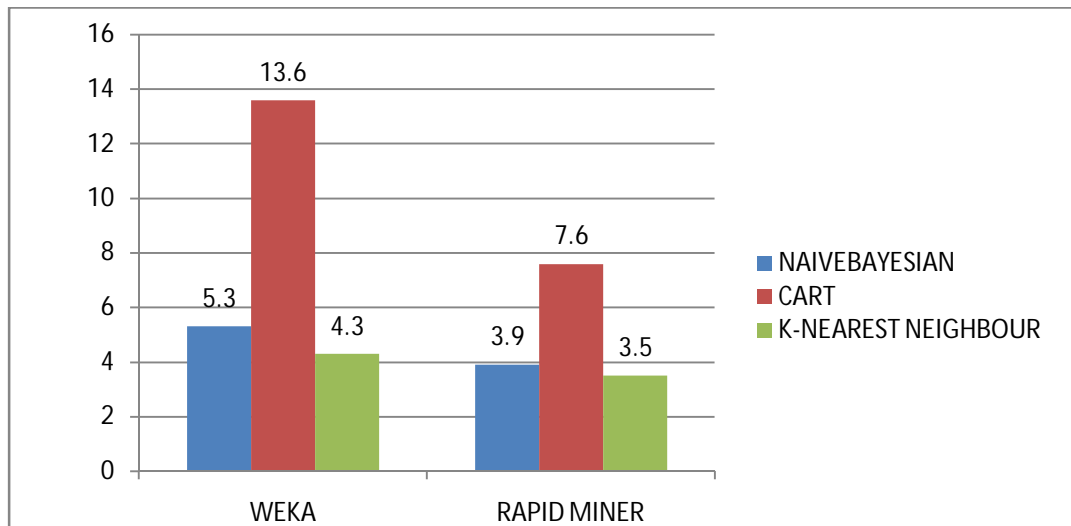
Fig 4. Shows the accuracy of various classification techniques in weka and rapid miner tools gives almost same accuracy of each algorithm. The total number of taken input instance is 253. The naïve bayesian has correctly classified instance is 240 so the accuracy of naïve bayesian is around 94 in both weka and rapid miner. The simple cart algorithm shows 88.9 accuracy in weka and 89.72 in rapid miner tool because the simple cart has correctly classified the instance is only 225. The K-nearest neighbour gives the 95.6 accuracy in weka and 96.3 in rapid miner which is 242 instances are correctly classified by K-nearest neighbour. According to weka and rapid miner KNN has high performance compared to other two algorithms. The K-nearest neighbour gives the maximum accuracy and minimum error rate and cart algorithm give the lowest accuracy and highest error rate in both tools. Compared to naïve bayesian and simple cart the K-nearest neighbour has correctly classified the instances with better accuracy.



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016



**Fig 5. Error rate of classification techniques**

The above Fig 5 shows the error rate of classification techniques which helps to know the minimum error rate belongs to which classification algorithm. The error rate of naïve bayesian is 5.3 in weka and 3.9 in rapid miner which means 13 instances are incorrectly classified by naïve bayesian. The simple cart algorithm has highest error rate in weka as well as rapid miner such as 13.6 and 7.6 error rate is occurring because 28 instances are incorrectly classified by simple cart algorithm. The K- nearest neighbour gives around 3 to 4 range of error rate in both weka and rapid miner tools which are 11 instances are only incorrectly classified by K- nearest neighbour. Among these three classification algorithms the simple cart gives a maximum error rate, but the k-nearest neighbour shows the minimum rate in both tools.

## VI. CONCLUSION AND FUTURE WORK

The web is the huge repository of web documents which is a very difficult task to retrieve the exact data from web pages. Web mining technique is very helpful in mining the data from web pages. In this paper, the data preprocessing of weblogs is done effectively, which the preprocessed weblogs are taken as input in pattern discovery. In pattern discovery the k- nearest neighbour classification algorithm performs well compared to naïve bayesian and cart algorithm. The k- nearest neighbour algorithm shows a maximum accuracy and minimum error rate in both weka and rapid miner tools. It's very helpful to find the frequent web users from web logs of the specific website. The future work of this paper to use various classification algorithms to be applied on web logs which is used to predict the page rank of website such as commercial or non-commercial web pages.

## REFERENCES

1. Chitraa, V, A. Selvadoss Thanamani., "A novel technique for sessions identification in web usage mining preprocessing", International Journal of Computer Applications 34.9, pp.23-27, (2011):
2. Bina Kotiyal, Ankit Kumar, Bhaskar Pant and R. H. Goudar., "Classification Technique for Improving User Access on Web Log Data", Intelligent Computing, Networking, and Informatics, Advances in Intelligent Systems and Computing 243, DOI: 10.1007/978-81-322-1665-0\_111, \_ Springer India 2014.
3. Jagriti Chand , Abhishek Singh Chauhan, and Ashish Kumar Shrivastava., "Review on Classification of Web Log Data using CART Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 80 – No 17, October 2013.
4. Adeniyi, D. A., Z. Wei, Y. Yongquan., "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics 12.1, pp. 90-108, 2016.
5. 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.
6. ERRITALI, Mohammed, and Hanane EZZIKOURI., "Pretreatment of web log files", Journal of Information Sciences and Computing Technologies 2.1, pp.108-121, 2015.



ISSN(Online): 2320-9801  
ISSN (Print) : 2320-9798

# International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 7, July 2016**

7. Ramya, C., G. Kavitha, and Dr KS Shreedhara., "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process", arXiv preprint arXiv:1105.0350, 2011.
8. Sait, Abdul Rahaman Wahab, and Dr T. Meyappan., "Data Preprocessing and Transformation Technique to Generate Pattern from the Web Log", International conference on Computer Science and Information Systems (ICSIS'2014), Oct. 2014.

## **BIOGRAPHY**

V. Vidyapriya, M.sc., Mphil., is an Associate Professor in PG and Research Department of computer science, Quaid-E-Millath Govt College for women (Autonomous), Anna Salai, Chennai-02. Her research interest focuses on web mining.

V. Pushpa is a M.Phil research scholar in PG and Research Department of computer science, Quaid-E-Millath Govt College for women (Autonomous), Anna Salai, Chennai-02. she received her Master degree in April 2015.