User Identification in Web Usage Mining using Data Mining Techniques with Nominal Distance Function and K-D Tree

G.Vijaiprabhu[1], Dr.K.Meenakshisundaram[2]
Ph.D. Research Scholar [1], Associate Professor[2]
Department of Computer Science[1,2]
Erode Arts and Science College, Erode[1,2]
[1]gvprabhu7@gmail.com, lecturerkms@yahoo.com[2]

***Abstract*- Web mining aims at deriving actionable knowledge from web. The web data is collected from Web servers, clients, proxy servers, or server databases. Web mining use web information with the help of data mining techniques to extract useful patterns from the web. It is divided in three main key domains. The content of web pages are analyzedby content mining, web access log areanalyzed by web usage mining and the link of the web pages are analyzed by structure mining. Among this, usage mining is widely used to discover the usage patterns from web log files. This research paper primary focuseson web usage mining to identify or classify users and status codeby using classification technique in data mining. Each user is identified and classified according to IP address specified in the log file and the status code is verified to improve the site. For classification, KNN with nominal distance function with K-D tree search algorithm is used instead of using traditional Euclidean function. This proposed work classifies the user and status code to identify the user and verify the response code during browsing. The work is implemented in Rapid miner tool and assessed with suitable evaluation metrics.**

***Keywords* – KNN classification, Nominal distance function, K-Dimensional tree.**

I. INTRODUCTION

Mining the data discovers information by [3] analyze massive set with various perceptions and extracting useful information via procedures. This is the most motivated research area to find out variant patterns. The main goal is to discover the knowledge hidden in data. Due to enormous growth of data, mining is at drastic level in each field. Getting the right information from data is the most challenging task. Many academicians and industry researchers are engaged on the process of knowledge mining due to abundance of data. It is the core step of Knowledge discovery procedure. The recent aspects and development promotes the rapid growth of KDD and DM [8]. Web mining is the application of data mining to reveal patterns from the World Wide Web. It is the largest data base, growing in unsystematic way. The pages are linked each other, but are not organized logically. During some course of period millions of web pages are added to web and undergo changes daily. This leads to information overloading. So in this situation, getting desired information or particular details is burden. Therefore a very efficient and effective technique is needed to extract or access the required information. The main challenge behind is the extraction of information with less attempt and time. Another major issue is the relevancy of the information. Proper management of data improves the retrieval efficiency. Web Mining is the widely accepted method for this. To satisfy the requirements of web crawlers one of the most used functional techniques in web mining is to analyze the user browsing patterns through web usage mining. The three broad areas in web mining are: Web content, Web Structure and Web Usage mining.

A. Web usage mining

The process of finding patterns and information from [4] server logs to have the idea on the user activity including where the users are from, how many users clicked on which site and the types of activities being done.

B. Web Content Mining

The extraction of desired information from the unstructured raw data is referred to as Web content mining. A set of information extraction tool is used to identify and collect content. It is the process of extraction of information from web document that may be in any of the format video, audio, text or structured records.

C. Web structure mining

It includes the process of assessing the nodes and structure of a site through the use of graph theory. There are two concerns that can be obtained here. One is the structure of a website and how it is connected to other sites and the document structure of the website that how each page is connected.

All the above categories come under web personalization which consists of five modules user profiling, Log analysis, Content management, website publishing and information acquisition and searching.

II. WEB USAGE MINING

Web usage mining refers to the automatic detection and analysis of patterns in blog data and it relatesthe data collected and [9] generated as a result of learners interactions with resources like websites and blogs. The aim is to capture, model, and assess the behavioral patterns and profiles of the users on the basis of interacting with the website. The revealed patterns are finally represented as collections of pages, objects and resources with frequently accessed groups of users with common desires.

Phases in Web Usage mining

A. Data Collection

A Web log files records information when a [15] user submits a request to a server. It is a text file which is created automatically, when a user requests a page. It is a file on which the server records information every time a user requests a resource from a site. When a user sends request to the server, the databases will be retrieved. At the same time , the user session including the URL, Client's IP address, accessing date and time , query stem will be recorded in the logs. A log file resides at three different areas such as web server log, proxy server log and client browser log.

B. Preprocessing

It integrates the databases and makesraw information into understandable and consistent format. The information stored in web logs is processed as it has insufficient and noisy data. It is done in early step by removing redundancy, useless, error, incomplete, inconsistency. There are many e-sources and web usage mining analyze data logs, site address, login information, access logs, cache, cookies etc. It includes methods like cleaning and User, session identification.

C. Pattern Discovery

It is the key component in web usage. It includes the algorithms and procedures from data mining, machine learning and pattern recognition. A variety of methods are used to find hidden data information on Web server logs. It includes methods like association rules, statistical classification clustering and sequential patterns.

D. Pattern Analysis

In this stage, repeated patterns are eliminated and relevant and meaningful patterns are found using Structured Query Language knowledge query mechanism and On-line Analytical Processing a multi-dimensional data cube, Usability analysis a modeling technique to accessing the behavior of user on the web site.

III. LITERATURE REVIEW

Adeniyi*et al* [1] present a study of automatic usage and recommendation system based on current user behavior through click stream data with the newly developed 'Really Simple Syndication' reader website, in order to provide relevant information to the individual without explicitly asking for it. The K-Nearest-Neighbor classification method is trained to use in on-line and Real-Time to identify clients/visitors click stream data at a particular time. To achieve this, the file was extracted, cleansed, formatted and grouped into meaningful session. The result shows that the classifier is transparent, consistent and have high tendency to possess desirable qualities and easy to implement.

AnandanBellie*etal*[2] analyses behavior of the university students to improve the quality of service of the internet. The quality of the internet is to addressed as the daily usage of net is increased and this can be solved while the existing user behavior pattern is known. This paper takes the web log data form th e university computer lab and employs K_-Means algorithm in Weka tool for assessing similar characteristics students while accessing the web. This analysis is helpful for the management of university internet infrastructure and web page personalization.

Jyothi*et al* [5] design a recommendation system and automatic web usage data mining which is based on current user behavior with click stream data. The K-Nearest-Neighbor algorithm is trained on-line to identify clients and visitors click stream data that meets the needs of the specific user at the particular time. This paper put forth a hybridization of traditional KNN and ANN which leads to the improvement in accuracy. In traditional KNN only distance between given two users is calculated. Here, newest accuracy of KNN is calculated by finding out the distances within all the users. This work is carried out in MATLAB and the results shows that hybridization of KNN and ANN gives more accuracy than when it is evaluated separately.

Kaviarasan*et al* [6] propose a method that combines usage patterns extracted from server logs with detailed semantic data that characterizes the content of the corresponding pages. this work develops a method to extract and analyze frequent semantic navigation patterns which are fed into a recommendation engine is proposed. By integrating usage and Web pages' detailed semantic information in the personalization process there is a possibility to increase the recommendation accuracy. The proposed method combines two research areas Semantic Web and Usage Mining. The research conducted an extensive experimental evaluation that confirms the recommendation accuracy increases with the integration of semantic and usage data. The results show that the proposed method achieves better accuracy than a usage based model.

ManishaKumari*et al* [7] did a review with classification technique nearest neighbor in web mining. It is applied on the log details and the performance of the algorithm is measured. This algorithm is biased by the value of fixing K. The only approach available is to run the algorithm for different K and choose the best one. There is no separate training phase even though the sheet is split into training and test. All the computations are done in the test phase. The review of this work can be implemented in discovering user navigation pattern, recommendation system, Computer behavior studies and content improvement.

Vedpriya*et al* [10] proposed a new model for predicting the user behavior from the data accessed from web log server, proxy server and client-side cache. K-means clustering and Regression Analysis algorithms are used to predict the future accessing of web pages. Using the web log files and user's current navigation pattern, the proposed system predicts next web files to the user in form of recommendation list. K-means clustering is used to assign weights to the accessed pages and regression analysis is used for prediction. Finally, the performance is evaluated in terms of accuracy, error rate, time and space complexity.

Vidyapriya*et al* [11] proposed a method to identify web users from weblogs using supervised learning method. This paper intends pattern discovery using various classification techniques to determine highest accuracy and lowest error rate using Naïve Bayesian, CART, k-nearest neighbour. The primary objectives of this paper is to identify the interest of user access pattern from the weblogs defining specific website. This paper deals with data preprocessing and pattern discovery. This dataset is taken from the specific website with 4193 raw log entries. From the results it is revealed, the k- nearest neighbour displays the highest accuracy and lowesterror rate in the weka tool.

IV. METHODOLOGY

A. Existing Methodology (KNN)

Existing method uses K-Nearest neighbor algorithm to classify the users. The algorithm has a fixed number of neighbors to vote in the process of classification for an instance which is identified by 'k', where 'k' is a positive integer. The 'K' is fixed by the user. There is no standard rule used for initialization. The only way is to initialize randomly or run the algorithm for variant 'K' valuation and pick the most suited one with high accuracy. The algorithm is a non-parametric lazy learning method that doesn't require any prior knowledge regarding classification. It yields the closest records from training samples which have highest priority and can be used both in discrete and continuous data. The procedure of classification starts with a set that contains certain number of attributes. The series is divided into dual as training and testing. Training is given as input to the algorithm while testing is used to assess it. The division can be done using various methodology like 'Random sampling, Percentage split, Cross validation, Hold-out method' etc.Fig 1.shows the nearest neighbor with k=2.

Distance function used in KNN

Euclidean Distance Function - The length [14] of the line segment (c_1, d_1) and (c_2, d_2)(two dimension) is calculated by the equation 1. It can be implemented for one dimension to n-dimension.It can also be known as L2 norm.

$$E = \text{SQRT}((c_2 - c_1)^2 + (d_2 - d_1)^2) \quad \text{--- (1)}$$

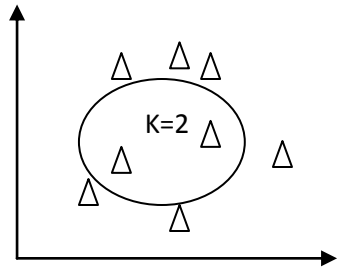

Fig. 1 KNN with 2 nearest neighbor

Search algorithm used in KNN for sorting and searching of neighbor

Linear search – It is a sequential search algorithm to find an element in the array. It searches the element one by one until the element is found or the whole list is searched. It makes only two comparisons in each iteration. Hence it takes high processing time. Also it makes more comparisons to search an element.

Procedure KNN

Step 1: Determine parameter 'K' randomly.
Step 2: Calculate the difference among the instance and all the training samples by Eq. (1).
Step 3: Sort the range and determine the closest point on the basis of K^{th} minimum length using linear search algorithm.
Step 4: Select the category or vote of the nearest neighbor.
Step 5: Return the mode of 'K' labels.

Advantages

The KNN algorithm is easy to implement and interpret the result. It is applicable to classification and regression problem and able to handle multi class problem.

Disadvantages

Euclidean norm leads to high computation with squared calculation which leads to less accuracy while classifying the lengthy data like IP address and linear search algorithm takes high processing time as it search the element one by one.

B. Proposed Methodology (EKNN)

The Enhanced KNN (EKNN) method introduce a nominal distance function to calculate the difference between two points and use K-dimensional tree for search algorithm in order to improve the accuracy and minimize the processing time.

Proposed distance function used in KNN

Nominal distance function –Hamming distance is used as a nominal distance function. It takes two points with equal length. It gives the number of bit positions in which the two bits are different. The Hamming distance between two strings, c and d is denoted as H_d (c, d). It is used in several applications including information theory, coding theory and cryptograpy. The data are classified with minimum hamming distance.

Hamming distance of A and B = H_d(A XOR B) ---(2)

Example

1. Input: n1 = 9, n2 = 14
   9 = 1001, 14 = 1110

H_d (n1, n2) = 3

2. Input: n1 = 4, n2 = 8
   4 = 0100, 8 = 1000
   H_d (n1, n2) = 2

3. Input: n1 = Kathir, n2 = Karthi
   H_d(n1, n2) = 4

Proposed Search algorithm used in KNN

K-Dimensional Tree - A K-D Treeis a space partitioning data structure for organizing points or data in a K-Dimensional space. The non-leaf node in K-D tree divides the space into two parts, called as half-spaces. Points in the left space represents left subtree and points in the right represents right subtree. In a 2-D Tree, the root have an x-aligned plane, the root's children have y-aligned planes, the root's grandchildren have x-aligned planes, and the root's great-grandchildren have y-aligned planes and so on. Points are aligned by selecting the median of the values with heap or merge sort to have a balanced k-d tree.

Determination of a point position whether it is left right subtree

If the root node is aligned in plane A, then the left subtree will contain the smaller value than the root node and the right subtree will contain all points that are greater-equal to that of root node. An example is shown in Fig 2.

Example K-d tree

Points: (2, 3), (9, 6), (4, 7), (5, 4), (7, 2), (8, 1)

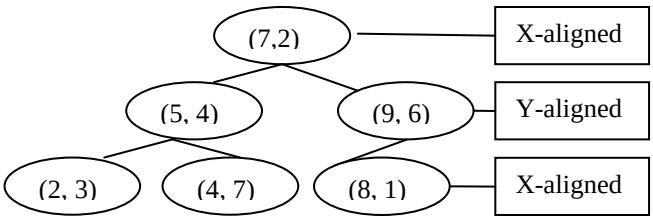

Fig.2 K-D Tree

Procedure EKNN

Step 1: Determine parameter K.
Step 2: Calculate the difference among the instance and all the training samples by Eq. (2).
Step 3: Sort the range and determine the closest point on the basis of K^{th} minimum length using K Dimensional tree algorithm.
Step 4: Select the category or vote of the nearest neighbor.
Step 5: Return the mode of 'K' labels.

Advantage

- Reduce the searching time with less comparison.
- Analyze multidimensional dataset effectively.

V. RESULT AND DISSCUSSION

The Dataset is taken from Kaggle[13] repository with four fields User ID, Login Time, User Communication and Status code. The Classification is done on the nominal attribute User Id and Status code.This dataset has the file format of "Common Log Format". [12] The Standardized text format is also known as "NCSA Common Log Format" mostly used by the web servers while generating server log files.

*Common Log Format:*Host Identifier, Date, Request, Status, Bytes.

Example: 128.2.0.1 user-identifier frank, [11/Nov/2010:11:45:26 -0900], "GET /apache_pb.gif HTTP/1.0", 300, 2346.

A. User Identification

Table I list out the performance of existing KNN and proposed EKNN with five measures for User identification in which Accuracy, Error rate, Recall and Precision are in percentages and Processing Time in seconds and it is shown in Fig 3 and Fig 4.

Table I. User Identification

Classified User/IP	Parameters	Proposed Method EKNN for User Identification	
		Existing KNN	Proposed EKNN
10.128.2.1	Accuracy	89.01	94.52
10.129.2.1	Error rate	10.99	5.48
10.130.2.1	Recall	88.72	93.85
10.131.0.1	Precision	89.15	95.02
10.131.2.1	Processing Time in seconds	14.91	9.56

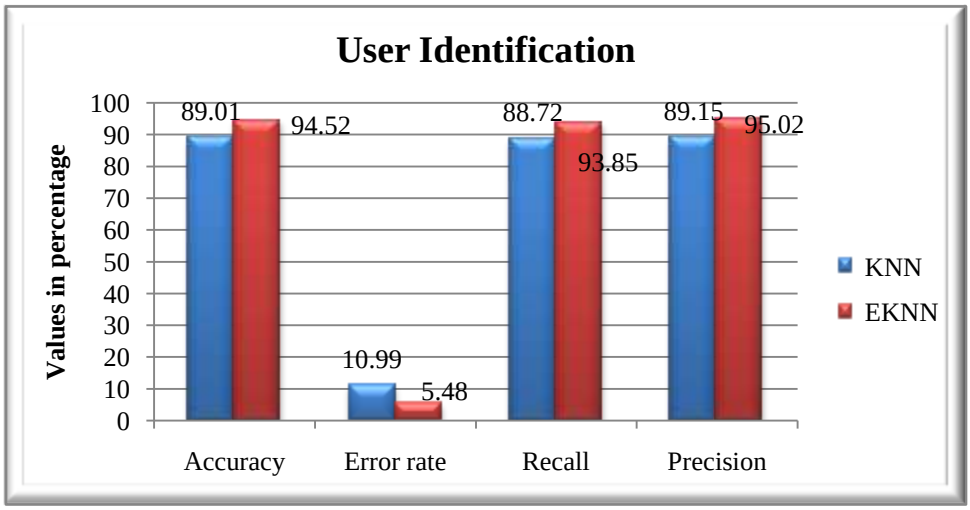

Fig 3. User identification for Accuracy, Error rate, Recall, Precision

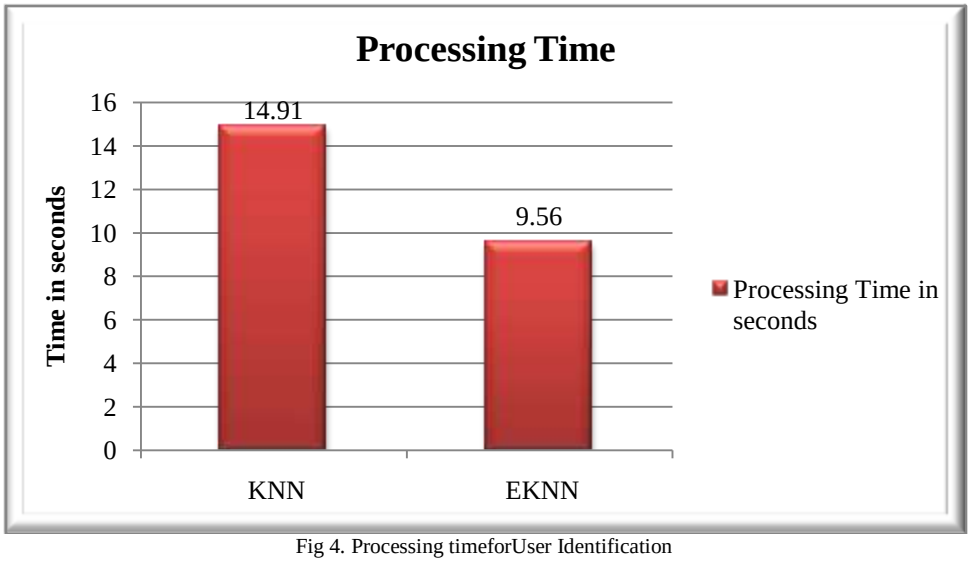

Fig 4. Processing timeforUser Identification

The observation in Fig. 4, 5 shows the proposed EKNN gives better results than KNN in terms of accuracy and recall and precision. The error rate and processing time is also reduced in EKNN for User identification.

B. Status code identification

Table II list out the performance of existing KNN and proposed EKNN with five measures for Status code identification in which Accuracy, Error rate, Recall and Precision are in percentages and Processing Time in seconds and it is shown in Fig 5 and Fig 6.

Table II. Status Code

Classified Status code	Parameters	Proposed method EKNN for Status code Identification	
		Existing KNN	Proposed EKNN
200 206 302 304 404	Accuracy	90.45	96.32
	Error rate	9.55	3.68
	Recall	89.72	95.52
	Precision	90.97	96.85
	Processing Time in seconds	12.03	6.21

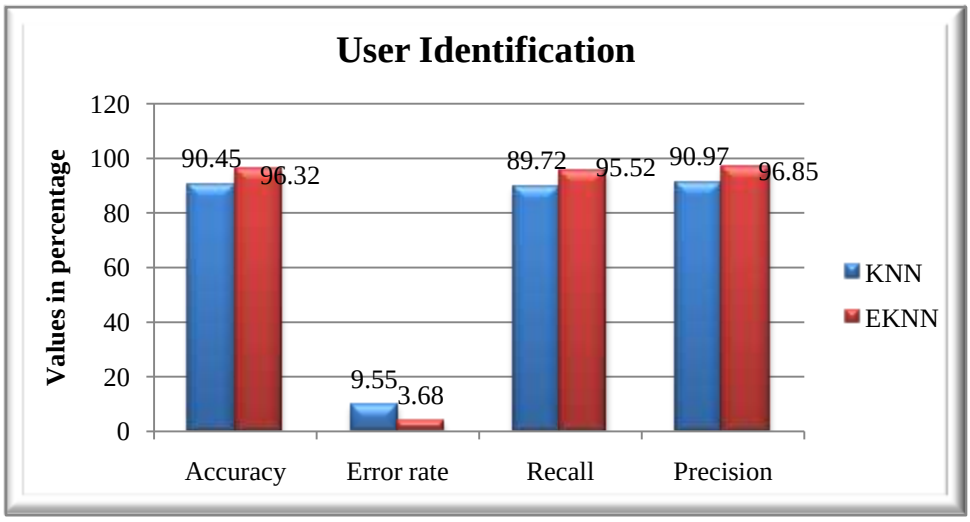

Fig 5. Status code identification for Accuracy, Error rate, Recall, Precision

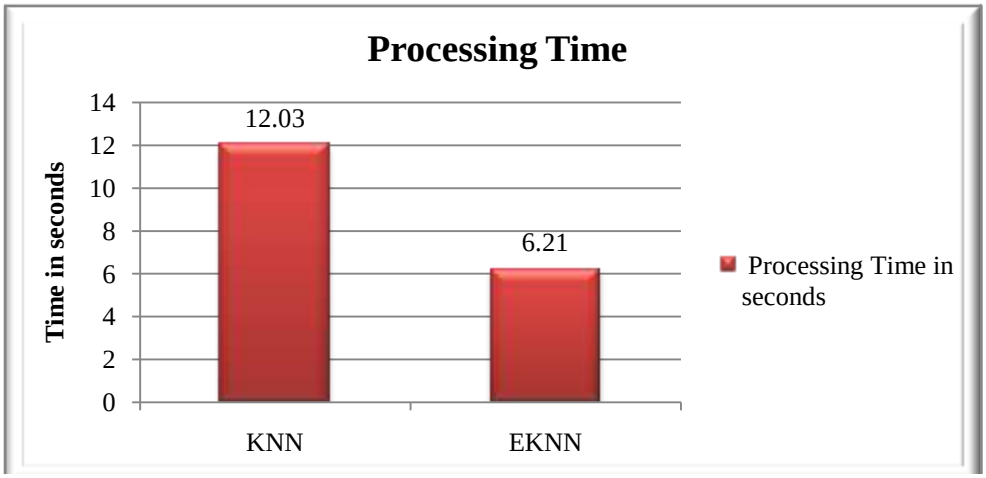

Fig 6. Processing timeforStatus codeIdentification

The observation in Fig. 4, 5 shows the proposed EKNN gives better results than KNN in terms of accuracy and recall and precision. The error rate and processing time is also reduced in EKNN for Status code verification.

VI. CONCLUSION

Web mining discovers knowledge from web data which comprise of web documents, hyperlinks, usage logs, etc by using data mining techniques. This research focus on Web usage mining to discover interesting usage patterns from web data, in order to understand and to better serve the needs of the user with web-based applications. The supervised learning technique K-Nearest Neighbor (KNN) is applied with nominal distance function and K-Dimensional tree search algorithm instead of the traditional Euclidean distance and linear search algorithm for user identification and status code verification. Thisenhancement in EKNN improves the accuracy and reduces the processing time.

In future, the work may be extended by implementing other search algorithms and distance function to further improve the accuracy and minimize the processing time.

REFERENCES

[1] Adeniyi D.A, Wei .Z, Yongquan .Y, " *Automatedweb usage data mining and recommendation system using K-Nearest Neighbor classification method*", Applied Computing and Informatics, 2016.

[2] AnandanBellie, "*Web Usage Analysis of University Students to Improve the Quality of Internet Service*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4Issue 5, May 2015.

[3] Bing Liu, "*Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*", second edition, Springer.

[4] JoshilaGrace .L.K, Maheswari .V ,DhinaharanNagamalai, " *Analysis of Web Logs and Web User in WebMining*", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.

[5] Jyoti, Jagdeepkaur, "*Recommendation System with Automated Web Usage Data Mining by Using k-NearestNeighbour (KNN) Classification and Artificial Neural Network (ANN) Algorithm*", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 5 Issue VIII, August 2017.

[6] Kaviarasan .. , Hemapriya .K , Gopinath .K, " *Semantic Web Usage Mining Techniques for predicting Users'Navigation Requests*", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 5, May 2015.

[7] ManishaKumari, SaritaSoni, "*A Review of classification in WebUsageMining using K- Nearest Neighbor*",Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 5, 2017.

[8] Margaret H. Dunham, Sridhar .S, " *Data Mining: Introductory and Advanced Topics*", Pearson Education.

[9] SahajChavda, Saurabh Jain, NikunjPanchal, Manisha Valera, " *Recent Trends and Novel Approaches in WebUsage Mining*", International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 04April -2017.

[10] VedpriyaDongre,JagdishRaikwal,"*An Improved User Browsing Behavior Prediction using RegressionAnalysis on Web Logs*", International journal of computerapplications, Volume 120 – No.19, June 2015.

[11] Vidyapriya V., Pushpa .V, "*Identifying Web Users from Weblogs Using Classification Algorithms*", International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 7,July 2016.

[12] en.wikipedia.org/wiki/Common_Log_Format

[13] www.kaggle.com/shawon10/web-log-dataset.

[14] en.wikipedia.org/wiki/K_nearest_neighbors_ algorithm.

[15] en.wikipedia.org/wiki/Web_mining.