Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 1405-1416
© Research India Publications
http://www.ripublication.com

# A Review of classification in Web Usage Mining using K- Nearest Neighbour

**Manisha Kumari**
*Research Scholar Computer Science, BBAU University,
Lucknow, Uttar Pradesh, India.*

**Sarita Soni**
*Assistant Professor Computer Science
BBAU University, Lucknow, Uttar Pradesh, India.*

## Abstract

Web Usage Mining came into account due to proliferation of information sources available on internet. With the help of Web Usage Mining we are able to extract useful information from the corpus. Classification is widely used method in pattern discovery phase of Web Usage Mining. Classification is a supervised learning; it can be used in designing of models describing data classes, where attribute class is required in the construction of classifier. Classification is useful in categorizing the object that has similar functionality. Classification technique uses many algorithms as classifier among them Nearest Neighbour (KNN) method is a very simple, most popular, highly efficient and effective algorithm. The output of K-NN classifier is a class membership. An object is classified on the basis of number of vote of its neighbours. The object is being assigned to the class most common among its k- nearest neighbour. This article provides a review of Web Usage Mining systems, its classification technique and widely used KNN classifier. We will also emphasize on advantages and disadvantages of K-NN algorithm.

**Keywords:** Web usage mining, K- Nearest neighbour, Pattern discovery, Classification.

## I. INTRODUCTION

In recent years, the advance in computer and web technologies and the decrease in their cost have expanded the means available to collect and store data. As an intermediate consequence, the amount of information (Meaningful data) stored has been increasing at a very fast pace. Web mining is a process that uses many data analysis tools to discover patterns and relationships in data that we are using to make a valid prediction. Web mining can be broadly defined as a discovery and analysis of useful information from the World Wide Web. Web usage mining is the process of extracting useful information from server web logs. Information regarding interested web users provides valuable information to web designer for quickly respond to their individual needs. The information gathered through the web is further evaluated by traditional data mining techniques such as clustering, classification and association.

Classification algorithm can be used to classify interested users. Classification is simply trying to group records in a class. It is actually a predictive task. K- Nearest Neighbour is one of the most popular algorithms used as classifier. KNN algorithm is a type of instance – based learning or lazy learning algorithm where the function is only approximated locally and all computation is deferred until classification. An object is classified by a majority vote of its neighbours. For example If it walks like a duck, quacks like a duck and looks like a duck, then it’s probably a duck. The k-nearest neighbor classifier is capable of producing useful and good classification and recommendation to the client choices with optimal values of k.

## II. WEB MINING IN BRIEF

Based on the different emphasis and different ways to obtain information, web mining can be divided into three major parts: web Content mining, Web Structure mining and Web Usage mining. Web contents mining can be described as the automatic search and retrieval of information and resources available from millions of sites and online databases through search engines/ web spiders. Web structure mining operates on the Web’s hyperlink structure.

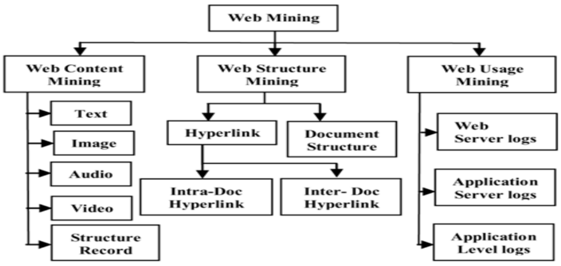

**Fig 1:** Web mining taxonomy

### A. Web Usage Mining;-

Web usage mining is an application of data mining techniques to discover interesting usage pattern from web data, in order to understand and better serve the need of Web - based applications. It can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular web site.

## B. How to perform Web Usage mining:-

Web Usage Mining consists of three phases pre-processing, pattern discovery and pattern analysis. Web Usage Mining is achieved by discovering the secondary data derived from the interaction of the users while surfing on the web. It gathers useful usage data thoroughly, filter out irrelevant usages data, establish the actual usages data, discover interesting navigation patterns and display the navigation patterns clearly. The secondary data includes the data from the proxy server logs, browser logs, web server logs, user profiles, user sessions, user queries, registration data, bookmark data, mouse clicks and scrolls, cookies and any other data which are the results of these interactions. The log files can help us answer questions such as “from what search engine are visitors coming? What pages are the most and least popular? Which browsers and operating systems are most commonly used by visitors? Visitor’s behaviour are analysed and interpreted through some data mining techniques are association rules, path analysis, sequential analysis, clustering and classification.

The main task of web usage data is to capture web browsing behaviour 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. As our work concentrates on web usage mining, it is the application of data mining techniques to discover usage patterns from web data. Data is usually collected from user's interaction with the web, like web/proxy server logs. Usage mining tools discover and predict user behaviour, in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service.

The major problem with Web Usage Mining is the nature of the data they deal with. With the growth of internet, Web Data has become huge in nature and a lot of transactions are taking place in seconds. Apart from the volume of data, the data is not completely structured. It is in a semi structured format so that it needs a lot of pre-processing before the actual extraction of the required information.

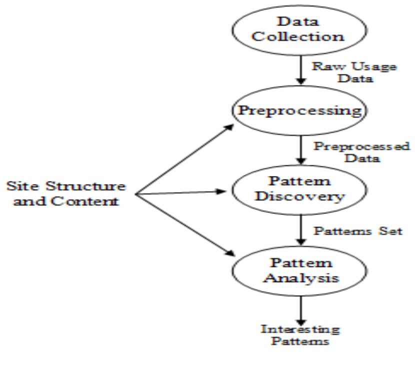

**Fig2:** Web usage mining phases

### B.1 Data Collection:-

The web log files on the web server are primary source of data for Web Usage Mining. Data can be collected from the following three locations.

- Web Servers
- Web proxy servers
- Client browsers

Web servers can be configured to write different fields into the log file in different formats. The most common field used by web servers are the followings: IP Address, Login Name, User Name, Request Type, Status, Bytes transferred, Referrer, Visiting Path, Path traversed, Timestamp, page last visited, success rate, User agent, URL etc.

### B.2 Data pre-processing:-

The data collected from web server log is often incomplete and create uncertainty. Pre-processing is important phase  web usage mining process in order to clean, correct and complete input data and to mine the knowledge effectively . Pre-processing phase takes 80% time of whole process.

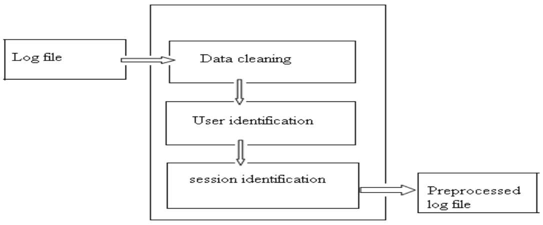

**Fig 3:** Web log pre-processing

1. **Data Cleaning:** In the process of data cleaning, elimination of irrelevant items can be accomplished by removing unwanted view of images, graphics, multi media etc. This process minimizes the file to a great extent.

**2. User Identification and Session Generation**: After data cleaning, unique users must be identified. User identification is done by (i) User login information. (ii) to use cookies for identifying the visitors of a web-site by storing a unique ID. (iii) The same IP but different user agent means a new user. The user agents are said to be different if it represents different web browsers or operating systems in terms of type and version. Identification of the user sessions is very important because it largely affects the quality of pattern discovery result. Session is time duration spent on the web pages. It is using timestamp details of web pages.

**3. Data Conversion:** It is conversion of the log file data into the format needed using mining algorithms.

### B.3 Pattern Discovery:-

Discovery of desired patterns and to extract understandable knowledge from pre-processing data is a difficult task. Here we will briefly describe some techniques to discover patterns from processed data.

**1. Association Rules:** Association rule generation can be used to relate pages that are most often referenced together in a single session. It can discover the correlations between pages that are most often referenced together in a single server session/user session.

**2. Sequential Patterns:** By using this approach, Web marketers can predict the future visiting patterns which will be placing advertisements aimed at certain user groups. The disadvantage of it is difficult to find out the interesting patterns if huge amount of data is present.

**3. Clustering:** Clustering is a technique which groups a set of items together having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters can be discovered *i.e.* usage clusters and page clusters. Clustering of users refers establish groups of users exhibiting similar browsing patterns. On the other hand, clustering of pages will discover groups of pages having related content. This information is useful for Internet search engines and Web assistance providers.

**4. Classification:** Classification is done to identify the characteristics that indicate the group to which each case belongs. This pattern can be used in both the cases i.e. to understand the existing data and to predict how new instances will behave. A classification technique follows three approaches statistical, machine learning and neural network.

**B.4 Pattern Analysis:-** The final step of the entire Web Usage Mining process is Pattern Analysis.

The objective of this procedure is to select the interesting patterns and filter out uninteresting patterns. The patterns are analysed using techniques such as OLAP techniques, Data and Knowledge Querying and Usability analysis. In OLAP techniques, the results of pattern discovery into data cube after that OLAP operation are performed. Data and Knowledge Querying methods use a tool like SQL. After this, visualization techniques such as graphing patterns or assigning colours to different values are used to highlight overall patterns.

## III. CLASSIFICATION ALGORITHMS IN PATTERN DISCOVERTY PHASE:-

The need and requirement of the user's of the websites to analyze the user preference become essential due to massive use of internet. Classification techniques are to be applied on the web log data and the performance of these algorithms can be measured.

There are many algorithms used as a classifier, In this section we are comparing which are mostly used in pattern discovery phase of Web Usage Mining.

- ID3
- C4.5
- Decision Tree Classifier
- Naive Bayes Classifier
- Support Vector Machine
- Rule Based
- K- Nearest Neighbour
- Artificial Neural Network

| S. No. | Algorithms | Advantages | Disadvantages |
|---|---|---|---|
| 1. | C4.5 | 1). It takes less time to build model.<br>2). Produces accurate result.<br>3). It has short search time.<br>4). Can use both discrete and continuous values<br>5). Deals with noise | 1). Empty braches<br>2). Insignificant branches<br>3). Over fitting<br>4).Does not work with small training data |
| 2. | ID3 | 1). Produces more accurate result than C4.5<br>2). Uses nominal attributes for classification with no missing values<br>3). Detection rate is increase and space consumption is reduced | 1). It has long searching time<br>2). Takes more memory than C4.5<br>3). Sometimes it may generate very long rules which are very hard to prune |
| 3. | Support Vector Machine | 1). Produces very accurate results<br>2). Less over fitting, robust to noise<br>3). Good at text classification when high dimensional data to be classified | 1). Only good for binary classification<br>2). High complexity and extensive memory requirements for classification in many cases<br>3). Speed and size requirement both in training and testing is more |

| | | | |
|---|---|---|---|
| 4. | Naive Bayes | 1).Simple to implement<br>2).Great computational efficiency and classification rate<br>3). It predict accurate results for most of the classification and prediction problems | 1). Requires large number of records to obtain good results.<br>2). It stores all the training samples<br>3).The precision of algorithm decreases if the amount of data is less |
| 5. | K- Nearest neighbour | 1).Easy to implement<br>2). Suitable for multi model classes<br>3). Training is very fast<br>4). Robust to noisy training data<br>5). Zero cost of the learning process | 1).Time to find the nearest neighbour in a large training data set can be excessive<br>2).It is sensitive to noisy or irrelevant attributes<br>3). Performance of algorithm depends on the number of dimensions used<br>4).Memory limitation |
| 6. | Artificial neural networks algorithms | 1). It is easy to use, with few parameters to adjust<br>2). A neural network learns and reprogramming is not needed<br>3). Easy to implement<br>4). Applicable to a wide range of problems in real life | 1).Requires high processing time if neural network is large<br>2). Difficult to know how many neurons and layers are necessary<br>3). Learning can be slow |

We briefly discussed about classification algorithms and we see it is not easy to justify which is better. Most analytical problem involves making a decision. A classification technique helps in the analysis which is most insightful and directly links to an implementation of a roadmap about the user/customer. K-nearest neighbour is widely used in classification problems in the industry. To evaluate any technique we commonly look for the following important aspects:

1. Easy to interpret output
2. Calculation time
3. Predictive power

KNN algorithm fairs across all parameters of considerations. It is commonly used for its low calculation time and ease to interpret output.

## IV. K- NEAREST NEIGHBOUR CLASSIFIER

### A. Basics

According to the nearest neighbour technique the classification of an unknown data tuple is accomplished by analysing the classes of its nearest neighbours. KNN algorithm employs this principle of nearest neighbour technique. But in case of KNN algorithm a fixed number of nearest neighbours are allowed to vote in the process of classification of an unknown data tuple which is identified by k, where k is a positive integer. When k=1 then the unknown data tuple is classified as the class of the training data tuple which is most nearest to it. The k-nearest neighbour algorithm (KNN) is an intuitive yet effective machine learning method for solving conventional classification problems. KNN is non parametric lazy learning algorithm. KNN algorithm does not require any prior knowledge regarding data set for classification. It performs classification purely on similarity basis. It is considered as a lazy learning algorithm because it does not build a model or function previously, but yields the closest k records of the training data set that have the highest similarity to the test. KNN can be used both in discrete and continuous decision making known as classification and regression respectively.

The procedure of classification in KNN starts with a data set. The data set is constituted of certain number of attributes that define a data set. The data set is divided into two sets: training set and test set. Training set is given as input to the algorithm while test set is used to . The division of the data set can be done using various methods such as hold-out method, random sampling, cross validation etc. KNN classifies any new tuple by using training data tuples similar to it. Due to this KNN is also called local learner. There is no explicit training phase in KNN. It stores all the training tuples given to it as input without doing anything. All the computations are done at the time of classification of a test tuple. In KNN algorithm, the training tuples can be viewed as a set of data points in an n-dimensional space, where n dimensions are the set of n attributes describing the data set. When an unknown tuple comes for classification, we have to find out the k most nearest data points to it in the n dimensional space. To find the k most nearest data points to the unknown tuple various distance metrics are used for example Euclidean distance, Minkowski distance, Manhattan distance.

### B. Distance used in KNN

The three famous distance functions used with KNN are (i) Euclidean Distance : D(x, y) = $((\sum_{i=1}^{m}|x_{i-}y_i|)^2)^{1/2}$

(ii) Manhattan Distance: D(x, y) = $\sum|x_{i-}y_i|$

(iii) Minkowski Distance: D(x, y) = $((\sum_{i=1}^{m}|x_{i-}y_i|)^r)^{1/2}$

### C. Mathematical model of KNN

We present a mathematical model for knn algorithm and show that knn only makes use of local prior probabilities for classification.

For a given query instance $x_t$, knn algorithm works as follows:

$$y_t = \underset{c\in\{c1,c2,\dots cm\}}{\arg max} \sum_{x_i \in N(x_t,k)} E(y_i, c) \qquad (1.1)$$

Where $y_t$ is the predicted class for the query instance $x_t$ and m is the number of classes present in the data. Also,

$$\mathrm{E(a,b)} = \begin{cases} 1 & \text{if a = b} \\ 0 & \text{else} \end{cases} \qquad (1.2)$$

N(x, k) = Set of k nearest neighbour of x.

Eq. 1.1 can be written as: $y_t =$

$$\arg\max\{\sum_{x_i\in N(x_i,k)} E(y_i\ ,c1), \sum_{x_i\in N(x_i,k)} E(y_i\ ,c2), \dots.\sum_{x_i\in N(x_i,k)} E(y_i\ ,cm) \qquad (1.3)$$

$$y_t = \arg\max\{\sum_{x_i\in N(x_i,k)} E\frac{(y_i\ ,c1)}{k}, \sum_{x_i\in N(x_i,k)}, E\frac{(y_i\ ,c2)}{k}, \dots\dots\dots., E\frac{(y_i\ ,cm)}{k}\} \qquad (1.4)$$

And we know that

$$p(c_j)_{(x_t,k)} = \sum_{x_i\in N(x_t,k)} \frac{E(y_i,c_j)}{k} \qquad (1.5)$$

Where $p(c_j)_{(x_t,k)}$ is the probability of of occurrence of jth class in the neighbourhood of $x_t$. Hence eq. 1.4 turns out to be

$$y_t = \arg\max\{p(c_1)_{(x_t,k)}, p(c_2)_{(x_t,k)}, \dots\dots..p(c_m)_{(x_t,k)} \qquad (1.6)$$

It is clear from Eq. 1.6, that KNN algorithm uses only prior probabilities to calculate the class of the query instance. It ignores the class distribution around the neighbourhood of query point.

### D. Algorithm

Input Parameters: Data set, k

Output: Class membership

Step 1: Store all the training tuples.

Step 2: For each unseen tuple which is to be classified

A Compute distance of it with all the training tuples using Euclidean Distance.

B. Find the k nearest training tuples to the unseen tuple.

C. Assign the class which is most common in the k nearest training tuples to the unseen tuple.

End for

### E. How to choose value of K in KNN

However, to apply KNN, we need to choose an appropriate value for k, and the success of classification is much dependent on this value. Thus, the KNN method is biased by k. There are many ways of choosing the value of k , but a simple one is to run the algorithm many times with different k values and choose the one with the best performance or the proper choice of k depends on data set.

### F. Advantages and Drawbacks

***Advantages***

KNN has many main advantages : simplicity, effectiveness, intuitiveness and competitive classification performance in many domains. It is robust to noisy training data and is effective if the training data is large.

***Drawbacks***

It suffers from two major drawbacks:

1. It uses only *local* prior probabilities to predict instance labels, and hence does not take into account, class distribution around the neighbourhood of query instance. This results in undesirable performance on imbalanced data sets. The performance of kNN algorithm over imbalanced data sets can be improved, if it uses this information while classifying instances.

2. It is a lazy learner i.e. it uses *all* the training data at the runtime, and is hence slow.

## V. CONCLUSION

From the last decade, World Wide Web has become one of the prominent media. Web is becoming the necessity of the business and organizations into many aspects. WUM can be used to discover interesting user navigation pattern, which can be applied to real world problems such as recommendation system, User/Computer behaviour studies, Web site/ page improvement etc. This paper has provided a survey about WUM processes and one of its classification technique KNN. KNN is very simple and if compared with other algorithms KNN again maintains its efficiency. KNN is a satisfactory classification technique used in WUM, but is a lazy learner and the accuracy depends on the value of k. There are bundles of paper available about KNN. After analysing these papers we can be concludes that there are two approaches to improve the performance of KNN. (1) By alteration in the distance metric. (2) Hybrid version of KNN. So, we can conclude that the performance of WUM can be improved to a great extent by making modification in the KNN algorithm.

## REFERENCES

[1] Pawel Weichbroth Mieczyslaw owoc Michal Pleszkun “Web User Navigation Pattern Discovery fro WWW Serve Log Files” (2012).

[2] Tasawar Hussain, Dr. Sohail asghar, Dr. Nayyer Masood” Web Usage Mining: A Survey on Preprocessing of Web Log Files” (2012).

[3] M Agosti “Web Log Mining: A Study of User Sessios” (2007).

[4] Web Mining for Web Personalization MAGDALINI EIRINAKI and MICHALIS VAZIRGIANNIS (2003).

[5] R Cooley, B. Mmobasher and J. Srivastava “Data preparation for mining World wide Web browsing patterns”, Journal of knowledge and Information systems, (1), 1999.

[6] S. Dhawan and S. Goel “Web Usage Mining: Findings Usage Patterns from Web Logs” (2013).

[7] S. B. Imandoust and M. Bolandraftar “Application of K- Nearest Neighbor (KNN) approach for Predicting Economics Events: Theoritical Background” (2013).

[8] L. K. Joshila Grace, V. Maheshwari and Dhinaharan Nagmalai”Analysis of Web Logs and Web User in Web Mining” (2011).

[9] Kavita Sharma, Gulshan Shrivastava and Vikas Kumar “Web Mining Today and Tomorrow” (2011).

[10] Navin Kumar Tyagi, A. K. Solanki and Sanjay Tyagi “An Algorithmic approach to data preprocessing in Web Usage Mining” (2010).

[11] G. Shivaprasad, NV Subba Reddy and U. Dinesh Acharya “Knowledge Discovery from Web Usage Data: An Effective Implementation of Web Log Preprocessing Techniques” (2015).

[12] Lya Hulliyyatus Suadaa “A Survey on Web Usage Mining Techniques and Applications” (2014).

[13] Mahendra Pratap Yadav, Mhd Feeroz and Vinod Kumar Yadav “Mining the Custmor Behaviour using Web Usage Mining in E-commerce” (2012).

[14] Nandita Agarwal and Prof. Anand Jawdekar “Used-Based Approach for Finding Various Results in Web Usage Mining” (2016).

[15] Sunena and Kamaljit Kaur “Web Cultural Mining and Enhancing User Access on Web Using Culture” (2016).

[16] Suharjito, Diana and Herianto “Implementation of Classification Technique in Web Usage Mining of Banking Company” (2016).

[17] B. Kotsiantis, I. D. Zaharakis, P. E. Pintelas “Machine learning: a review of classification and combining techniques” (2007).

[18] Ms. Aparna Raj, Mrs. Bincy, Mrs. T. Mathu “Survey on Common Data Mining Classification Techniques” (2012).

[19] K. Q. Weinberger, J. Blitzer, L. K. Sau l,” Distance metric learning for large margin nearest neighbor classification” (2005).

[20] Thair Nu Phyu “Survey on Classification Techniques in Data Mining” (2009).

[21] T. Cover and P. Hart “Nearest Neighbor Pattern Classification” (1967 ).