International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 807-812 www.ijaem.net ISSN: 2395-5252

# A survey on Algorithms used in Web Usage Mining

[1,] Sheth Kevalya, [2] Maniyar Dhruval, [3] Ranpara Viraj,
*L. E. College (Diploma), Morbi*
*L. E. College (Diploma), Morbi*
*L. E. College (Diploma), Morbi*

Date of Submission: 15-02-2023 Date of Acceptance: 25-02-2023

**ABSTRACT:-** Web data mining is a specialization in data mining and it is used to extract the required information and knowledge from the ocean of large amount of web pages. The web mining has some mining categories like web content, web usage and web structure mining. In this paper the focus is on web usage mining which would be used to record and analyze user access data on the web and collect data in the form of usage logs. After visited any website user leaves some useful information such as visiting time on website, internet protocol (IP) address, visited page etc. All these information is collected, analyzed and store in usage logs. It helps to understand the user behavior and improves website structure. In this paper, one can learn about various type of web usage mining algorithm. Web usage mining technique incorporates various type of algorithm like: Decision trees, Naïve Bayesian classifier, K-nearest neighbor classifier, etc. One can gain basic knowledge of implementing the Web Usage Mining Techniques.

**Keyword:-** web data mining, web usage mining, Decision trees, Naïve Bayesian classifier, K-nearest neighbor classifier.

## I. INTRODUCTION:-

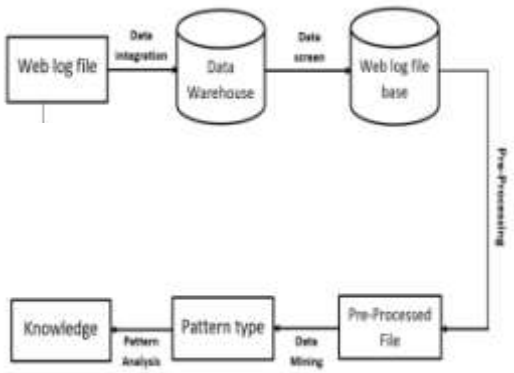
**Data mining:-**

Data extraction is a process for analysing usable information and extracting data large data stores, involving various models, smart methods, algorithms and tools. This process can help them analyse data, user behaviour and forecast future trends[1]. Data mining is about finding hidden knowledge, unexpected models and new rules in big databases. Research into databases and information technology has resulted in an approach to storing and handling this valuable data for later decision-making [2]. It is also referred to as the process of knowledge discovery, the exploration of knowledgefrom data, the extraction of knowledge, or the analysis of data/models. The aim of this technique is to find models which were until then unknown. Once these models have been found, they can be used to make certain business development decisions [3].

**Web mining:-**

Web mining is one of the varieties of strategies use in data mining. The main purpose of web mining is to automatically extract information from the web.

For discovering useful data (films, tables, audio, images etc.) from the web different techniques and tools are used [1]. Web mining is the Integration of facts collected by traditional data mining methodologies and strategies with Information gathered over the world wide web. Web mining helps to improve the exposure of web search engine via identifying the web Pages and classifying the web documents. The web mining has few mining classes such as web content, web usage and web structure [4]. Web content mining is the method of extracting meaningful knowledge from web pages or documents. Web architecture mining refers to the process of deriving information from the connection between the organizational structures of the web. Web usage mining is the extraction of information through a user's interaction with the web [5].



International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 807-812 www.ijaem.net ISSN: 2395-5252

**Web Usage mining: -**

Web usage mining facilitates in locating the user desires by using analysing the internet server log files to make the directors of the internet websites to regulate their net web page to attract greater quantity of users [6]. Web usage mining is the utility of data mining techniques to find out usage patterns from web facts, for you to understand and better serve the needs of web-based packages. Internet usage mining includes 3 stages, particularly pre-processing, sample discovery, and sample analysis [7]. Web usage mining is normally use to file user records at the web and keep inside the form of person logs. When person go to any internet site consumer put some useful records like IP deal with of person device, go to time on internet site and visited web page on internet site and a few different statistics. All this required knowledge is stored in consumer logs and it allows in recognizing user behaviour and in development of website structure.

**How to perform web usage mining: -**

Web usage mining particularly contains three phases: pre-processing, sample discovery and sample analysis. Web Usage Mining is achieved by discovering the secondary data derived with the communication of the users (while surfing on the web). It gathers useful utilization facts very well, filter irrelevant usage data, create the actual usages data, notice thrilling navigation patterns, and show the navigation patterns occurring in reality. The secondary data includes the data from the proxy server logs, web server logs, browser 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. Vacationer's behaviours are analysed and understood from a few statistics mining techniques. These are affiliation regulations, route analysis, sequential evaluation, clustering and category [8].

The most important task of web usage mining is to capture web browsing behaviour of users from a particular web site. Web usage mining can be classified according to types of usage data observed. In current context, the usage data is web log data, which maintains the information about the user navigation. With respect to 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 determine and predict user behaviour, in order to help the designer to improve the web site, to attract more visitors in your web site, or to give regular users a personalized and adaptive service [8].

**Phases of web usage mining: -**

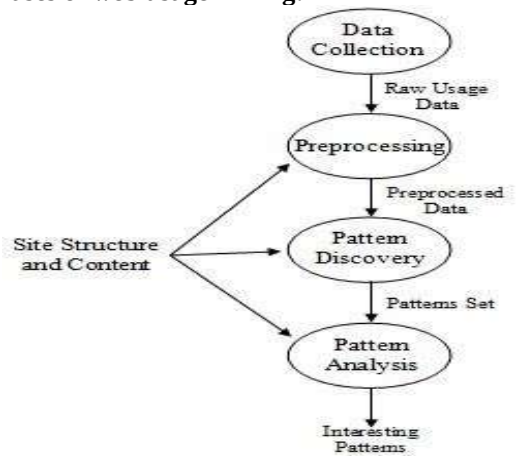



**Fig: Web usage mining phases**

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

- Web Servers
- Web proxy servers
- Client browsers

**Data pre-processing: -** The data collected from web server log files is often incomplete and create uncertainty. Pre-processing is important phase of web usage mining process in order to clean, correct and complete input data and to mine the information successfully. Pre-processing phase takes 80% time of whole process. through the pre-processing phase log file is passed from the following steps:

- Data Cleaning
- User Identification and Session Generation
- Data Conversion

**Pattern Discovery: -** Discovery of desired patterns and to abstract meaningful information from pre-processing data is a difficult task. some techniques to determine patterns from processed data are:

1. Association Rules
2. Sequential Patterns
3. Clustering
4. Classification

**Pattern Analysis: -** The last step of the entire Web Usage Mining process is Pattern Analysis. The main objective of this procedure is to select the interesting patterns and filter out uninteresting

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 807-812 www.ijaem.net ISSN: 2395-5252

patterns. The patterns are analysed using techniques such as OLAP techniques, Data and Knowledge Querying and Usability analysis.

**Web usage mining algorithm:**

**1)Naïve bayes**

The massive growth of web usage is creating notable growth in information which leads to effort in extracting useful information. In huge area WWW (world wide web), large number of contents is available in different format like text, videos, image and audio. That information is available in review sites, forums, blogs, and social media. The service of sentiment analysis systems, this formless information could be automatically converted into organized data of public opinions about products, services, brands, or any topic that people can direct ideas about. This information can be very valuable for commercial applications like marketing analysis, product reviews, product feedback and customer service. Web content mining is the mining, extraction and combination of useful information, data and facts from Web page content. Here, identification or determining designs from huge data sets is done, further these designs enable you to forecast user opinion [9].

$$P(A_i|C, A_j) = P(A_i|C)$$

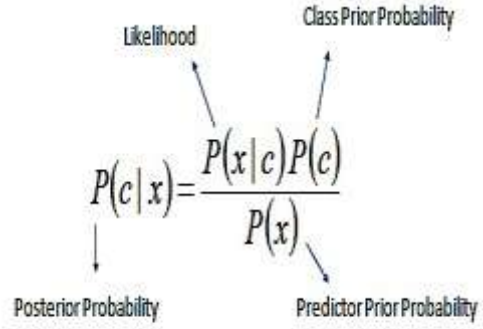
Naive bayes is a popular algorithm for organizing text. The Bayesian Classification represents a supervised learning technique as well as a numerical method for classification. The classifier is a simple yet powerful algorithm for the classification task. Naive Bayes denotes the strong independent assumptions in the model, rather than the restricted distribution of each feature. A Naive Bayes model accepts that each of the features it uses are conditionally independent of one alternative given some class. Naive Bayes classifiers are highly scalable, requiring several parameters linear in the number of features in a learning problem [9].

Bayes' rule offers the formula for the probability of Y with some specified feature X. In the real-world problems, we hardly find any case where there is only one feature.[10]

When the features are independent, one can spread Bayes' rule to what is called Naive Bayes which assumes that the features are independent that means altering the value of one feature doesn't impact the values of other variables and this is why we call this algorithm "NAIVE".[10]

Naive Bayes can be applied in wide domains like face recognition, weather prediction, Medical Diagnosis, News classification, Sentiment Analysis, and a lot more.[10]

When there are multiple X variables, we make simpler it by assuming that X's are independent, so




$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

P(c|x): posterior probability of class (c, target) provided predictor (x, attributes). This represents the probability of c being true, provided x is true.

P(c): is the prior probability of class. This is the observed probability of class out of all the observations.

P(x|c): is the likelihood which is the probability of predictor-given class. This represents the probability of x being true, provided x is true.

P(x): is the prior probability of predictor. This is the observed probability of predictor out of all the observations.

Naive Bayes classifier gives great results when we use it for textual data analysis. Such as Natural Language Processing. It is special case of text mining normally focused on classifying opinion type in score, it isn't very accurate but it is useful. The classification focuses on two sentiments i.e., positive and negative. This work has been implemented in Python language which runs on developing platform named Anaconda. In this, have occupied datasets of user's feedback from publicly available data. The simplest way to produce features from text is to separate the text up into words. Each word in the user feedback will then be a feature that have controlled with. In order to do this, first divide the sentences based on whitespace. Then count up how many times each word arises in the negative feedback and similar for positive feedbacks. To guess the possibilities on the user's feedback in test.csv and train.csv, import the file because the probabilities were generated from it and the algorithm had previous knowledge

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 807-812 www.ijaem.net ISSN: 2395-5252

regarding the data it's predicting on. Give the label as positive and negative, as the classification label got to know the accuracy used in the nltk.classify.util.accuracy along with the coding requirement. To do all this, we'll need to calculate the probabilities of each class occurring in the data, and then make a function to calculate the classification.[9]

The result has been made through the complex feedback from number of people, those are analysed and the large data of people classified consequently to feature label as positive and negative. The user feedback on a website, the datasets acquire the data site and does text classification through the naive bayes classifier. That classifies the sentiment from the sentences as positive and negative by the features [9].

**2)K-nearest neighbour classifier**

K-nearest neighbour (KNN) is a type of managed learning algorithm used for both regression and classification. KNN attempts to guess the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is close to the test data. The KNN algorithm computes the possibility of the test data belonging to the classes of 'K' training data and class holding the highest possibility will be selected. In the case of regression, the value is the mean of the 'K' certain training points [11]. Classification technique uses several algorithms as classifier among them. Nearest Neighbour (KNN) method is a very simple, most popular, very efficient algorithm. The result of K-NN classifier is a class relationship. An object is classified on the basis of number of votes of its neighbours. The object is being allocated to the class most common among its k- nearest neighbour [8].

## K- NEAREST NEIGHBOUR CLASSIFIER

### A. Basics

According to the k-nearest neighbour method the arrangement of an unknown data tuple is accomplished by analysing the classes of its nearest neighbours. KNN procedure employs the principle of nearest neighbour procedure. However, in case of KNN algorithm a fixed wide variety of nearest neighbours are suited to vote in the procedure of type of an unknown statistics tuple which is understood by k, in which k is an effective integer. When k=1 then the unknown data tuple is classified as the class of the training data tuple which is nearest to it. K-nearest neighbour 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 maximum similarity to the test. The process of category in KNN starts with a dataset. The data set is created to ensure variety of characteristics that outline a data set. The data set is separated into sets: training set and take a look at set. Training set is specified as input to the algorithm. The division of the data set can be done using various methods such as hold-out method, random sampling, cross validtion etc.

KNN organizes any new tuple by using information from tuples like it. Due to this KNN is likewise known as neighbourhood learner. There is no explicit training phase in KNN. It contains all the training tuples given to it as input without doing anything. All the calculations are done at the time of classification of a test tuple. In KNN set of rules, the training tuples may be observed as a fixed records of factors in an n-dimensional space, where n dimensions are the set of n attributes describing the statistics set. When an unknown tuple derives for classification, we have to find out the k nearest data points to it in the n dimensional space. To find the k nearest data points to the unknown tuple several distance metrics are used. For example: Euclidean distance, Minkowski distance, Manhattan distance.

### B. Distance used in KNN

The three famous distance function used with KNN are

I. Euclidian Distance: $D(x, y) = ((\Sigma^{m}_{i=1}|x_i - y_i)^2)^{1/2}$

II. Manhattan Distance: $D(x, y) = \Sigma| x_i - y_i|$

III. Minkowski Distance: $D(x, y) = ((\Sigma^{m}_{i=1}|x_i - y_i)^{r})^{1/2}$

### 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:

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

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,

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

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

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

$$\arg max\left(\sum_{x_i\in N(x_i,k)} E(y_i, c1), \sum_{x_i\in N(x_i,k)} E(y_i, c2), \ldots \sum_{x_i\in N(x_i,k)} E(y_i, cm)\right) \quad (1.3)$$

$$y_t = \arg max\left(\sum_{x_i\in N(x_i,k)} E\frac{(y_i, c1)}{k}, \sum_{x_i\in N(x_i,k)} E\frac{(y_i, c2)}{k}, \ldots\ldots, E\frac{(y_i, cm)}{k}\right) \quad (1.4)$$

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 807-812 www.ijaem.net ISSN: 2395-5252

And we know that:

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

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

$$y_t = \arg\max\{p(c_1)_{(x_t,k)}, p(c_2)_{(x_t,k)}, \ldots\ldots p(c_t$$

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.

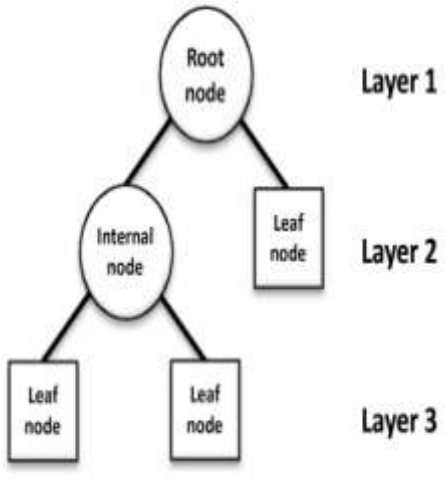
C. Assign the class which is most common in the k nearest training tuples to the unseen tuple.

**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.

**3)Decision tree**

As the computer technology and computer network technology are more and more increasing, the amount of data in information industry is getting higher and higher. It's far vital to examine this massive number of records and extract beneficial know-how from it. Decision tree classification method is one of the most popular data mining techniques. In decision tree 'divide and conquer' technique is used as basic learning strategy. A decision tree is a structure that includes a root node, branches, and leaf nodes Every internal node represents a take a look at on a characteristic, every branch denotes the final results of a take a look at, and every leaf node holds a class label. The uppermost node in the tree is the root node [12].




A decision tree is a flowchart alike tree structure, in which every inner node represents a check on an attribute, each department denotes an outcome of the take a look at, class label is denoted by each leaf node (or terminal node). Given a tuple X, the attribute values of the tuple are verified against the decision tree. A path is traced from the root to a leaf node which holds the class calculation for the tuple. It is easy to change decision trees into classification rules. Tree models in which the target variable can take a hard and fast set of values are known as classification trees. In this tree shape, leaves denote class labels and branches denote conjunctions of features that cause those class labels. Decision tree can be created comparatively fast compared to other methods of classification [12].

**Example of decision tree:**

There are several algorithms used to build decision Trees CHID, CART, ID3, C4.5, and others.

- CHID (Chi-square–Automatic–Interaction–Detection): is an important decision tree learning algorithm to handle minor attributes only. It is a supplement of the automatic communication detector and theta automatic communication detector procedures.
- CART (Classification - regression tree): is the most common algorithm in the numerical community. In the fields of statistics, CART helps decision trees to increase reliability and receipt in additional to make binary splits on inputs to get the purpose.
- ID3 (Iterative Dichotomise 3): It is an easy way of decision tree algorithm. The calculation used to build the tree is data gain for splitting criteria. The development of tree stops when all samples have the same class or data gain is

not greater than zero. It fails with numeric attributes or missing values.

- C4.5 is the ID3 enhancement or extension that presented by the same author. It is a combination of C4.5, C4.5-no-pruning, and C4.5-rules.It uses gain ratio as splitting criteria. It is a best choice with numeric attributes or missing values.[13]

## II. CONCLUSION

Web mining comprises a series of techniques which focuses at obtaining intelligence from data of web. It is useful obtaining useful information effectively and making business decisions. With more than 2 billion web pages it is difficult to gather the analytics but due to web mining it is very feasible. One can get clean, required and structured data using these techniques.

## REFERENCE

[1]. Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview Jawad Mughal
[2]. Book :- Data mining Pieter Adriaans
[3]. DATA MINING TECHNIQUES AND APPLICATIONS Mrs. Bharati M. Ramageri
[4]. Mining Web Content Using Naïve Bayes Classification Analysis K.S. JeenMarseline
[5]. Understanding the Classification of Data Mining and Web Mining, Gehad Abdallah Amran, Hassan Faisal Aldheleai, Hussein Al-Sanabani
[6]. Web usage Mining for Exploring User Needs and Interest,Dr. V. Govindasamy , V. Akila , D. Dinesh
[7]. Web usage mining: discovery and applications of usage patterns from Web data
[8]. Jaideep Srivastav, Robert cooley , Mukund Deshpande
[9]. A Review of classification in Web Usage Mining using K- Nearest Neighbour ManishaKumari ,SaritaSoni.
[10]. Mining Web Content Using Naïve Bayes Classification Analysis ,K.S. JeenMarseline
[11]. Naive Bayes Algorithm: A Complete guide for Data Science Enthusiasts ,Anshul Saini
[12]. K-Nearest Neighbor From: medium.com
[13]. A Survey on Decision Tree Algorithms of Classification in Data Mining , Himani Sharma, Sunil Kumar
[14]. Comparison of Classification Modelling Algorithms in Web Usage Mining, Srujani J PritiBadar
[15]. Popular Decision Tree Algorithms of Data Mining Techniques: A Review , Radhwan H. A. Alsagheer, Abbas F. H. Alharan, Ali S. A. Al-Haboobi