I. INTRODUCTION

1.1 Background of the Research

World Wide Web (WWW) is the largest source of online data that reflects a virtual company in this internet era. The information available is massive and it is increasing at an enormously high rate. The volume and dynamic nature of the WWW coupled with the lack of an integrated structure or scheme, makes information retrieval a tedious and challenging task. As the sources of internet usage are growing rapidly, it is necessary for users to use automated tools to find desired information [137]. Users require systems at client side and server side for mining data and extracting the desired information. For a specific user only small information from the web is useful and while the rest of the information is irrelevant at best. The process of accumulating user requests is an automatic process performed by the web server in the form of web log which can be analyzed to understand user behavior.

User behavior analysis is used in different domains such as e-commerce, e-learning, health sectors etc., Web personalization offers many functions such as simple user salutation, to more complicated functions such as content delivery as per users interests. Recommendation systems are another common application where the interested web pages are recommended to users. User Access Patterns are represented as association rules from web logs by a Formal Concept Analysis [148]. The use of user behavior analysis or pattern analysis has found a wide variety of uses in various fields to improve overall system and also to design the system.
Applications of Web Usage Mining in Different Domains

The system finds its applications in various domains. They are listed as follows:

- **E-Commerce** is the main source of shopping now. For Customer Relationship Management, customer’s details should be stored and analyzed for the continued success of business. Usage mining techniques are very useful to focus on customer attraction, customer retention, cross sales and customer departure.

- **System Improvement** is done by understanding the web traffic behavior by mining log data so that policies are developed for Web caching, load balancing, network transmission and data distribution.

- **Site Modification** is a process of modifying the web site and improving the quality design and contents based on the interest of users. Web Pages can be modified or linked as per customer behavior.

- **This technology has been used by government agencies to classify threats and fight against terrorism.** The predicting capability of mining applications can benefit society by identifying criminal activities. Patterns for detecting intrusion fraud, attempted break-ins are also provided by mining and widely used for crime investigation. Activities like hacking, internet fraud, fraudulent websites, illegal online gambling, virus spreading, child pornography distribution & cyber terrorism are easily identified based on the patterns detected by web mining techniques and are apprehended accordingly.

- **It can also be implemented in Intranet and Extranet for resource optimization within the organization.**
In e-learning, it allows educators to understand the access behavior of learners, compare learners and their access patterns and cluster similar learners.

62.2% of internet users visit Health related websites. Understanding the user’s interests and needs helps these websites to give relevant information to the patients.

1.2 Existing Tools for Analyzing Web Log

There are many public and shareware tools available for log or traffic analyzers whose functionality is limited to produce statistical reports. Some of them are Analog, WebLogs, WebLog, Ststat. All these analyzers provide users with reports. On the other hand another tool named Follow 2 tracks user sessions and presents specific information about the individual user.

For e-commerce log analysis alone many systems are available like Accrue, Elytics, Lumio, NCR, WebTrends etc. These systems analyze products purchased, advertisements and click-through rates which provide key indicators to the marketers of a company. Enterprise Insight, a product by Epiphany, contains tools for analyzing Web and Commerce server logs. It shares data with Epiphany, personalization engine and integrates web visitor information with customer data from other operational systems.

1.3 Categorization of Web Mining

Web mining is categorized into three types depending on the web data used for mining process. They are

- Web Content Mining
- Web Structure Mining
- Web Usage Mining.
1.3.1 Web Content Mining

Web content mining [60] is the task of discovering useful information available on-line from the content data of the web pages. Content data correspond to the collection of facts of a web page to convey information to the users. The aim of Web content mining is to provide an efficient mechanism to help the users to find the information they seek. There are different kinds of Web content like multimedia data, structured documents, semi-structured documents and unstructured data which can provide useful information to users. Web content mining includes the task of organizing and clustering the documents and providing search engines to access the different documents by keywords, categories, contents etc.

1.3.2 Web Structure Mining

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting two related pages. In addition, the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Thus, Web Structure Mining can be regarded as the process of discovering the structure of information from the Web.

1.3.3 Web Usage Mining

Web usage mining, also known as web log mining is the application of data mining techniques to large web log repositories in order to discover useful knowledge about a user’s behavioral patterns and website usage statistics that can be used for various website design tasks. Usage data also captures the identity or origin of Web users.
The main source of data for web usage mining consists of textual logs collected by numerous web servers all around the world. There are three phases in web usage mining.

1. Preprocessing is the first phase where data cleaning, user identification, session identification, path completion and feature selection are carried in the web log data.

2. Pattern discovery is the important phase where various data mining techniques like statistical analysis, association, clustering, pattern matching etc. are used to process the data.

3. Pattern analysis is the third phase in which uninteresting rules are filtered out from the patterns discovered. Analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations [34].

1.4 Web Log Preprocessing

The first and foremost step of web log mining is preprocessing which lays the foundation of the entire web log mining. It acts as the key of quality assurance which impacts the rule and pattern produced by the data mining algorithm. It is time consuming and takes 80% of mining process [113]. Transformation of the raw click stream data into a set of user profiles is the goal of preprocessing [38]. The main task of data preprocessing is to select standardized data from the original log files, prepared for user navigation pattern discovery algorithm. A variety of heuristics for preprocessing tasks such as merging and cleaning, user and session identification etc are available and there are a number of unique challenges in each of these tasks [123]. A lot of research works has been carried out in this preprocessing area for grouping sessions and transactions, which is
used to discover user behaviour patterns. The result of data preprocessing will directly affect the accuracy and reliability of the algorithm processing results [163].

1.4.1 Data Cleaning

Log data consists of irrelevant entries along with the user’s requests. If a user requests to view a particular page along with server log entries, graphics and scripts are downloaded in addition to the HTML which is irrelevant for mining. So jpeg, gif files, sound files are removed to improve data quality to analyze it. Not all the requests are fulfilled by the server. The failed requests are identified by status code in a log entry. The requests with status code less than 200 and the ones with greater than 299 were removed, since only the successful entries lies within this range. Robot requests or spider navigations if any are eliminated.

1.4.2 User Identification

Web user identification is one of the most challenging steps in the process of web usage mining. In the case of simple market basket analysis, the customer is identified exactly by his/her customer ID. In the analysis of navigation patterns it is difficult to know which downloaded page belongs to which user. The second problem is that the same user can use multiple computers, and many people can use the same computer. The third problem is proxy servers which hide relevant information about unique users as multiple computers appear on the internet using the same IP address through the proxy server. The simplest method to find is to group similar IPaddress and browsing agent fields in the entries.

1.4.3 Session Identification
Sessions are defined as a set of pages visited by a user of one particular visit to a website in a specified duration. A user may have a single session or multiple sessions during a period. Once a user has been identified, the click stream of each user is divided into sessions. There are three methods in session identification based on time and navigation. The simplest methods are time oriented in which one method is based on total session time and the other is based on single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. The default time given by Cooley is 30 minutes [128]. The second method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes then the second entry is assumed as a new session. The third method is a referrer based method is proposed on the basis of navigation in which referrer URL of a page should exist in the same session. If no referrer is found then it is the first page of a new session.

1.4.4 Path Completion

Path completion step is carried out to identify missing pages due to caching and usage of ‘Back’ key by the user while browsing. Path Set is a set of incomplete accessed pages in a user session. Using the referrer-based method, the reference length of pages in the complete path is modified by considering the average reference length of auxiliary pages. This is estimated in advance through the maximal forward references and the reference length algorithms. Transactions can be identified from complete path set [157].

1.4.5 Session Reconstruction

This is an additional step for the proposed research work to transform the user sessions into a matrix format which simplifies the process of mining. After path completion, a matrix is formed with web pages as columns and user sessions are rows. A numeric entry is made whenever a page is browsed and for every revisit it is incremented [26].
1.4.6 Feature Selection

The high dimensionality of data can cause data overload. Added to this, if there are a lot of features, it is possible that many of them are irrelevant for data mining operations. The complexity is increased and the analysis results are not accurate because of the redundant and irrelevant features. Moreover, memory and computation time increases and it gives an unnecessary load to the systems. The solution for these problems is the reduction of data dimension.

1.5 An Impression on Dimensionality Reduction

Due to the complex nature of data and the increase in the volume of data, research in mining has taken a sweep to a greater extent. The use of multidimensional data will result in more noise, complex data, and the possibility of unconnected data entities. To reduce processing time and to alleviate the limitations of dimensionality during mining process, data reduction is needed. Dimension reduction is a technique that is widely used for various applications to solve the issues related with dimensionality.

Dimension reduction is important in cluster analysis, which not only makes the high dimensional data addressable and reduces the computational cost, but can also provide users with a clear picture and visual examination of the interesting data [18]. Many emerging dimensionality reduction techniques have been proposed in literature. For example, Local Dimensionality Reduction (LDR) approach tries to find local correlations in the data, and performs dimensionality reduction on the locally correlated clusters of individual data and dimensionality reduction adaptively adjusted and integrated with the clustering process [46][29]. There are four major reasons for performing dimension reduction, which are depicted in figure 1.1. Each reason can be referred to as a distinctive sub-problem [94] and they are as follows.
- Decreasing the training(model) cost
- Increasing the performance
- Reducing irrelevant dimensions
- Reducing redundant dimensions

The cost involved in mining algorithms is calculated on the basis of time required for the algorithm to run and the size of the dataset. Performance is measured in terms of the accuracy of the learning model. Reduction of redundant dimensions and of irrelevant dimensions can be further divided into two sub problems: Feature selection and Record selection.

![Taxonomy of dimension reduction problem](image)

**Fig 1.1: Taxonomy of dimension reduction problem**
Record selection is the selection of relevant and important records which aid the learning process better than others [7]. Feature selection is to identify and select features in the dataset as important, and discard any other feature as irrelevant and redundant information. Since it reduces the dimensionality of the data, it holds out the possibility of more effective and rapid operation of data mining algorithms and accuracy of future classification can be improved. There are three types of Feature selection algorithms in machine learning literature: filter methods, wrapper methods, and embedded methods.

The wrapper methods assess subsets of variables according to their usefulness to a given predictor by conducting a search for a good subset using the learning algorithm itself as part of the evaluation function. It has a higher risk of over-fitting and is very computationally intensive, especially in building the classifier and has a high computational cost. Embedded methods perform variable selection as part of the learning procedure and are less computationally intensive than the wrapper method but are specific to a learning machine. Filters, being motivated by the data distribution, perform feature selection independently of any particular classifier. In this method the optimal subset is found by using an empirical risk estimate for a particular classifier. It easily scales to very high-dimensional datasets, is computationally simple, fast and is independent of the classification algorithm [109].

Several dimensionality reduction methods are available in literature which have been used in several applications [40]. Many methods use the second order statistics, and the variance or covariance of the data in the optimization of the objective function. Some of the methods are Principal Component Analysis, Multivariate Linear Regression, Partial Least Squares, Factor Analysis and Canonical Correlation Analysis. Second-order
methods reduce the redundancy of the data by using a simpler representation but do not give a better representation since the latent variables extracted by them are not independent. Rather, they are correlated. In recent years, principal component analysis (PCA) and Independent component analysis (ICA) are widely used in various applications due to their significant performance [82]. Both are linear dependent methods which takes only the linear dependencies between variables.

1.5.1 Principal Component Analysis

Principal Component Analysis is a filter method widely employed for feature selection using the linear dimension technique. The transformation takes place linearly by retaining the characteristics of the original data set. It is useful in image processing applications [109] for numeric attributes. It performs an orthogonal transformation on the input space, to produce a lower dimensional space in which the main variations are maintained.

1.5.2 Independent Component Analysis (ICA)

ICA is a rich statistical technique developed for digital signal processing applications to reveal hidden factors for measurement of signals. It also acts as a powerful tool for analyzing text document data if the documents are presented in a suitable numerical form. ICA has received considerable interest in recent years because of its versatile applications such as source separation, channel equalization, speech recognition and functional magnetic resonance imaging, face recognition, telecommunication, predicting stock market performance and financial market data mining [56]. This method finds a linear transformation in which the extracted components are mutually independent from each other and it focuses only on the statistical independence of features in the high dimensional data. It has been used for dimension reduction and representation of word histograms[82].
Blind Source Separation (BSS) problem is an area where ICA found its use initially to find some hidden sources from the observed mixture data. These hidden sources are assumed to be mutually independent and may give a simpler and better representation of the data. They are called the independent components. The process of extraction of the independent components from the observed mixture data is called independent component analysis.

**Definition for Independence**

Independence is defined by the probability densities in mathematics between two scalar-valued random variables $x$ and $y$\cite{56}. The variables $x$ and $y$ are said to be independent if information about $x$ does not give any information on $y$ and vice versa.

Let us denote by $p_1(y_1,y_2)$ the joint probability density function of $y_1$ and $y_2$.

$$p_1(y_1) = p(y_1,y_2)dy_2$$\hspace{1cm}(1.1)

An analogous formula holds for $p_2$. Then $y_1$ and $y_2$ are independent if and only if the joint probability density function

$$p(y_1,y_2) = p_1(y_1)p_2(y_2)$$\hspace{1cm}(1.2)

This definition extends naturally for any number ‘$n$’, of random variables, in which case the joint density must be a product of $n$ terms.

**General Procedure of ICA**

To extract new patterns or best patterns for the purpose of reducing the dimensions of patterns space and to achieve better performances ICA is used. Each and every feature is normalized by calculating mean and standard deviation. The Absolute mean is calculated for each and every row. The Independent matrix is created by
comparing and shrinking the values. Finally the independent matrix is multiplied with the original matrix and the mean is calculated for all attributes. The mean is compared with a threshold and the attributes which are less than the threshold are selected.

1.5.3 Feature Selection in Web Log Data

Generally, dimension refers to the number of attributes or features in a matrix. In this work, dimensionality reduction is referred as the measurement and elimination of the unimportant web pages in the session matrix—the ones that are not frequently used by the users.

1.6 Clustering in Mining

Various definitions exist for a cluster since it is application specific. A few definitions given by [63] are as follows.

“A set of entities which are alike, and entities from different clusters are not alike”

“An aggregation of points in the space such that the distance between two points in the cluster is less than the distance between any point in the cluster and any point not in it”

Clustering is the division of data into groups of similar objects. It disregards some details in exchange for data simplification. Informally, clustering can be viewed as data modeling concisely summarizing the data, and, therefore, it relates to many disciplines from statistics to numerical analysis. The grouping of different things is a natural process for human beings. There exist numerous natural examples of different classifications of living things in the world [82]. In web portal design, taxonomies are often created to describe categories and subcategories of topics found on the Web site. Many objects existing with different names are the results of unsupervised categorization processes made by humans who divide objects into separate classes by using their observable
characteristics. Clustering follows unsupervised learning for which no prior knowledge is expected on any object classifications, and plays an important role in a broad range of applications like Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. This indicates the prominence of clustering in mining research, pattern recognition and statistics.

1.6.1 Clustering Methods

The Clustering techniques can be applied to data for numerical data, categorical data or a mixture of both. There are many issues in clustering since there is no standard solution to normalize the data. Many similarity measures are available. The tremendous growth in the volume of data is a big challenge for clustering activity. To overcome these issues many clustering methods were developed. Since the results of the algorithm are different with different features, the method should be selected based on the application and data type.

The clustering algorithms are categorized into 6 categories in general. They are Partitioning, Hierarchical, grid-based, density-based, model-based and graph-based clustering algorithms [14]. The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. Similarity is the key measure to find distance or similarity between data points. Types of clustering algorithms are depicted in figure 1.2.
Hierarchical method creates hierarchical decomposition of objects and it is classified into agglomerative (bottom-up) approaches and divisive (top-down) approaches based on how the hierarchy is formed. Agglomerative clustering starts with single point clusters and recursively merges two or more of the most appropriate clusters. When requested, a number of k clusters are obtained and the process stops its execution. Divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster till the k number of clusters is obtained.

Partitioning method is useful for applications where a fixed number of clusters are required and it classifies objects into several one-level clusters, where each object belongs to exactly one cluster, and each cluster has at least one object. It is divided into numerical methods and discrete methods.

Density-based method is based on the number of objects in neighborhood – density and is used to discover clusters with arbitrary shape. It typically regards clusters as dense regions of objects in the data space that are separated by regions of low density. The algorithm
finds characteristic descriptions for each and every group which represents a class or a concept. It attempts to optimize the fit between the given data and some mathematical models.

**Grid Based Algorithm** performs all the operations in a grid structure which is formed by partitioning the space into a finite number of cells. The main advantage of the approach is its fast processing time.

**Graph Based Algorithms** is a subset of nodes in a graph such that every two nodes in the subset are connected by an edge and it can be a prototypical form of cluster. This algorithm finds its application in biological networks such as Protein-Protein interaction, community detection in social networks etc.

**1.6.2 Hard Clustering and Soft Clustering**

This is another categorization of clustering based on the nativity of an object to a cluster. Clusters can be interpreted as subsets on the basis of set theory. Methods can be classified as soft or hard. Hard clustering, otherwise called crisp methods, is based on classical set theory and requires that an object either belongs to a cluster or not. It partitions the data into a specified number of mutually exclusive subsets.

Soft Clustering also called as Fuzzy clustering methods, however, allow the objects to belong to several clusters simultaneously, with different degrees of membership. Every point has a degree of belonging to the clusters, as in fuzzy logic. It is more natural than hard clustering in many applications. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. The discrete nature of the hard partitioning also causes difficulties with algorithms based on analytic functional.
1.6.3 Defuzzification

Defuzzification is the process of producing crisp sets from resultant fuzzy sets and corresponding membership degrees at the end of clustering. It is typically needed in fuzzy control systems. There are many methods available in defuzzification as Lambda-cut sets, Centroid method, Height method, Weighted average method, Mean–max method, Centre of sums, Centre of largest area. The simplest but least useful defuzzification method is to choose the set with the highest membership.

1.6.4 Importance of Clustering in Web Mining

Due to the exponential growth of the World Wide Web and its users a lot of issues are to be solved for searching, interacting, doing business etc., Mining is the best solution for various issues and to explore new dimensions for websites. Finding Users interests through their behavior is an active research area to increase the performance of websites.

Clustering is a technique to group together a set of web data items having similar characteristics [61]. Clustering can be performed on users or page views. Web data clustering is the process of grouping web data into “clusters” so that similar objects are in the same class and dissimilar objects are in different classes. Clustering analysis in web usage mining intends to find the cluster of users, pages, or sessions from web log file, where each cluster represents a group of objects. User clustering is designed to find user groups who have common interests based on their behaviors, and it is critical for user community construction. Web mining has obvious fuzzy characteristics, so fuzzy clustering is better suited for the web mining [6].
One important research point in web usage mining is the clustering of web users and their browsing navigation patterns. There has been an increased demand for understanding of web-users due to Web development and the increased number of web based applications. Web designers understand the users better and provide suitable customized services to the users based on their navigation pattern analysis.

1.7 Clustering Users Navigation Patterns

Web log data preprocessing has resulted in the definition of user navigation patterns. By mining web user’s historical access patterns not only the information about how the web is being used, but also some demographics and behavioral characteristics of web users could be determined to perform market segmentation in e-Commerce applications or provide personalized web content to the users. The navigation path of the web users gives valuable information about the user interests.

Clustering is a process of grouping of users navigation patterns into ‘clusters’ in a way that similar patterns are placed in the same group or classes, while dissimilar user patterns are placed in different groups or classes. Clustering web usage data is different from the traditional clustering due to the nature of data type. Therefore, there is a need to develop specialized techniques for clustering analysis based on Web usage data.

The first step is to determine the attributes that should be used to estimate similarity between user sessions. From feature selection the relevant attributes are selected for clustering. Then, the “strength” of the relationships between the navigation pattern values are estimated. The most popular measures that are used are Euclidean distance, Mahalanobis distance, Cosine measure, and Jaccard coefficient. Finally, clustering algorithms (hierarchical or partitional) are applied in order to determine the
clusters to which each user session will be assigned. The following flowchart gives a pictorial representation of the preprocessing and clustering steps in web log data analysis. The input for clustering is the navigation patterns with selected features which are selected from the feature selection methods.

Fig 1.3: Flowchart for Clustering Process

1.7.1 Fuzzy clustering in User Navigation Patterns

Traditional clustering approaches generate partitions in which each user navigation pattern data belongs to one and only one cluster. Hence, the clusters in a hard clustering are disjointed. Fuzzy clustering extends this notion and suggests a soft clustering schema [53]. The fuzzy set theory provides a very powerful soft partition which divides the users into different groups and various algorithms exists on soft clustering. Fuzzy clustering produces overlapping cluster partitions and it is widely studied and applied in various areas.
**Fuzzy C Means**

The fuzzy c-means (FCM) clustering algorithm is the best known and most powerful methods used in cluster analysis. In FCM, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. User navigation patterns on the edge of a cluster may be in the cluster to a lesser degree than user navigation patterns in the center of cluster.

The procedure of FCM is briefed as follows. The process aims at minimizing the objective function of FCM with centroid values calculated at each iteration

\[
J_{FCM} = \sum_{K=1}^{N} \sum_{i=1}^{c} (u_{ik})^q d^2(x_i, v_i) \tag{1.3}
\]

where ‘d’ is the distance between centroid and pattern. A fuzzy matrix is created and updated with new membership values given as

\[
u_{ik}^{(b+1)} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{k,i}}{d_{k,j}} \right)^{\frac{q-1}{q}}} \tag{1.4}
\]

In fuzzy c-means, the centroid of a cluster is the mean of all user navigation patterns points, weighted by their degree of belongingness to the cluster:

\[
c_k = \frac{\sum_x w_k(x)^mx}{\sum_x w_k(x)^m} \tag{1.5}
\]
The degree of belonging, \( w_k(x) \), is related inversely to the distance from pattern to the cluster center as calculated on the previous iteration of the user navigation patterns from different sessions. ‘m’ is a parameter to control how much weight is given to the closest center of the data. The process repeats until the difference between two membership values converge.

### 1.7.2 Penalized Fuzzy C Means

Penalized Fuzzy C-Means is an extended algorithm of Fuzzy C Means to incorporate spatial context in the clustering. To achieve this objective function fuzzy c means is penalized by a regularization term. The new objective function of the PFCM is defined as follows:

\[
J_{PFCM} = \sum_{k=1}^{N} \sum_{i=1}^{c} (u_{ik})^q d^2(x_k, v_i) + \gamma \sum_{k=1}^{N} \sum_{j=1}^{N} (1 - \mu_{ij})^q w_{kj} (1.6)
\]

An iterative algorithm for minimizing the above objective function is derived by evaluating the centroids of similar data points. The constrained optimization in function will be solved using one Lagrange multiplier in the following equation (1.7)

\[
\eta_q = \sum_{k=1}^{N} \sum_{i=1}^{c} (u_{ik})^q d^2(x_k, v_i) + \gamma \sum_{k=1}^{N} \sum_{j=1}^{N} \sum_{i=1}^{N} (u_{ik})^q (1 - \mu_{ij})^q w_{kj} + \lambda (1 - \sum_{i=1}^{c} u_{ik}) (1.7)
\]

Initially the cluster centroids of each of the navigation patterns are fixed randomly, And also values for number of clusters, a fuzzification parameter and an error value is initialized. Membership values are calculated as follows.

\[
\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left( \frac{d^2(x_k, v_l) + \gamma \sum_{j=1}^{N} (1 - \mu_{ij})^q w_{kj}}{(q-1)} \right)^{1/(q-1)}} (1.8)
\]
The cluster centroids for user navigation patterns are calculated (1.9) iteratively until convergence takes place in membership.

\[ V_i^* = \frac{\sum_{k=1}^{N} (u_{ik})^q x_k}{\sum_{k=1}^{N} (u_{ik})^q} \]  

(1.9)

When the algorithm has converged, a defuzzification process takes place to convert the fuzzy partition matrix \( U \) to a crisp partition. A number of methods have been developed to defuzzify the partition matrix \( U \), among which, the maximum membership procedure is the most useful one in many applications. The procedure assigns the user navigation patterns \( k \) to the class \( C \) with the highest membership

\[ C_k = \arg_{i}(\max(u_{ik})), i = 1, ... c \]  

(1.10)

1.7.3 Bolzano-Weierstrass Theorem

The Bolzano–Weierstrass theorem is a set of theorems about convergence in a finite-dimensional Euclidean space \( \mathbb{R}^n \). The theorem is named after Bernard Bolzano and Karl Weierstrass. There are many theorems about bounded sets and to a list a few

**Theorem 1:**

“Every bounded sequence has a convergent subsequence” [45]

This is a nested interval theorem to imply the intersection of all the intervals \([a_n, b_n]\) is a single point ‘w’. The theorem is proved with a sequence of ‘n’ numbers. Let \( \{w_n\} \) be a bounded sequence. Then, there exists an interval \([a_n, b_n]\) such that \( a_i \leq w_n \leq b_n \) for all \( n \). A sequence of intervals \( \{[a_n, b_n]\} \) can be obtained by mathematical induction as follows.

1. for each \( n \), \([a_n, b_n]\) contains infinitely many terms of \( \{w_n\} \)

2. for each \( n \), \([a_{n+1}, b_{n+1}] \subseteq [a_n, b_n]\)

3. for each \( n \), \( b_{n+1} - a_{n+1} = \frac{1}{2} (b_n - a_n) \)
**Theorem 2:**

“Every bounded infinite set in $\mathbb{R}^n$ has an accumulation point”

For $n=1$, an infinite subset of a closed bounded set $S$ has an accumulation point in $S$. For instance, given a bounded sequence $a_n$ with $-C \leq a_n \leq C$ for all $n$, it must have a monotonic subsequence $a_{n_k}$. The subsequence $a_{n_k}$ must converge because it is monotonic and bounded. Because $S$ is closed, it contains the limit of $a_{n_k}$[156].

**Applications of Bolzano–Weierstrass theorem**

The Bolzano–Weierstrass theorem finds its application in economics to prove the existence of Pareto efficient allocation. When no allocation is possible in the matrix of consumption bundles for agents in an economy, then it is termed as Pareto efficient allocation.

**1.8 Classification in Mining**

Classification is the task of mapping a data item into one of several predefined classes [61]. The process involves building a learning model which is used to classify a class of objects to determine the class label of a new object whose class is not determined. Since the class label of each training sample is provided, this process is known as supervised learning.

Classification is divided into binary classification and multi-class classification on the basis of the number of classes in the problem[118]. The binary classification categorizes instances into exactly one of two classes and multi-class classification deals with more than two classes. Depending on the number of labels assigned to an instance, classification is also divided into single-label classification and multi-label classification.
In single-label classification, one and only one class label is to be assigned to each instance, while in multi-label classification, more than one class can be assigned to an instance. If a problem is multi-class, for instance categories of students as slow learners, average, fast learners on the basis of their performance, it can be either single-label, where exactly one class label can be assigned to an instance. If a classification is multi-label, an instance can belong to any one, two or all of the classes. In an internet marketing scenario, a customer can be classified as ‘not a customer’, ‘once visited customer’ and ‘a regular customer’ based on their browsing patterns and discovered rules for attracting the customers with exclusive special offers [17]. Based on the probability of class assignment, classification can be divided into hard classification and soft classification. In hard classification, an instance can either be or not be in a particular class, without an intermediate state while in soft classification, an instance can be predicted to be in some class with some likelihood. Classification algorithms consist of three phases, a training phase that consists of labeled records, a test phase using previously unseen labeled records, and a deployment phase that classifies unlabeled instances [50].

1.8.1 Classification methods

Classification methods are classified as the following types.

**Decision Tree based Methods** are constructed beginning with the root of the tree and proceeding down to its leaves. Rules can also be extracted from decision trees easily [120] [121].

**Rule-based Methods** are one of the most popular techniques and are expressed in the form of an IF-THEN rule. Depending on the type of rule, it is discriminated as association rule or predictive rule learning.
Memory based reasoning method is a computational model of human reasoning in which the results of the reasoning process depend heavily on the contents and the current state of memory. The contents of memory are a collection of interrelated descriptions of concepts, events, scripts, algorithms, etc. The state of the memory makes it possible to gain access to some descriptions and renders others inaccessible.

Neural Networks based methods are effective in large databases with known prior examples and it is particularly effective for predicting events.

Naive Bayes methods is a Bayesian network or a statistical model to represent random variables and their conditional dependencies via a directed acyclic graph.

Vector Machines are machine learning methods with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. They are categorized as supervised learning models and they are useful for large data sets.

1.8.2 Classification in Web Usage Mining

Web Usage mining is the application of mining techniques in web log collected in the web server. This analysis is useful since data is intrinsic and loyal. The analysis of user behavior provides valuable information for web designers to quickly respond to their individual needs and for their efficient organization of the website. It helps organizations in decision making. The organizations are interested in developing a profile of users belonging to a particular class or category. This requires a study of the user navigation pattern that best describes the users interest. The patterns are grouped into classes or categories. The major goal is to predict the target class based on grouped data. This would help the organizations to better satisfy the requirements of a
particular customer. Web Usage Mining classification is usually used to construct profiles of users belonging to a particular class or category. A new approach to classify user patterns by modified naïve Bayesian classifier is proposed [37].

1.9 Ensemble classification

Training all the data in classification leads to an increase in time complexity and leads to inaccurate results for voluminous data. Since web logs are voluminous in nature and to overcome these issues, ensemble is an appropriate model for classification. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions [39]. The base of this concept is the well known Condorcet’s Jury Theorem [129], for democracy, which refers to a Jury of voters who need to make decision regarding a binary outcome. A number of classifiers are integrated and the classification results are predicted by either majority voting or boosting process. This method has applications in a wide range of real world problems, such as object detection, distributed computing, large-scale data, privacy-preserving applications, pattern recognition and many other fields. Due to multiple learner system there is an increase in accuracy and reliability of the classification since the failure of few learners may not affect the overall learning. The uncorrelated errors are eliminated by averaging the results. The ensemble methods are efficient, accurate and time saving and are independent of the errors of individual learners. A summary of advantages are as follows

- Enhanced accuracy and robustness over single classification methods.
- Combining predictions of an ensemble is often more accurate than the individual classifiers that make them up.
- Divide and conquer technique decomposes a complex learning task into small tasks to increase the efficiency.

- Reduce the Statistical, Computational and Representational shortcomings of standard learning algorithms [39].

1.9.1 Building Blocks for Ensemble Classification

The building blocks for ensemble classification [129] are as follows

- A training set, which can be described in different ways, and it is the dataset for ensemble learning.

- A base inducer is an induction algorithm that obtains a training set and a classifier is formed which represents the generalized relationship between the input attributes and the target attribute.

- Diversity Generator is the component which is responsible for generating diverse classifiers.

- Combiner component is responsible for combining the classification of the various classifiers.

1.9.2 Types of ensemble models

Ensemble classifiers are learning algorithms that construct a set of many individual classifiers called base-learners and combine them to classify test data points by sample average. It is now well-known that ensembles are often much more accurate than the base-learners that make them [13]. Depending on the number of techniques applied for ensembling it is categorized as homogenous and heterogeneous. Homogeneous ensemble
methods use the same base learner on different distributions of the training set. A heterogeneous ensemble is an ensemble with a set of base classifiers that consist of models created using different algorithms.

The ensemble methods are divided into two main types depending on the interrelationship between classifiers as Dependent models and Independent models. The combination technique differs from model to model.

1.9.2.1 Independent Models

The process is done concurrently and called as concurrent model. In this methodology the original dataset is transformed into mutually exclusive datasets from which several classifiers are trained. The results are combined by using combination methods and these methods are independent of the induction algorithms and different inducers can be used with each dataset. Bagging is a well known independent ensembler.

Bagging improves generalization performance by reducing variance of the base classifiers. The performance of bagging depends on the stability of the base classifier. If a base classifier is unstable, bagging helps to reduce the errors associated with random fluctuations in the training data. However, if a base classifier is stable, it may not be much of an improvement. Hence it degrades the performance. Finally the output is combined with a weighted average method. The most important aspect is its decreased susceptibility to over-fitting in noise data.
1.9.2.2 Dependent models

The process is sequential and so is appropriately termed as a sequential model. It is also called Incremental Batch Learning since the output of the first iteration is given as prior knowledge to the second iteration. The learning algorithm uses the current training set together with the classification of the former classification for building the next classifier. The last iteration’s classifier is taken as final classifier. Boosting is a well known model of this type of classification.

Boosting is a general method for improving the accuracy of any given learning algorithm. Training starts with one base learner at a time. A misclassified data subset is retrained with higher weights in the next iteration. The process is an iterative one until the best classification is obtained.
1.10 Machine Learning Methods for Classification

Machine learning is a branch of artificial intelligence which analyzes data and information is acquired from data. The core of machine learning deals with representation and generalization. Representation is the process of representing data and evaluating functions on data instances and generalization is the property that the system will perform well on unseen data instances.

Machine learning is used for a number of applications. For example images of faces are analyzed and categorize into normal people, criminals etc., Object identification, pattern identification, are implemented successfully with machine learning systems. For example, a machine learning system could be trained on email messages to learn to distinguish between spam and non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders. There are various machine learning methods that exist among which Vector Machines is an emerging and better classification algorithm and follows the supervised method for classification.
1.10.1 Support Vector Machines

Support vector machines is a popular and promising supervised learning method. Support Vector Machines is one of the best approaches to data modeling due to its accuracy and modeling that is based on Structural Risk Minimization from statistical learning theory. The principle behind this theory is to learn from finite data and control the generalization ability of learning machines to reduce the errors. The data is separated by a hyper plane in the n-dimensional space. The data which are closest to the Hyperplane are called support vectors which act as representatives. The Hyperplane that separates two groups of data is expressed in the objective function

\[ w \cdot x + b = 0 \]  \hspace{1cm} (1.12)

where \( x \) is a set of training vectors, \( w \) represents vectors perpendicular to the separating Hyperplane and \( b \) represents the offset parameters which allows for an increase in the margin. For n-dimensional data n-1 Hyperplane are to be introduced. A positive slack variable \( \varepsilon \) is introduced in the objective function in order to add some flexibility in separating the categories\[150\]. The improvised objective function is

\[ y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i \]  \hspace{1cm} (1.13)

where \( \varepsilon \geq 0 \). For non-linear classification kernel functions are used instead of every dot product. It maps the input space into a high dimensional space through a non-linear transformation. The kernel mapping provides a common base for most of the commonly employed model architectures, enabling comparisons to be performed \[21\]. Support Vector Machines require the solution for the objective function
\[ y_i(w \cdot \phi(x_i) + b) \geq 1 - \varepsilon_i \]  (1.14)

where \( x_i \) is set of training vectors and are mapped into a higher dimensional space by the kernel function \( \phi \). In this higher dimensional space, by estimating maximal margins, SVM finds a linear separating Hyperplane. The efficiency of the model depends on the selection of kernel, the kernel's parameters and margin parameter.

**Pitfalls in SVM Methods**

- When the training set increases there is a linear increase in the number of support vectors and it is dense.
- Predictions are not probabilistic.
- Time complexity due to the estimation of error by cross-validation technique.
- There exists a limited number of Kernel functions that are suitable.
- Complexity in Multiclass classification problems.
- Difficulty in the interpretation of parameters for a solved model.

Due to these drawbacks in SVM, Relevance Vector Machine is considered for this research. As expected the results are more promising.

**1.10.2 Relevance vector machine**

Relevance Vector Machines (RVM) is a probabilistic vector machine learning model that uses Bayesian inference to obtain parsimonious solutions for regression and probabilistic classification. The RVM has an identical functional form to the support vector machine, but provides probabilistic classification. Tipping [99] introduced the Relevance Vector Machine (RVM) which makes probabilistic predictions and yet retains the
excellent predictive performance of the support vector machine. RVM is better than SVM on several aspects such as sparseness and tuning. The number of relevance vectors are much smaller than support vectors which give better sparseness and the regularization parameter tuning during training phase is not needed. The main drawback observed is the involvement of highly nonlinear optimization process in training phase.

**Procedure of RVM**

Supervised learning techniques make use of a training set that consists of a set of sample input vectors \( \{x^n\}_{n=1}^N \) together with the corresponding targets \( \{t^n\}_{n=1}^N \). The targets are basically real values in regression tasks or class labels in classification problems. It is typically desired to learn a model of the dependency of the targets on the inputs from the training set, so that accurate predictions of it can be made for previously unseen values of \( x \). Commonly, these predictions can be based on some function \( y(x) \)
defined over the input space formed by a kernel function centered at the different training points

\[ y(X;w) + \epsilon = \phi w + \epsilon \]  

(1.15)

While this model is similar in form to the support vector machines, the kernel function here does not need to satisfy the Mercer’s condition, which requires a continuous symmetric kernel of a positive integral operator. Multi-kernel RVM is an extension of the simple RVM model. The simple procedure of RVM is illustrated in Figure 1.6.

1.11 Particle swarm optimization

Particle swarm optimization (PSO) is a population based stochastic optimization technique[75]. The motivation behind this technique is the social behavior of birds flocking or fish schooling. A group of birds randomly searching for food in an area where there is only one piece of food in the area being searched is a prime example. The technique is used to solve the optimization problems. In PSO, each single solution is a “bird” in the search space and is termed as a “particle”.

1.11.1 Procedure for Particle swarm optimization

All the particles have fitness values which are evaluated by the fitness function to be optimized, and have velocities which direct the flight of the particles. The particles fly through the problem space by following the current optimum particles through many iterations. In each iteration, the particle is updated by following the two “best” values. When a particle takes part of the population as its topological neighbors, the best value is a local best and is called lbest. The second “best” value, global best, also called as gbest, is tracked by the particle swarm optimizer, and is the best value obtained so far by any
in the population. PSO is used for optimization through its particle’s velocity and position change. The velocity of each particle in the swarm is updated using the following equation:

$$v_i^{k+1} = iwv_i^k + c_1 r_1 * (pbest_i - kx_i^k) + c_2 r_2 * (gbest - kx_i^k)$$  \hspace{1cm} (1.16)

where, \(v_i^k\) is the velocity of kernel parameters of agent \(i\) at iteration \(k\), \(iw\) is the inertia weight of PSO algorithm, \(c_1\) & \(c_2\) are weight factor of current particles to optimize local training samples and are user supplied coefficients, \(r_1\) & \(r_2\) are random values regenerated for each velocity update. Each of the three terms of the velocity update equation has a different role in the PSO algorithm. The first term is the inertia component, responsible for keeping the particle moving in the same direction it was originally heading. The second term \(c_1 r_1 * (pbest_i - kx_i^k)\) called the cognitive component, acts as the particle’s memory, causing it to tend to return to the regions of the search space in which it has experienced high individual fitness. The third term \(c_2 r_2 * (gbest - kx_i^k)\) called the social component, causes the particle to move to the best region the swarm has found so far.

The random values have a stochastic influence on the velocity update. This stochastic nature causes each particle to move in a semi-random manner, heavily influenced in the directions of the individual best solution of the particle and global best solution of the swarm. There are various advantages exists in PSO technique

- Only the most optimist particle transmits information into other particles and the search is carried out by the speed of the particle without mutation or any cumbersome activities.
- Optimization ability is more and the process completes easily with calculation.
Due to its advantages, simplicity and simple implementation, the algorithm can be used widely in the fields such as function optimization, the model classification, machine study, neutral network training, the signal procession, vague system control, automatic adaptation control etc.

1.12 Objectives and Scope

The primary objective of this research is to design a classification system which classifies a new user with the previous users group to help them in their browsing experience by finding desired information in a quick and accurate manner. To achieve this primary objective, the following sub-goals were formulated.

(i) To select relevant features from the user session matrix which consists of navigation patterns as rows and web pages as columns otherwise termed as features.

(ii) To design and develop a novel clustering methodology, that can be used to group users into similar groups.

(iii) To develop a better multi type classification model based on ensembling to classify a new user from the groups of navigation patterns.

(iv) To develop a model to fit in heterogeneous websites without the knowledge of web pages or site map.

1.13 Structure of Thesis

The present research work focuses on the navigation pattern analysis using an ensemble of relevance vector machine classifiers with Bolzano–Weierstrass based clustering technique. Before clustering the dimensionality reduction of the web log data through Quantum Skew Divergence based ICA is done.
The rest of this thesis is organized as follows. Chapter 2 reviews various works done previously in different phases of web log and navigation pattern analysis. In Chapter 3 we present in detail the proposed framework of the research. It discusses the preamble and the framework briefly. Chapter 4 demonstrates the three phases of proposed research work. It explains how features are selected with proposed algorithm in the first phase. It is shown how the clustering of navigation patterns takes place by the proposed method in the second phase. And it is also shown how classification is done using the proposed optimized methods in the third phase. Experiments are carried out and results are documented in chapter 5. Finally, Chapter 6 concludes this thesis and outlines directions for future work.