**CHAPTER 1**

**INTRODUCTION**

1. **INTRODUCTION TO WEB MINING**

Now a day’s Web is most certain branch of world and even web surfing is a very essential action for most of the users who will be online. Along with rapid development of World Wide Web, Information overload and confusion with information has becoming obstructions. Here we collect web log data and apply the data cleaning process and also identify the user based on IP address. We will apply clustering technique such as K-means and SOM algorithm to form a cluster. Based on their interest similar web page can be recommend to users and also owner can analyze his web site and improve his business. Web usage mining technique to get knowledge from web server log files where all user navigation history is registered. The important aspects in this are data exploration and analysis of activity and preferences of users.

**1.2 WEB USAGE MINING**

Web usage mining mainly absorbs the blind explore of user access prototypes from one or added web servers. Many companies depend upon internet to carry out their commerce and accumulate, big amounts of data in their day by day operations. This kind of information is produced usually by web servers and stored in server access log files. By this many companies can improve relationship between customer and manager by giving them accurately what customer need. Accordingly companies can recognize the requisites of the client and they can react to customer wants faster and as needed. They can also amplify effectiveness and productivity based on top of the outlines created depending on this. Web usage mining uses one among many data mining practices towards finding out of valuable access patterns from web server logs, whereas web log data is a proof of every URLs contacted by users on a particular web site. Each entry of web log file consists of access time, IP address, URL visited, REFERRER (Web page visited just prior to the current one), etc

Web Usage Mining (WUM) is an emerging field of research and is most likely to find out new knowledge in Internet-based business. WUM applications are also being used in some famous websites to understand customer’s profiles and their performance, accordingly to know strengths and weaknesses of their website. Here we represent a simple introduction of WUM lifecycle, a data mining technology and WUM implementation. The main aim is to understand clustering algorithms that can be worked on data.

**Source of data in Web Usage Mining**

**Web server logs**

Log file consists of registered information which includes about user request history (i.e. Data regarding the application, including client IP address, ask for date and time, which page appealed, HTTP code, bytes provided, user agent, and referrer are usually registered). This huge statistics can be merged into a particular file, or even separated into distinctive log files, such as a contact log, error log, or referrer log. Server logs usually do not typically assemble user specific information. These folders are not accessible to public Internet users, but accessible only to the webmaster or managers.

**Server Log file format**

The main Apache HTTP server will usually have two log layouts, as common log format and also combined log format. The component of the log entities are as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **VISITOR IP ADDR** | **Rfc 391** | **TIME and DURATION** | **STATUS CODE** | **VISITOR REFERRING WEBSITE** | **BROWSER TYPE** |
| 66.249.67.142 | - - | [09/Aug/2009:00:34:05 +0530] | 200 | "http://www.enggresources.com/classroom/" | Mozilla |

Fig1 : Server Log File Format

***Visitor IP Address****:* This piece of data provides the Internet Protocol (IP) address of theguests who visited the website in concern.

Eg : 66.249.67.142

***Rfc931*:** This kind of information is highly unreliable and should never be used excepton rigidly controlled internal networks. Hence is not available as identity of the client, which is represented as “hyphen” and can be determined by the client machine.

Eg: - -

***Time and duration***: This showsthe time and period meant for how lengthy the website was accessed by an individual user.

Eg: [09/Aug/2009:00:34:05 +0530]

***Status code*:** This is the status code which is sent by server to the user. Thisinformation is very precious, as it reveals whether the request resulted in a winning response, and in an indicated redirection, or whether an error sourced by the client, or an error in the server.

Eg : 200 - successful

***Visitor Referring Website:***The submitting website offers the information or URL of thewebsite that was referred by that website in deliberation.

Eg: “http://www.enggresources.com/classroom/"

***Browser Type:***This gives the data of the particular form of browser that was exercised forcontacting the website.

Eg: Mozilla

***Cookies:***It is a piece of information provided by server to user.

**1.3 WUM PROCESS**

WUM is a process similar to exploitation in the data mining with different data source types and tools used. Three phases that are included in Web usage mining starts from preprocessing, pattern discovery and then pattern analysis. In the initial preprocessing phase, information contained in a data source is converted into data abstraction. In the second phase, various algorithms and techniques can be applied for pattern discovery. Patterns and rules generated in second phase are analyzed in the third as in order to filter out uninteresting rules.

**Data Collection**

Within this phase, usage data is collected from diverse resources like web servers, by clients connected to a server, or from middle sources such as proxy servers and packet sniffers.

Data of a typical web server is as given in fig (1), and following data is a sample data of the first row of the log file: 202.83.50.209 - - [26/Jul/2009:00:38:37 +0530] "GET /images/bgmain.gif HTTP/1.1" 200 144 "http://www.enggresources.com/profile.php?puid=8907" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; GTB6; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) ; .NET CLR 2.0.50727)"

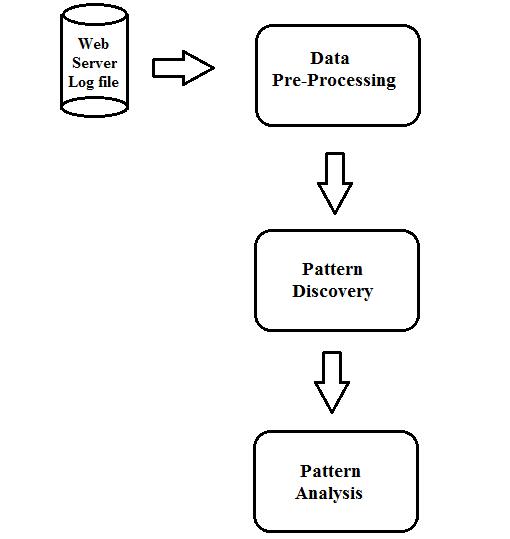


Fig2: WUM process

**Data Preprocessing**

Data preprocessing mainly involves various types of procedures completed on raw data to organize it for a further processing method. Commonly it is used as a groundwork data mining practice. Data preprocessing renovates the data interested in a layout that will be extra effortlessly and proficiently processed further by the client. This is the phase where data are uncontaminated from noise, their inconsistencies are determined, and they are integrated and combined in order to be used as key in to the further stage of pattern detection. The methods which are used here can provide client a data expansion. The different jobs of data preprocessing are:

**Data Cleaning***:*

The first phase in data preprocessing is to dirt free the raw web data. During this step the existing data are examined and inappropriate or redundant items are eliminated from the dataset. Inappropriate records are deleted throughout data cleaning. Since the objective of WUM is to get required pattern. Below are two kinds of redundant records which should be removed:

* The records with filenames extension of GIF, JPEG, CSS.
* The records with noisy data or uncompleted queries are removed.

**User Identification**:

The recognition of each user who accesses a website is one of the most essentialissues for the victory of a personalized website. The simplest advance is to assign a different user to each special IP identified in the log file. Cookies are also useful for identifying the guests of a website by storing an ID, which is created by the web server for each user visiting the website.

**Session Identification**:

The identification of customer sessions also is given significant attention in WUM process, as sessions instruct the navigational actions of the users and they are most important for prototype discovery. A client session is a restricted deposit of pages visited by the identical user contained by the extent of one exacting visit to a web site.

**Pattern Discovery**

In this phase, required information is determined by classifying users according to their navigational activities. The purpose of classification is to identify the differentiating distinctiveness of predefined groups, based on a position of instances, e.g. users of each class. Categorization is the procedure to plan a data entry into one of numerous predefined divisions. This involves withdrawal and collection of aspects that best explains its properties of a agreed class or type. The task for realizing the patterns offer some techniques as numerical analysis, association regulations, in order pattern analysis, clustering and so on.

The intention of clustering algorithm is to organize a assortment of data objects in to clusters, such objects within a cluster are more associated to each other then they are in other clusters. In our approach we make use of Self Organizing Map algorithm for cluster optimization and are also presented to personalize web Page cluster of end users. Similarly we make use of k-means algorithm to frame the clusters.

**K-means clustering technique**

K-means is another such well known and most used clustering method. Usually K-means is used to cluster very large set of data. Clustering is nothing but dividing complete data into groups as clusters where data objects in each cluster will have some similarity. K-means is used by many because of its very fast convergence property. K-means is one among the numerical and iterative method hence very simple and fast. Here initially it calculates the space between each of data object and group center in every iteration. So it is to be proved that K-means is a effective clustering method and suitable to produce globular clusters. As K-means is partitioning and iterative method the clusters formed will be independent and compact. There will be random selection of K centers and next starts with taking each data object which is nearest to that. And usually Euclidean distance is considered to find out distance between each data object and cluster center. So in this way similar data objects are brought together to form clusters. And this will be continued until the criterion becomes small.

**Self Organizing Maps**

SOM is used to cluster the data based on similarity along with an idea of assigning the same number of instances to each session. SOM is one of the neural network algorithms which not only cluster the data but also reduces the dimensionality of data. Here SOM makes use of weight matrix that is obtained from previous steps. Weight matrix includes weights calculated for each session. So here we illustrate the feasibility of using Self Organizing Map to mine the web log data of user navigation. This helps web masters to better understand the characteristics along with navigational behaviors of web users who visit their pages. SOM is one of the effective algorithms. Next is to analyze the pattern that is nothing but visualizing the clusters formed by SOM as well as K-means clustering methods.

**Pattern Analysis**

This is the concluding step in the WUM process. Following the preprocessing and pattern detection, the attained usage patterns are analyzed to clean uninteresting information and take out the useful information. High-level data mining uncertainty languages need to be developed to permit users to describe data mining jobs by facilitating the pattern of the relevant sets of data for scrutiny, the domain knowledge, the sorts of knowledge to be mined, and the clauses and constraints to be imposed on the discovered pattern.

Pattern analysis enables us to do the automatic recognition of patterns in data as of the same source and make calculations of new data coming from the similar source. Pattern analysis can be done here using java swings software which is user friendly and also simple to use.

**Visualization of clusters**

As we are breathing in a data-driven period with fast rising technologies, there survives a bulk of data produced, captured and also preserved for a range of purposes. So there is a call for of efficient, sufficient tools to visualize the data and dig out knowledge from them. This is feasible with different visual analytic methods.

Incoming into visual analytics, that maintains analytical reasoning by interactive visual boundaries. Visual analytics deals with capabilities of human visual information gathering and also abundant processing power of computers. Visual analytics are applied to the chore of transforming data addicted to valid information and finally to produce the awareness. Many visual analytic systems combine both interactive visualization and also computational analysis. This helps for a complete understanding and decision making in very complex and huge data sets. Visual analytics contain various tools and techniques that help user to extract efficient knowledge from complex data and communicate effectively.

**Visualization tool used**

**JFree Chart**: JfreeChart is an release source library extended in Java. And also it is at no cost complete Java free chart library which is easy for developing team to display qualified class charts in their functions. JFreeChart's characteristic deposit consists of:

* A regular and fine-certificated API, sustaining a wide sort of graph types.
* A stretchy propose that is simple to expand, and aims together server-side and client-side purposes.
* It is a shore up for numerous output types, including Swing and JavaFX workings, image files (PNG and JPEG), and vector graphics file formats (PDF, EPS and SVG).
* JFree Chart is an unwrap source or, more exactly at no cost software. It is scattered beneath the requisites of the GNU Lesser General Public License (LGPL), allows use in proprietary purposes.

And here we make utilize of this useful tool to compare two well known clustering algorithms. Particularly we used pie chart and also line chart to analyze the results of cluster formation by both algorithms. And hence we can also have clear comparison between them.

**Line chart:** A line chart or line diagram displays information as a sequence of data positions or markers united by straight line sectors. Line diagram gives you an idea about how the data changes at equal time incidence.  They are destined to evaluate two disconnect variables and later these variables are planted on an axis. In the end, we will get a chart with a line that starts from a unchanging point across a chart going up and doing in relative to data.

**Pie chart:** A pie diagram which is otherwise called a pie outline is a kind of chart generally utilized as a part of blend with rates. An extensive circle is isolated into divisions relying upon those rates and every segment speaks to a piece of the entirety. In a pie outline, the circular segment length of every part division is intended to be corresponding to the rate it should speak to.

**1.4 OBJECTIVE OF THE PROJECT**

The foremost objective of this project is to know the clear comparison between two widely used clustering methods such as Self organizing maps and k-means clustering algorithms.

**1.5 Organization Of Project Work**

**Chapter 1: Introduction:** This chapter tells about the introduction to Web mining along with its types and mainly about Web usage mining process. And even the Motivation, Objectives of the project.

**Chapter 2: Literature Survey:** This chapter gives information about the Background of Web Usage mining, many applications of clustering methods such as SOM and k-means. And even about the importance of visualization in many areas.

**Chapter 3: System Requirement Specification:** This chapter speaks about user characteristics, software and hardware conditions, functional requirements and non-functional prerequisites and feasibility revise.

**Chapter 4: System Design:** This chapter deals with the architecture of the system and complete run of the project is signified by professional data flow diagram.

**Chapter 5: Implementation:** This chapter contracts with the strides involved in the making of the task exertion. It is characterized through the module portrayal and calculation.

**Chapter 6: Testing:** This chapter principally contracts with the differed sorts of the experiments to affirm the quality of the undertaking.

**Chapter 7: Result analysis and comparison:** This chapter deals analyzing of results from both clustering methods and later have clear comparison between both.

**Chapter 8: Conclusion and Future Work:** This chapter fundamentally the rundown of the whole venture and it additionally recommends a portion of the improvement thought which couldn't be concealed because of imperative of time and assets.

**CHAPTER 2**

**LITERATURE SURVEY**

Literature survey is primarily completed with a specific end goal to investigate the foundation of the present scheme which serves to figure out defects in the current framework & guides on which unsolved issues we can work out. Along these lines, the accompanying subjects delineate the foundation of the undertaking as well as reveal the issues and blemishes which propelled to propose arrangements and chip away at this venture.

**2.1 Literature survey**

There are several existing work related to web mining, web usage mining. Due to some drawbacks existing technologies cannot be implemented everywhere. Web usage mining supports clustering based on user behavior .There are several types of clustering mechanisms in existing system.

Due to continuous growth of users as well as size of shared documents on web site there is necessary of automatic adjusting of web contents. To meet the expectations of users, we can make use of server log files. And mining of such related data with accuracy leads to discovery of web user navigational data. Weichbroth and Owoc provide skeleton for mining of web navigational data models in order to know the running of web usage [1].

Some research helps to know the capacity of web usage mining on e-education domain. Seal et.al utilized of educational mining of data practices to analyze learner actions. Preprocessing is considered to be crucial part of data mining process. Main point is to develop a data preprocess method functional to model logs based on SCORM. They define various static variables along with SCORM content tree and also multidimensional graphics [2].

Web usage mining has conventionally stressed on individual queries and even queries which leads to web sites or web page stay. Here Hollink et.al tried to characterize web sites in terms of meaning of queries that guide to know about large data sets on web. So they demonstrated exploitation of such links for effective mining of data along with how patterns can be used effectively [3].

Web mining is nothing but discovering and extracting of helpful information from World Wide Web. This engages automatic detection of data objects from web servers. This mainly assists many organizations to know the rate of each customer and even marketing strategies with effectiveness of promotional acts. Suresh, Padmajavalli discussed significance of data preprocessing was along with required steps [4].

In this they evaluated the feasibility of using SOM to extract web navigational log data and also provided a device to visualize the user navigation. Kate et.al extended a LOGSOM system that uses SOM to arrange the web pages into two dimensional maps. And is mainly based on users navigational behavior. Hence the resulting map also helps web masters to better understand behavior of web users as being a visual analysis tool [5].

Now a day’s data clustering has been most important application issue in data mining. SOM is one such widely used to cluster the multivariate data. In SOM we have basic assumption on data density and unsupervised learning. SOM works by analyzing neighbor neurons and their corresponding activities to form the geometry of clusters. Here Alfredo et.al also discussed issues related to SOM clustering and also filtering, watershed transforms. Proposed new gradient based visualization [6].

SOM clustering algorithms can also be used in Map Reduce using Spark platform, where this is a popular programming environment to process large data sets. SOM is mostly used as it is unsupervised learning and also useful for clustering large data sets. Shi Na et.al also reported on performance of SOM in Map Reduce in terms of classification of accuracy and also speed up using 100 million points different cores [7].

Cluster analysis is one of the core parts in data mining. Always the way of clustering influences the outcomes thoroughly. This document talks about typical k-means clustering and also evaluates its short comings. As it needs to calculate the space between cluster center and data pattern its efficiency is not high. So Zhu et.al proposed new improved k-means to avoid this and which stores data in every iteration and be able to use in subsequent iteration [8].

K-means is one of the fast convergent methods but will always have basic condition. So they proposed another novel clustering along with k-means as Affinity Propagation (AP). produces a fine set of clusters with rapid speed. Taking these both properties of k-means and Affinity Propagation method Boutsids et.al presented new clustering strategy which provides less errors than AP and k-means [9].

**CHAPTER 3**

**SYSTEMS REQUIREMENTS**

Software requirement condition is a key file, which outlines the groundwork of the item headway practice. It records the necessities of a structure and additionally has a depiction of its genuine segment.

**3.1 REQUIREMENT ANALYSIS**

Requirement analysis is the important part of project. Once the requirements are known then we can provide solution for so far problems that have occurred in existing system. This contains all the duties that go into investigation, scoping and explanation of new or changed system

**3.2 FUNCTIONAL REQUIREMENTS**

A functional prerequisite describes a function of a software scheme and provides system behavior with specific functionalities or inputs. Based on the proposed problem definition following are the functional requirements.

* Based on the previous navigational patterns of the user, he needs to be provided with suitable recommendations.

Next link clicked by the user should be predicted accurately so that access delay of the next web page can be decreased significantly.

**3.3 NON-FUNCTIONAL REQUIREMENTS**

**Usability**

This project can be used to analyze the websites in efficient manner compared to others hence it is user friendly.

**Scalability**

In this project we are using a web usage log file of 40 mb, it can be extended for large files.

**Reliability**

The highest viewed link information generated by the k-means and SOM algorithms are reliable.

**Adaptability**

This project is very user friendly and hence we can alter the project to get any other related information.

**3.4 HARDWARE REQUIREMENTS**

The hardware we are utilizing for this project is as follows

* + Processor : Intel core I3
  + RAM : 4GB
  + Hard Disk : 500GB

**3.5 SOFTWARE REQUIREMENTS**

The software we are using in this project is as follows

* + Operating System : Windows 7/8
  + Technologies used : Java
  + JDK : Version 1.7
  + Eclipse : Luna

**3.6 RESOURCE REQUIREMENT**

**ECLIPSE IDE**

Eclipse IDE is one of the greatest tool that is widely used in present situation. At present situation, even to build a small application, developers should and must use Eclipse

Versions of eclipse are

1. Galileo
2. Ganymede
3. Helios
4. Indigo
5. Juno
6. Kepler
7. Luna

Among these versions Galileo is first version and Luna is the latest version. In this project I have used Luna version.

Important features of eclipse are

* **Technology Independent:** We can use any coding language like c, c++, java, php, .net, C# etc.
* **User Friendly:** We can use this application very easily. For example if we get some exception while writing code then just hover on that particular line then it immediately resolves it by surrounding try catch block.
* **Graphical User Interface:** All the functionality has been included well in advance and presented to user using GUI instead of writing commands.
* **Time efficient:** It executes the code as efficiently as other application such as edit plus, notepad plus but reduces the time while writing compiling and executing commands.
* **Built-in Compiler:** As soon as we write a program it automatically compiles the program and after clicking run it immediately executes and shows errors on console if there is some syntax error.
* **Separate Class files:** It maintains separate subfolder for maintain all class file generated after compiling the code, this separate folder helps Java Virtual Machine(JVM) to identify classes easily.
* **Easier Execution:** No need to write commands for execution, instead just click on run to execute the program.
* **Easier Management:** It internally manages all JVM structure and maintains a separate folder for each type of files.
* **Identify errors:** As we execute the compiled program, it displays errors along with exception type in server console.
* **Libraries:** Eclipse contains all built-in libraries in the form of jar file in JRE-Library folder.
* **Easier to write program:** In eclipse it is easy to write program because for example if we are writing a implementation class for some abstract method defined in super class then just hover on methods and select override the abstract method.
* **Import:** Importing package, class, interface, abstraction class and even other driver class jar files are easy.

**JFree Chart**:

The JFreeChart undertaking was started fourteen years prior, in February 2000, by David Gilbert. Today, JFreeChart is the most broadly utilized outlining library among Java engineers.

JfreeChart is a release source library extended in Java. And also it is at no cost complete Java free chart library which is easy for developing team to display qualified class charts in their functions. JFreeChart's characteristic deposit consists of:

* A regular and fine-certificated API, sustaining an ample sort of graph types.
* A stretchy propose that is simple to expand, and aims together server-side and client-side purposes.
* It is a shore up for numerous yield sorts, including Swing and JavaFX workings, picture documents (PNG and JPEG), and vector design record groups (PDF, EPS and SVG).
* JFree Chart is an unwrap source or, all the more precisely at no expense programming. It is scattered underneath the necessities of the GNU Lesser General Public License (LGPL), permits use in exclusive purpose.

The JFreeChart is famous for its efficient chart creation, and user-friendly installation setup.

And here we make use of this useful tool to compare two well known clustering algorithms. Particularly we used pie chart and also line chart to analyze the results of cluster formation by both algorithms. And hence we can also have clear comparison between them.

**Line chart:** A line chart or line diagram displays information as a sequence of data positions or markers united by straight line sectors. Line diagram gives you an idea about how the data changes at equal time incidence.  They are destined to evaluate two disconnect variables and later these variables are planted on an axis. In the end, we will get a chart with a line that starts from a unchanging point across a chart going up and doing in relative to data.

**Pie chart**: A pie diagram (or a circle outline) is a roundabout measurable chart, which is further separated into fragments to point up numerical extent. In a pie graph, the curve length of every cut (and likewise its focal point and zone), is relative to the limit it speaks to. While it is known for its likeness to a pie which has been fragmented, there are varieties on the mode it can be introduced. A pie diagram which is additionally known as a pie outline is a kind of chart routinely utilized as a part of mix with rates. A major circle is isolated into segments relying upon those rates and every portion speaks to some piece of the entirety. In a pie graph, the curve length of every portion is intended to be relative to the rate it should speak to.

**CHAPTER 4**

**SYSTEM DESIGN**

System Design is an answer for how to way to deal with the making of a framework. This vital stage gives the understanding and procedural points of interest fundamental for executing the framework suggested in the attainability study. The configuration step delivers an information plan, a structural outline and a procedural outline. The information outline changes the data space model made amid investigation into the information structures that will be obliged to execute the product. The compositional configuration characterizes the relationship among major auxiliary segments into a procedural depiction of the product. Source code created and testing is directed to coordinate and accept the product.

From an undertaking administration perspective, programming configuration is led in two stages. Preparatory outline is associated with the change of prerequisites into information and programming structural engineering. Configuration concentrates on refinements to the compositional representation that prompts point by point information structure and algorithmic representations of programming.

**Logical Design:**

The logical plan of an information arrangement is equivalent to an engineering proposal or conceptual view of an automotive. It explains the most important attributes and how they are related to one more. The outputs, inputs and relationship between the variables are designed in this phase. The purposes of database are exactness; integrity and winning recover from failure, isolation and security of data and good general performance.

**Input Design:**

The input proposes is the connection between users and the information scheme. It specifies the approach in which data go into the system for processing. It can guarantee the reliability of the system and construct reports from accurate date or it may effect in the output of error information.

Online data entry is available which accepts input from the keyboard and data is displayed on the screen for verification. While designing the following points have been taken into consideration. Input formats are designed as per the user requirements.

a) Interaction with the user is maintained in simple dialogues.

b) Appropriate fields are locked thereby allowing only valid inputs.

**Output Design:**

Each and every activity in this work is result-oriented. The mainly important feature of information structure for users is the output. Efficient intelligent output design improves the usability and acceptability of the system and also helps in decision-making. Thus the following points are considered during output design.

* What information to be present?
* Whether to display or print the information?
* How to arrange the information in an acceptable format?
* How the status has to be maintained each and every time?
* How to distribute the outputs to the recipients?

The system being user friendly in nature is served to fulfill the requirements of the users; suitable screen designs are made and produced to the user for refinements. The main requirement for the user is the retrieval information related to a particular user.

**Data Design:**

Data plan is the first of the three design actions that are carried out through software engineering. The collision of data formation on program structure and practical difficulty causes data design to contain a reflective influence on software value. The concepts of information thrashing and data generalization provide the establishment for an advance to data design.

**4.1 Fundamental Design Concepts**

**Abstraction:**

During the software devise, abstraction tolerates us to arrange and channel our process by rearranging structural reflections until the functional attributes; data streams and data stores have been recognized Data abstraction involves specifying legal operations on objects; representations and manipulations details are suppressed.

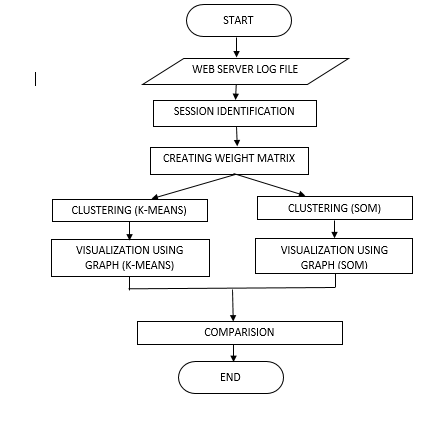
**Information Hiding:**

Information hiding is an essential design notion for software. When software scheme is proposed using the information hiding advance, every module in the scheme hides the inner elements if the processing behavior and modules communicating just through well-defined lines. Information hiding can also be used as the primary design practice for architectural sketch of a system.

**4.2 Problem Definition**

To have comparison between two most widely used clustering algorithms such as Self organizing map and k-means. By discovering user and sessions from the web server log file and constructing a hyper graph using session details, identifying homogeneous groups with the help of these clustering technique. And later to visualize the clusters formed.

* 1. **Proposed Architecture**

****Fig3: Architectural overview of the system

**4.4 System outline**

The system consists of three phases. They are:

* Preprocessing
* Clustering using SOM and K-Means algorithms
* Visualization

Each of the phases is discussed in detail in the following sections.

**Pre-processing**

Data Preprocessing is the first key connection of a Web log mining. Information preprocessing results will have an immediate effect on the behind issues led by the exchange distinguishing proof, the way examination, the model established. The standard of preprocessing is to change over the Weblog into reliable, complete and precise information supply, to meet the needs of Web information mining calculation usage procedure is required. On the other hand, as Web log record organization is not at all like from the conventional feeling of database or information distribution center which has a decent information structure, furthermore semi-organized, together with the presence of a mixed bag of grounds to the information in the logs fragmented. This has endeavored of pretreatment face numerous down to earth issues. Furthermore, information pre-handling has turned into the hardest undertaking in the Web information mining. The procedure of information pre-handling first includes multi-server log information accumulation and solidification, and afterward tidies up the log record to evacuate the photos, sound documents, multi-page system for the page, Web robots hunt logs, et cetera. At last, pre-treatment information is stacked into the information distribution center. The preprocessing has four stages such as

**Data Cleaning**: Removing the log entry that is not needed for the mining practice.

**User Identification:** Unique users are identified.

**Session Identification:** Segregates the page requests of all user into individual session

**Session Page View Matrix**: Creation of the session page view matrix, in the rows represents the sessions and columns the pages.

**CLUSTERING METHODS**

Clustering is a procedure of categorization of objects into dissimilar crowds or the screening of a data position into [subsets](http://en.wikipedia.org/wiki/Subset) (clusters), so that the data in each subset (perfectly) allocate various familiar trait - frequently according to several defined [distance measure](http://en.wikipedia.org/wiki/Metric_(mathematics)).

**K-means ALGORITHM**

The K-means algorithm involves randomly selecting K initial centroids where K is a user defined number of desired clusters. Each point is then assigned to a closest centroid and the collection of points close to a centroid form a cluster.

K-means is another such well identified and most used clustering process. Traditionally k-means is the most widely used algorithm for grouping items into sets with similar attributes. The k-means technique has been exposed to be useful in producing fine clustering results for several practical purposes. K-means clustering is a way familiarly used to repeatedly dividing the data set into *k* clusters. The input given to k-means algorithm is the session page-view matrix.

Since k-means algorithm is hard clustering, it classifies the data points into one and only one cluster. Hence if any of the data point is equally close to two of the cluster centers, k-means algorithm is forced to put that point into one particular cluster. This is one of the major disadvantages of k-means algorithm.

Working of K-means method:

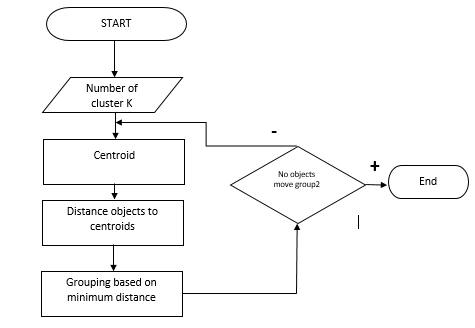


Fig4: K-means algorithm working

Usually K-means is used to cluster very large set of data. Clustering is nothing but dividing complete data into groups as clusters where data objects in each cluster will have some similarity. K-means is used by due to its rapid union property. K-means is one among the numerical and iterative method hence very simple and fast. Here initially it calculates the space linking each of data object and cluster center in every step. So it is to be proved that K-means is a effective clustering method and suitable to produce globular clusters. As K-means is partitioning and iterative method the clusters formed will be independent and compact. There will be random selection of K centers and next starts with taking each data object which is nearest to that. And usually Euclidean distance is regarded as to find out distance among both data entity and cluster center. So in this way similar data objects are brought together to form clusters. And this will be continued until the criterion becomes small.

# K-Means: Step-By-Step Example

As a simple illustration of a k-means algorithm, consider the following data set consisting of the scores of two variables on each of seven individuals:

|  |  |  |
| --- | --- | --- |
| Subject | A | B |
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

This data set is to be grouped into two clusters.  As a first step in finding a sensible initial partition, let the A & B values of the two individuals furthest apart (using the Euclidean distance measure), define the initial cluster means, giving:

|  |  |  |
| --- | --- | --- |
|  | Individual | Mean Vector (centroid) |
| Group 1 | 1 | (1.0, 1.0) |
| Group 2 | 4 | (5.0, 7.0) |

The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean. The mean vector is recalculated each time a new member is added. This leads to the following series of steps:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cluster 1 | | Cluster 2 | |
| Step | Individual | Mean Vector (centroid) | Individual | Mean Vector (centroid) |
| 1 | 1 | (1.0, 1.0) | 4 | (5.0, 7.0) |
| 2 | 1, 2 | (1.2, 1.5) | 4 | (5.0, 7.0) |
| 3 | 1, 2, 3 | (1.8, 2.3) | 4 | (5.0, 7.0) |
| 4 | 1, 2, 3 | (1.8, 2.3) | 4, 5 | (4.2, 6.0) |
| 5 | 1, 2, 3 | (1.8, 2.3) | 4, 5, 6 | (4.3, 5.7) |
| 6 | 1, 2, 3 | (1.8, 2.3) | 4, 5, 6, 7 | (4.1, 5.4) |

Now the initial partition has changed, and the two clusters at this stage having the following characteristics:

|  |  |  |
| --- | --- | --- |
|  | Individual | Mean Vector (centroid) |
| Cluster 1 | 1, 2, 3 | (1.8, 2.3) |
| Cluster 2 | 4, 5, 6, 7 | (4.1, 5.4) |

But we cannot yet be sure that each individual has been assigned to the right cluster.  So, we compare each individual’s distance to its own cluster mean and to  
that of the opposite cluster. And we find:

|  |  |  |
| --- | --- | --- |
| Individual | Distance to mean (centroid) of Cluster 1 | Distance to mean (centroid) of Cluster 2 |
| 1 | 1.5 | 5.4 |
| 2 | 0.4 | 4.3 |
| 3 | 2.1 | 1.8 |
| 4 | 5.7 | 1.8 |
| 5 | 3.2 | 0.7 |
| 6 | 3.8 | 0.6 |
| 7 | 2.8 | 1.1 |

Only individual 3 is nearer to the mean of the opposite cluster (Cluster 2) than its own (Cluster 1).  In other words, each individual's distance to its own cluster mean should be smaller that the distance to the other cluster's mean (which is not the case with individual 3).  Thus, individual 3 is relocated to Cluster 2 resulting in the new partition:

|  |  |  |
| --- | --- | --- |
|  | Individual | Mean Vector (centroid) |
| Cluster 1 | 1, 2 | (1.3, 1.5) |
| Cluster 2 | 3, 4, 5, 6, 7 | (3.9, 5.1) |

The iterative relocation would now continue from this new partition until no more relocations occur.  However, in this example each individual is now nearer its own cluster mean than that of the other cluster and the iteration stops, choosing the latest partitioning as the final cluster solution.

Also, it is possible that the k-means algorithm won't find a final solution.  In this case it would be a good idea to consider stopping the algorithm after a pre-chosen maximum of iterations.

C:\Documents and Settings\Student\My Documents\My Pictures\gg.bmp

SOM is used to group the data based on similarity and along with an idea of assigning the same number of instances to each session. SOM is one of the neural network algorithms which not only cluster the data but also reduces the dimensionality of data. Here SOM makes use of weight matrix that is obtained from previous steps. Weight matrix includes weights calculated for each session. So here we illustrate the possibility of using Self Organizing Map to mine the web record data of user navigation. This helps web masters to understand the characteristics along with navigational activities of web users who visit their pages. SOM is one of the effective algorithms.

Next is to analyze the pattern that is nothing but visualizing the clusters formed by SOM. And this is done by using visual studio and through line chart and pie chart for both algorithms. Later we will have clear comparison between both algorithms.

**VISUALIZATION OF CLUSTERS**

As we are in a data-driven period with fast raising technologies, there subsists a bulk of data generated, confined and also maintained for a variety of reasons.

So there is a need of capable, adequate devices to visualize the data and take out knowledge from them. This is possible with diverse visual analytic techniques.

Entering into visual analytic that holds analytical causing by interactive visual borders. Visual analytics deals with capabilities of human visual information gathering and also abundant processing power of computers. Visual analytics are applied to the duty of altering data into valid information and closing to synthesize the understanding. Many visual analytic systems combine both interactive visualization and also computational analysis. This helps for a complete understanding and decision making in very complex and huge data sets. Visual analytics contain various tools and techniques that help user to extract efficient knowledge from complex data and communicate effectively.

And here we make use of this useful tool to compare two well known clustering algorithms. Particularly we used pie chart and also line chart to analyze the results of cluster formation by both algorithms. And hence we can also have clear comparison between them.

**Line chart:** A line chart or line diagram displays information as a sequence of data positions or markers united by straight line sectors. Line diagram gives you an idea about how the data changes at equal time incidence.  They are destined to evaluate two disconnect variables and later these variables are planted on an axis. In the end, we will get a chart with a line that starts from a unchanging point across a chart going up and doing in relative to data.

**Pie chart**: A pie diagram (or a circle outline) is a roundabout measurable chart, which is further separated into fragments to point up numerical extent. In a pie graph, the curve length of every cut (and likewise its focal point and zone), is relative to the limit it speaks to. While it is known for its likeness to a pie which has been fragmented, there are varieties on the mode it can be introduced. A pie diagram which is additionally known as a pie outline is a kind of chart routinely utilized as a part of mix with rates. A major circle is isolated into segments relying upon those rates and every portion speaks to some piece of the entirety. In a pie graph, the curve length of every portion is intended to be relative to the rate it should speak to.

**CHAPTER 5**

**IMPLEMENTATION**

The implementation is a core or heart of the project. In this real world, as there are no species without heart, similarly without an implementation there is no use of developing the system.

Implementation means writing the program in order to work the system as per the requirement given by the customers. So in order to write a program we need to select a right platform and technology. The important thing to be noted is, the platform and technology we use to implement the system should be able to understand by each and every user. So let us try to know the appropriate the platform and technologies used in our project.

**5.1 PLATFORM USED**

The platform indicates the operating system being used. Nowadays there many operating system being used but maximum people used to prefer Windows because of its features. Today most of the people refer windows XP, windows 7 and windows 8. In our project we are going to use Windows 7 as a operating system.

**5.2 LANGUAGE SELECTION**

Technology means the language which is used to write a program. At present in IT industry most widely used language by developers is “JAVA”. There are certain features in java which cannot be afforded by any other language. Hence it used by maximum people. So in our project we have used Java.

Let us discuss in brief about all the language used and needs of software.

**JAVA**

At present java is father of all languages. Java is said to pure 100% programming language because each and every element is written within the class and can be accessed through the object. The object is real world entity which has state and behavior and in order to design the object in particular structure we need a class. In simple we can say class as a logical component and object as a physical component.

**Features of java**

* **Easy to write code:** As compared to c and c++ writing the java code especially in eclipse is easier.
* **Easy to find and fix errors:** Along with writing the code, debugging is also done along with the code itself.
* **Code reusability:** Java follows a slogan called “write once execute many times” means we can use the same code many times as and when it is needed.
* **Easier execution:** In eclipse there is simple method start button, on clicking that button it starts to execute the program.

Externally also we can execute the program, for example in command prompt we need to type the command

javac filename. java //for compilation

java filename //for execution

* **Easier development method:** Developing the application using java in eclipse is easier because it separates all class and executable files, which helps the jvm to easily identify class files.

**Advantages of using java**

Using java we can achieve or we can say advantages of java are

* **Abstraction:** Abstraction is a process in which we hide the internal unnecessary implementation but show the necessary implementation. For example consider the projector, in this case we are just interested in output of the projector means the contents it displays but we are least bothered about the how exactly internal mechanism is taking place to display the content. Hence it is abstraction.

Using abstraction we can achieve loose coupling, it is technique where change of code in implementation method never affects the end user.

* **Polymorphism:** Polymorphism is process where we one method call for showing multiple behavior at different situation. An example for polymorphism is playing cards game.

Here we have two kinds of polymorphism

Compile Time polymorphism and Runtime polymorphism

Compile time polymorphism is technology during which call to overloaded method is decided at compile time based on parameters passed to method.

Runtime polymorphism is a technique during which call to dominated method is resolved at runtime based on instance type

* **Encapsulation:** Encapsulation is wrapping code and data into single unit or we can define the technique of making the data elements as private and providing public ways to access the data elements.

**5.3 Data pre-processing**

**Data Cleaning**

In the data cleaning process, it erases the information which has nothing to do with the mining in the Web log. Client access log record contains a large number of the unessential or repetitive that does not have any effect on removal examination. These insignificant or excess records should be centered around expelling from client access records. In a medium-sized Web server, access log documents consistently can without much of a stretch achieve several megabytes. In the event that there is not the introductory tidy up work, investigation will prompt a moderate and wasteful procedure. Information cleaning is essential, on the grounds that just this log information that has the capacity precisely mirror the examples of client access can be valuable to investigate the exactness of the learning, get the model and the outcomes significant. Owing to stateless or connectionless character of the HTTP protocol, a user’s appeal to browse a page effects in numerous log entries while graphics and other drafts are downloaded beside with HTML file content. While, the main intention of web usage mining is on the way to have a plain portrait of the web user activities; hence removal is needed for files having the suffixes such as, jpeg. Jpg, gif, css, cgi, etc. Error rules are not applicable and not useful for mining. The status can be checked and they can be eliminated, if found irrelevant. There are four dissimilar categories of status codes –Success (series starts from 200), Redirect (series starts from300), Failure (series starts from 400), Server Error (series starts from 500). From these, all error codes can be eliminated say, 401 (failed authentication), 404 (file not found) which are not required for analysis process and they are cleaned from the logs. There is a call for to eliminate at least some of the data grounds using this cleaning process. The logs containing the keywords “Google” and “yahoo” are also removed. The URLs in the logs are represented in the standard format and the query string removed. Machines and web spider navigation models must be explicitly identified. Web robots are software plans that inspect a Web site to take out its comfortable. Spiders regularly trail all the hyperlinks starting a Web page. To remove robots, we can look in support of all hosts that have demanded the page “robots.txt”, which is confirmed by robot though browsing. Unique links are identified and indexed.

**User Identification**

User identification covenants with identifying unique clients to web server. Web client recognizable proof is a standout amongst the most difficult strides during the time spent web use mining. Several troubles must be succeeding, for example, customer side reserving, changing and shared IP locations etc. If there should be an occurrence of straightforward business sector wicker container examination, the client is recognized precisely by its client ID. Nonetheless, if there should be an occurrence of web clients, it is not inconsequential which page downloads have a place with which client. The same individual can utilize various PCs, and more persons can utilize the same PC also. Moreover, intermediary servers can stow away applicable data about one of a kind clients as various PCs show up on the web utilizing the same IP address through the intermediary server. For client recognizable proof a vital issue is the way precisely the clients must be recognized. It depends for the most part resting on the assignment for the mining procedure is accomplished. In specific cases the clients are distinguished just through their IP addresses. This will be able to give a satisfactory result to concise time periods (minutes or hours) or while the normal results from the information mining errand needn't bother with all the more decisively data about the one of a kind web clients. For instance if there should arise an occurrence of selecting much of the time went to pages for server side reserving, or repacking the following page of normal navigational ways, it is unessential, whether a page is left by two distinct people or by one human being twice. The input point is that the page is gone to twice. Nonetheless, if there should be an occurrence of a commercial, it is vital, whether two extraordinary individual has seen the page or not. In different cases a few heuristics are utilized for better distinguishing proof of the clients. In the distinctive techniques are assembled into two divisions, one is the class of the proactive systems and the added is that of the responsive strategies. Proactive procedures go for separating the clients before requesting the page solicitation while receptive systems endeavor to partner people with the log sections after the log is composed. Proactive methods can be straightforward client validation with structures, utilizing treats or utilizing element site pages that are connected with the program summoning them. Receptive procedures work with the recorded log documents just, and the distinctive clients will be recognized by their navigational examples, download timing grouping or some different heuristics in light of some supposition in regards to their conduct. For instance, web clients are recognized in light of their navigational examples utilizing bunching techniques. Client recognizable proof should be possible in different routes like utilizing IP addresses, User Agents, treats, direct verification etc. We assume grouping of IP & user agent to recognize user uniquely. User identification can also be done using client side cookies. But, due to privacy reasons, cookies can be disabled by user, and not every website employ cookies.

**Creating indexes for URL**

Here we are applying the indexes for distinct requests of every user. We are using ‘hash Table ‘data structure for storing the index value as an integer and user demanded page as a cord. After applying the indexes for all separate requests, the indexes and requested pages are written into file.

**ALGORITHM for Indexing**

**Input**: User recognized log record

**Output**: ordered document

For every User U

Add solicitation page to cluster A

End for

For every solicitation in cluster A

Uproot excess in exhibit

End for

For every solicitation in cluster A

Increase the value of Hash table

Compose solicitation page and list worth to IP index record

End for

**Session Identification**

Session identification is regarded as the further step. A session is a succession of requests completed by a single user with an exclusive IP address on a particular web domain throughout a specified period of time. Few heuristics used to identify sessions are

* **Timeou**t: it occurs when the time connecting page requests go beyond a certain limit, and it isassumed that the user is opening a new session.

Page stay time threshold =10 min, Session time out threshold =30 min.

**IP/Agent**: Each special cause type for an IP address symbolizes a dissimilarsessions.

**Referring page**: If the referring page file for a request is null, it is assumedthat the request is coming from a different session.

Time oriented approach**:** The most essential session definition accompanies Time Oriented Heuristics which are in light of time constraints on aggregate session time or page-stay time. They are partitioned into two classes regarding the edges they use: In the first, the term of a session is restricted with a predefined upper bound, which is normally acknowledged as 30 minutes. In this sort, another page can be attached to the present session if the time distinction with the first page doesn't abuse all out session length of time. Something else, another session is expected to begin with the new page demand. In the second time-situated heuristic, the time exhausted on any page is confined with an edge. This edge worth is acknowledged as 10 minutes. In the event that the timestamps of two successively got to pages is more prominent than the edge, the present session is ended after the previous page and once again session begins with the recent page.

**Session construction**

The session is phrased as “user activity over duration of time on web site”. At this point we identify chain of web page demands by sole users on meticulous web domains for a precise time phase. This is tenured as session construction.

**ALGORITHM for session construction**

**Input:** User recognized log record & ordered document

**Output:** Session record

For every User U in view of distinctive (ip+browser)

For every solicitation r of User U

If(current\_req\_time – first\_req\_time < 30 mins)

Add this solicitation to current session

Else

Make fresh session & Write session subtle elements (ip,time,req) to session document

curr\_req\_time=curr\_req\_time - 30

End For

End For

**Session Page View Matrix**

This step consists of the creation of the session page view matrix. In the session page view matrix, the sessions are represented as rows and the pages are represented as columns. In a particular session, if a particular page has been visited then its value is given as 1. And if the page is not visited, then the value is given as 0, in the respective cell. The session view matrix thus created is used in the later stages.

Web usage navigational data is always bulky and diverse. So to have proper pattern discovery of web usage navigational data this must be uniform, integrated and also comprehensive. Data preprocessing is initial step in data mining where it involves filtering of redundant and irrelevant data. And also there will be transforming and resolving any inconsistencies if any. And finally identifying unique users and constructing user sessions.

* 1. **PATTERN DISCOVERY USING K-means and SOM**

**K-means Clustering algorithm**

K-means algorithm is the main broadly used algorithm in data mining for grouping very large web usage data items into sets with similar attributes. The k-means method is an evolutionary method that expands its name since its method of procedure. The k-means method has been revealed to be useful in producing excellent clustering results for numerous practical purposes. K-means clustering is a routine frequently used to regularly partition the data set into *k* collections.

K-means is one among many partition based clustering advance where each group is correlated with a centroid (center point). Every mark is allotted to the cluster with the contiguous centroid and number of clusters K ought to be specified. The basic step by step procedure is very easy

**ALGORITHM:**

**Input:** The session weight matrix which is obtained in earlier, preprocessing stages of web usage data. Along with Number of clusters to be formed, cluster initialization and distance matrix obtained before by preprocessing of web usage data.

**Output:** various clusters with closest centroid. The k-means clustering algorithm of web usage data encompasses the following four strides.

**Step 1:** select K primary cluster centers (representing k groups) at random from the data set which includes around 2398 sessions formed in preprocessing steps.

**Step 2:** Assign all sessions to their nearby cluster (measuring as of the cluster center). This is ended by presenting a each data point *x* and estimate the similarity (distance) *d* of this input entries *y* of all cluster center *j*. the closest cluster center to a data point *x* is the cluster center with least distance to the data point x.

**Step 3:** compute the center of every cluster as the centroid of all the sessions in every cluster.

**Step 4:** If the latest centers are diverse from the previous ones, replicate steps 2, 3 and 4. Otherwise terminate the algorithm.

So K-means initiates with primary partition of K clusters and allocate patterns to these clusters in order to lessen or decrease the squared fault.

Some of the applications of K-means algorithms:

* It is moderately effective and quick. It registers result at O(tkn), here n is number of items or focuses, k is count of groups and t is count of cycles.
* k-means bunching can be connected to machine learning or information mining
* Used on acoustic information in discourse comprehension to change over waveforms into one of k classifications (known as Vector Quantization or Image Segmentation).
* Also utilized for picking shading palettes on antiquated graphical presentation gadgets and Image Quantization.
* K-implies calculation is helpful for undirected information disclosure and is generally straightforward. K-means has discovered across the board utilization in part of fields, extending from unsupervised learning of neural system, Pattern acknowledgments, Classification investigation, artificial knowledge, picture preparing, machine vision, and many more.

Very few disadvantages of K-Means algorithm are:

* When the quantities of information are not all that numerous, beginning gathering will focus the group essentially.
* The number of bunch, K, must be resolved before hand. Its drawback is that it doesn't yield the same result with every keep running, subsequent to the subsequent groups rely on upon the beginning arbitrary assignments.
* We never know the genuine group, utilizing the same information, in light of the fact that on the off chance that it is inputted in an alternate request it may deliver diverse bunch if the quantity of information is few.
* It is delicate to starting condition. Diverse introductory condition may deliver distinctive aftereffect of group. The calculation may be caught in the nearby idea.

**SOM Algorithm**

SOM is used to cluster the data based on similarity and along with an idea of assigning the same number of instances to each session. SOM is one of the neural network algorithms which not only cluster the data but also reduces the dimensionality of data. Here SOM makes use of weight matrix that is obtained from previous steps. Weight matrix includes weights calculated for each session. So here we illustrate the feasibility of using SOM to mine the web log data of user navigation. This helps web masters to superior understand the characteristics along with navigational performances of web users who visit their sheets. SOM is one of the effective algorithms.

The self-organizing map (SOM) is valid apparatus in examining stage of data mining. It ventures input space on prototypes of a short-dimensional regular grid that can be efficiently utilized to discover properties of the data.

**ALGORITHM:**

**Input:** Sessions formed using the concept hierarchy and link in sequence where each URL is assigned with a distinctive or sole index.

**Output:** Clusters representing the sessions with similar steering behavior.

**Step 1:** Choose casual values for the weight vectors.

**Step 2**: Select an input vector sample with a definite probability. And also there is a set of Web pages along with a set of user transactions.

**Step 3**: For every input vector, should calculate the Euclidean distance among the input vector and the weight vector. Winning neuron will be the index value of the weight vector having minimum Euclidean distance with the input vector.

**Step 4:** alter the weight vectors of all neurons by means of the update procedure

**Step5**: maintain with step 2 until no obvious alterations in the feature map are viewed.

**5.5 VISUALIZATION OF CLUSTERS**

**Visualization tool used**

**JFree Chart**:

JfreeChart is an release source library extended in Java. And also it is at no cost complete Java free chart library which is easy for developing team to display qualified class charts in their functions. JFreeChart's characteristic deposit consists of:

* A regular and fine-certificated API, sustaining a wide sort of graph types.
* A stretchy propose that is simple to expand, and aims together server-side and client-side purposes.
* It is a shore up for numerous productivity types, together with Swing and JavaFX workings, image files (PNG and JPEG), and vector graphics file formats (PDF, EPS and SVG).
* JFree Chart is an unwrap source or, more exactly at no cost software. It is scattered beneath the requisites of the GNU Lesser General Public License (LGPL), allows applying in proprietary purposes.

The JFreeChart is famous for its efficient chart creation, and user-friendly installation setup.

And here we make use of this useful tool to compare two well known clustering algorithms. Particularly we used pie chart and also line chart to analyze the results of cluster formation by both algorithms. And hence we can also have clear comparison between them.

**Line chart:** A line chart or line diagram displays information as a sequence of data positions or markers united by straight line sectors. Line diagram gives you an idea about how the data changes at equal time incidence.  They are destined to evaluate two disconnect variables and later these variables are planted on an axis. In the end, we will get a chart with a line that starts from a unchanging point across a chart going up and doing in relative to data.

**Pie chart**: A pie diagram (or a circle outline) is a roundabout measurable chart, which is further separated into fragments to point up numerical extent. In a pie graph, the curve length of every cut (and likewise its focal point and zone), is relative to the limit it speaks to. While it is known for its likeness to a pie which has been fragmented, there are varieties on the mode it can be introduced. A pie diagram which is additionally known as a pie outline is a kind of chart routinely utilized as a part of mix with rates. A major circle is isolated into segments relying upon those rates and every portion speaks to some piece of the entirety. In a pie graph, the curve length of every portion is intended to be relative to the rate it should speak to.

**CHAPTER 6**

**TESTING**

**6.1 SYSTEM TESTING**

Testing presents very significant function for quality declaration and to guarantee the consistency of the software. The plan of testing is to discover the maximum probable number of errors. In other words testing is to make out maximum number of faults with lowest amount of effort and sensible time period. The word testing itself indicates that we are testing something on the developed modules based on giving some known inputs so that we can get to know that we got desired output. Testing also depends upon the objects or modules we test and it also helps us to find logical errors.

All object oriented model must be tested for exactness, completeness and uniformity. The system must be tested with respect to efficient requirements and also with respect to non functional requirements and also with the component and interaction of modules with other modules that survive in the system. For an error free plan the developer would be fond of to determine all the test cases. Hence additional number of errors has to be detected with least number of test cases.

**6.2 UNIT TESTING**

In unit testing each module developed is tested separately for its accuracy, timeliness, simplicity etc. The various errors discovered during the unit testing may be the errors in data interface between routines, logical error in the algorithm and failure to account for various processing cases. It may also contain requirement document, design error and implementation. These errors are detected while running with sample data. The errors are then rectified and tested again. This continues till all the errors are removed and the user needs are fully satisfied. Each individual file involved in projectis tested by the developer by giving the input file. The input file contains numeric value. These inputs are validated by writing Java code. This testing is done at that time of development of the project.

**Test Case 1: Unit Testing of Module 1:**

**Table 7.1 Main Menu Unit Test Case**

|  |  |
| --- | --- |
| **Name of the Test** | Test Case for Main Menu window. |
| **Test Description** | A Test for displaying Selection Menu. It consists of all modules buttons to perform the operation. |
| **Sample Input** | Program Execution. |
| **Expected Output** | Display the Selection Menu Screen. |
| **Actual Result** | As Expected. |
| **Remarks** | Pass. |

**Test Case 2: Unit Testing of Module 2:**

**Table 7.2 Pre-processing Module Test Case**

|  |  |
| --- | --- |
| **Name of the Test** | Test Case for pre-processing Module. |
| **Test Description** | A Test to verify whether web navigational data is cleaned means the unwanted and redundant data is removed. And even check for formation of sessions. |
| **Sample Input** | Log file |
| **Expected Output** | Text files with separation of sessions and users. |
| **Actual Result** | As Expected. |
| **Remarks** | Pass. |

**Test Case 3: Unit Testing of Module 3:**

**Table 7.3 Cluster Formation Module Unit Test Case**

|  |  |
| --- | --- |
| **Name of the Test** | Test Case for Cluster Formation Module. |
| **Test Description** | A Test for displaying clusters of sessions using K-means and SOM algorithm. |
| **Sample Input** | Weight matrix with no of clusters. |
| **Expected Output** | Display the clusters. |
| **Actual Result** | As Expected. |
| **Remarks** | Pass. |

**Test Case 4: Unit Testing of Module 4:**

**Table 7.4 Visualization Module Unit Test Case**

|  |  |
| --- | --- |
| **Name of the Test** | Test Case for visualization Module. |
| **Test Description** | A Test to visualize the clusters formed by both algorithms. |
| **Sample Input** | Cluster of sessions. |
| **Expected Output** | Charts showing separate sessions. |
| **Actual Result** | As Expected. |
| **Remarks** | Pass. |

**Test Case 5: Unit Testing of Module 5:**

**Table 7.5 Comparison Module Unit Test Case**

|  |  |
| --- | --- |
| **Name of the Test** | Test Case for Comparison Module. |
| **Test Description** | A Test to compare the clusters formed by both methods. |
| **Sample Input** | Clusters formed before. |
| **Expected Output** | Difference between their clusters. |
| **Actual Result** | As Expected. |
| **Remarks** | Pass. |

**Test Case 6: Unit Testing of Module 6:**

**Table 7.6 sessions count Module Unit Test Case**

|  |  |
| --- | --- |
| **Name of the Test** | Test Case for session count |
| **Test Description** | A Test to check for same cluster in both methods |
| **Sample Input** | Same cluster |
| **Expected Output** | Sessions present in particular cluster. |
| **Actual Result** | As Expected. |
| **Remarks** | Pass. |

**Black Box Testing**

Black box testing is nothing but running the whole system, but what function is going on is unknown. This testing is otherwise called as structured test; in our project all the files are integrated and tested structurally.

**Functional Testing**

In this functional testing if the screen goes blank, they may start to wonder as to what is happening and the user could just do anything such as press the enter key a number of times, or switch of the system and so on, but if a message is displayed saying that the processing is in progress and asking the operator to wait, then these type of problems are avoided in the functional testing. The functions involved in our project are analyzed by calculating the performance and response time after giving inputs.

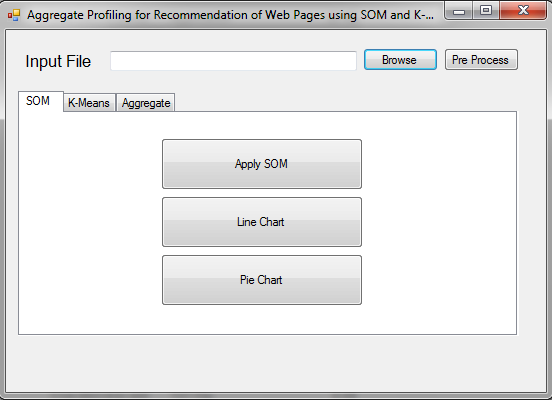
**6.3 ACCEPTANCE TESTING**

Testing is crucial to the victory of the scheme. System testing formulates a reasonable postulation that if all elements of the system are correct, the target will be successfully achieved. As the word itself says we are going to combine all the so far tested now and finally building up a system and test this system as a whole. If we get correct output then finally system can accepted.

**CHAPTER 7**

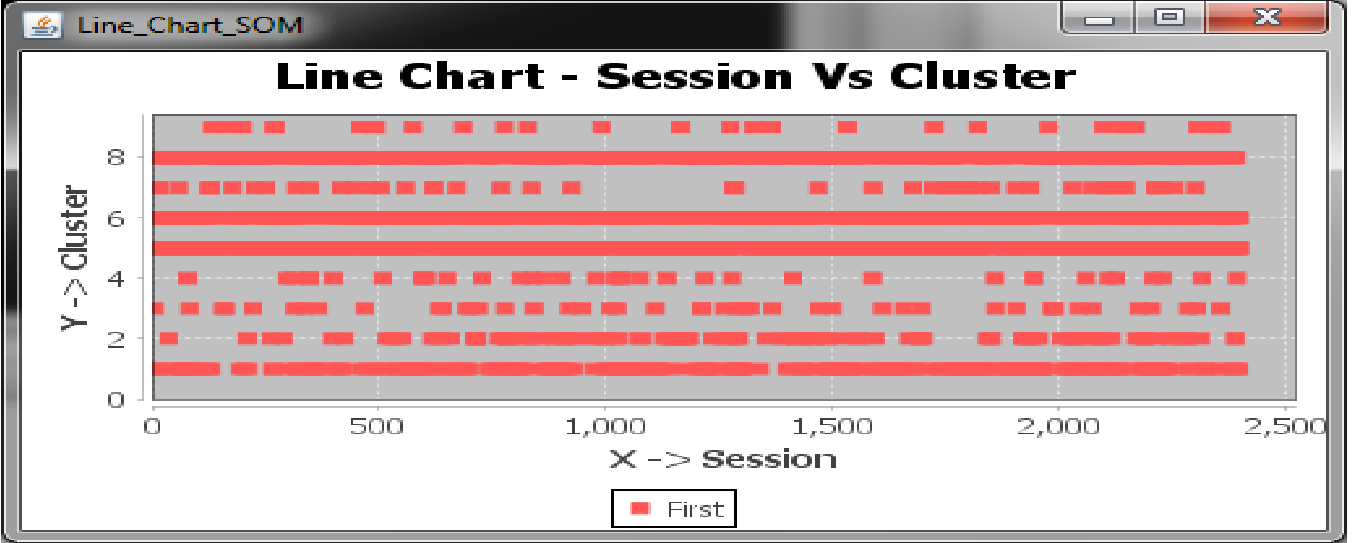
**RESULTS AND ANALYSYS**

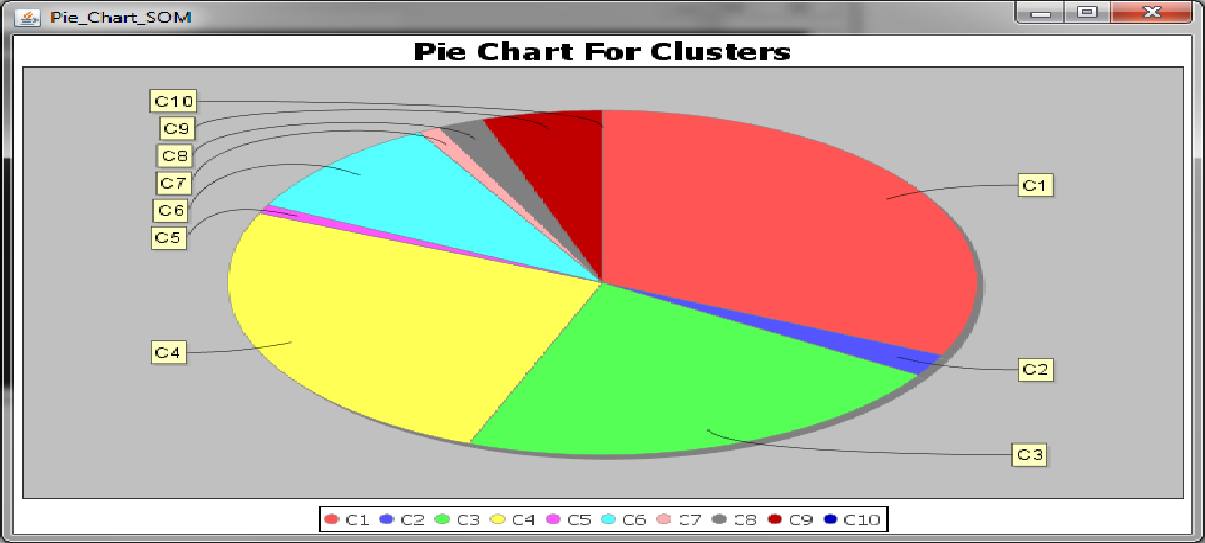
The subsequent snapshots characterize the outcomes or yields that we will attain after orderly implementation of the considerable number of modules of the framework.



**7.1 Results for SOM algorithm**

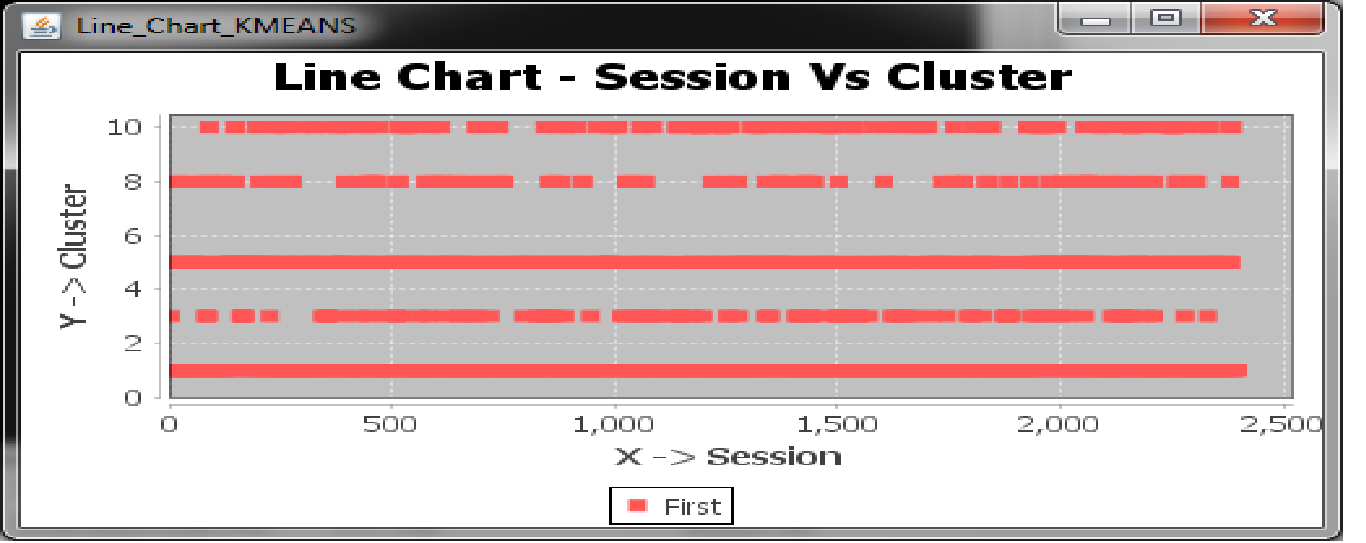
The result of self organizing map clustering algorithm on web usage data that is preprocessed before is as shown below in screen shots. These represent how sessions are separated into various clusters. There are total 2398 sessions after preprocessing of raw log data. And these are categorized into various clusters according to algorithm. As per the SOM algorithm some clusters are dense where as others contain very few or even null. This can be clearly understood by both graphs below. In line chart below no of sessions are plotted against each cluster. And in pie chart whole arc length is divided into no of clusters and each cluster shows the percentage of sessions it includes.

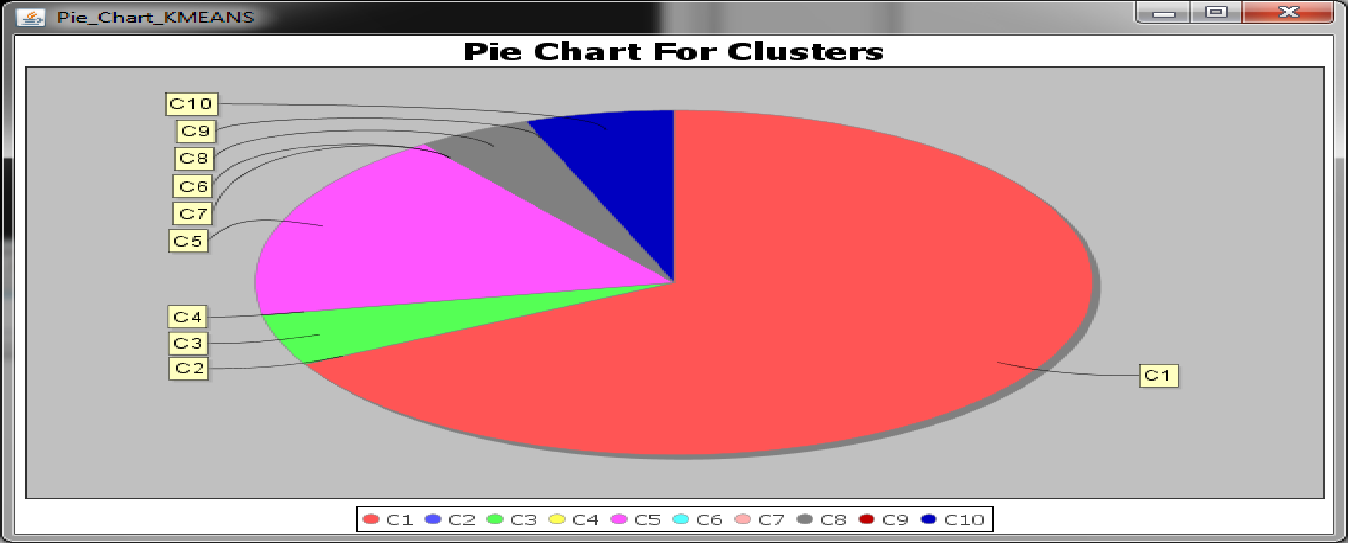
****

****

**7.2 Results for K-means algorithm**

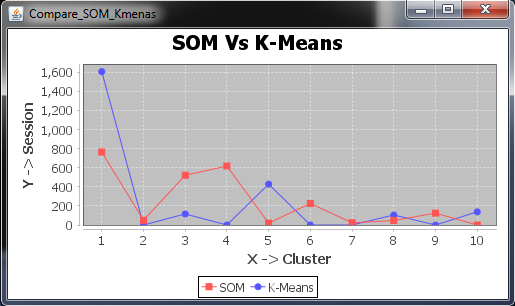
The results for K-means algorithms are as per in screen shots below. According to K-means algorithm the no of clusters are given initially and by taking weight matrix as inputs it moves the data into different clusters depending on weights of each. Here it is well represented using line chart and pie chart. In line chart it’s given sessions in particular cluster. And in pie chart it clearly shows the percentage of each cluster having various sessions in it.

****

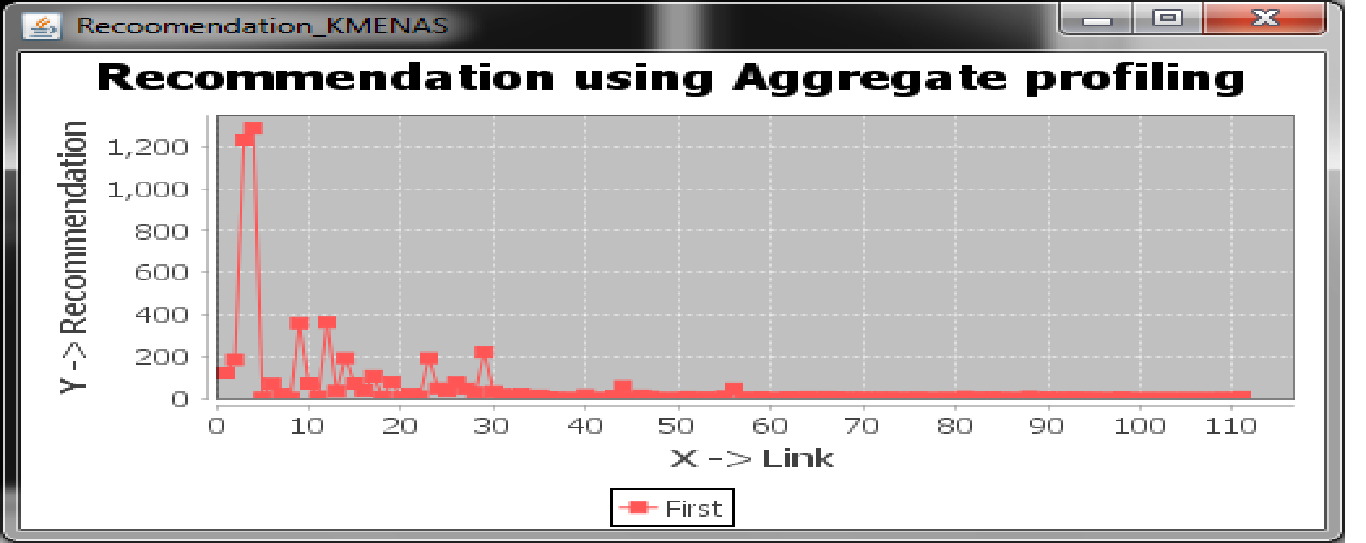
****

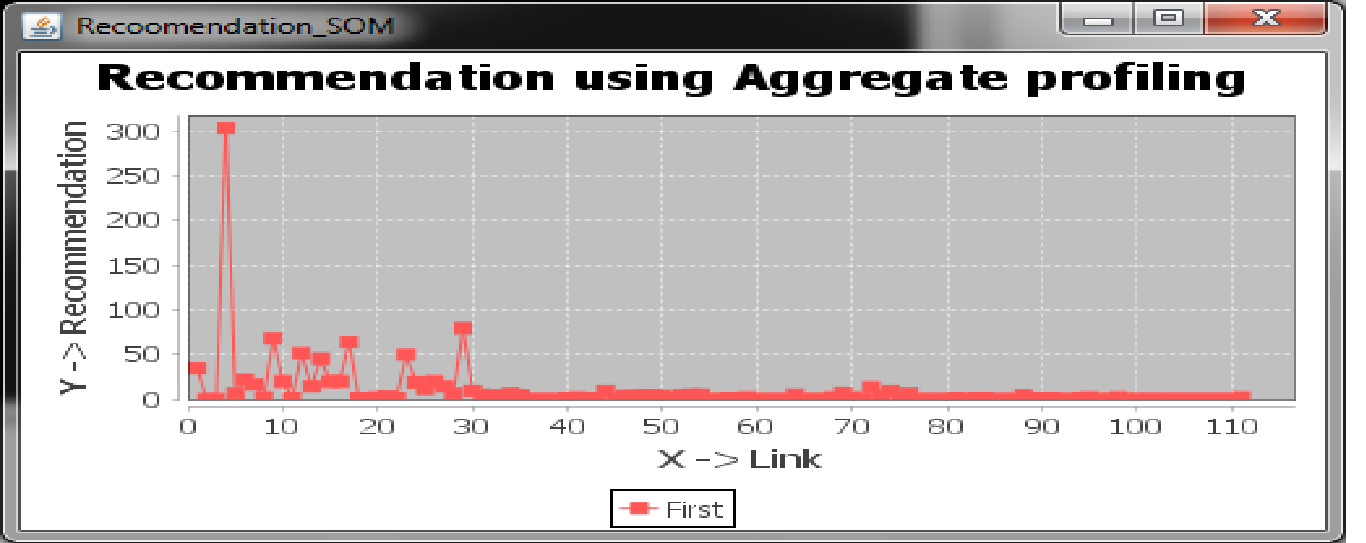
**7.3 Comparison between both algorithms**

This screen shot clearly explains the comparison between both well known algorithms as SOM and K-means clustering algorithm applied on web usage navigation data. Here red line indicates SOM algorithm and blue line indicates K-means algorithm. As per the comparison we can say clusters from both algorithms contain different data. That is sessions in various clusters are different in both clusters. That may vary according to steps followed and number of sessions in individual clusters. The various clusters contain different number of sessions. And even some of the clusters are null. This below line graph explains difference between both clustering algorithms.

****

In below screen shots it clearly shows the cluster formed by both methods and same cluster in k-means and SOM varies in having different sessions in it. We can easily find out the difference between them. Also the session count in both methods varies accordingly. Even some of clusters contains zero sessions in it as they attains minimal distance between centers. So here we consider cluster no 2 from both algorithms and accordingly it represents the various sessions in that cluster.





**Conclusion and Future Work**

The number of recommendations obtained using K-Means is very large with an anomaly that all the recommendations are part of a single session as the Aggregate Usage Profile values are 1 for almost all recommended URLs. Whereas Aggregate Usage Profiles for URLs obtained using SOM are better in terms of recommendation because their Aggregate Profile values lay between 0.33 and 0.89; indicating that more number of sessions have the same URLs (users with similar interests). By observation it is found that the recommendation process is better using SOM as compared to K-Means in terms of Aggregate Usage Profiling.

In future, we will work on the complexity and performance of recommendations generated along with the usage of other Web Usage Mining clustering algorithms to generate better Aggregate Usage Profiles with voluminous data sets; and develop algorithms to enhance and compare the performance of clustering processes on the basis of various index values by removing unwanted clusters.

**Bibilography**

[1] Paweł Weichbroth, Mieczysław Owoc, “Web User Navigation Patterns Discovery from

WWW Server Log Files”, 2012, Preprints of the Federated Conference on Computer Science

and Information Systems, pp. 1207–1212.

[2] Nawal Seal, Abdelaziz Marzak and Hicham Behja (2013). Web Usage Mining, International Journal of Computer Science , Volume 10, Issue 2, No 2, March 2013, pp. 45-48.

[3] Laura Hollink, Peter Mika, Roi Blanco, “Web Usage Mining with Semantic Analysis”, 2013, Proceedings of the 22nd international conference on World Wide Web, pp. 561-570.

[4] R.M. Suresh, R. Padmajavalli, “An Overview of Data Pre-processing in Data and Web Usage Mining”, 2006, 1st International Conference on Digital Information Management, pp. 458-495.

[5] Kate A. Smith, Alan Ng, “Web page clustering using a self-organizing map of user navigation patterns”, May 2003,Volume 35, Issue 2, pp. 245–256.

[6] Jose Alfredo F. Costal and Hujun Yin, “Gradient-based SOM Clustering and Visualisation Methods”, May 2010, The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1 – 8.

[7] Shi Na, Liu Xumin, Guan yong, “Research on k-means Clustering Algorithm”, April 2010, Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium , pp. 63-67.

[8] Yan Zhu, Jian Yu and Caiyan Jia, “Initializing K-means Clustering Using Affinity Propagation”, Aug 2009, Hybrid Intelligent Systems, 2009. HIS '09. Ninth International Conference on (Volume:1 ), pp. 338-343

[9] Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas, “Randomized Dimensionality Reduction for k-Means Clustering”, Nov 2014, IEEE Transactions on Information Theory (Volume:61 , Issue: 2 ) pp. 1045-1062.