Journal of Advances and Scholarly Researches in Allied Education
Vol. XIV, Issue No. 1, October-2017, ISSN 2230-7540

A Comprehensive Literature Review on Sequential Patterns Algorithms in Web Usage Mining

Aarti Pandey[1]* Prabhat Pandey[2] Ajitesh S. Baghel[3]

[1]Research Scholar, Dept. of Computer Science, A.P.S. University, Rewa, India

[2]OSD, Additional Directorate Higher Education, Division- Rewa, India

[3]Lecturer, Dept. of Computer Science, A.P.S. University, Rewa, India

Abstract – Web Usage Information is the main repository for network utility channelling that primarily consists of web client logs, proxy logs and server logs. As server logs have all but interchangeable structures and are promptly supplied to all domain server hosts, which is in general cooperative and needy repository in research on network utility channeling. In this review paper the basic approaches to extract blue prints in network logs comprise of analytical analysis, session recognition, client recognition, assortment and successive principle channelling.

Keywords: Web Usage Mining, Web Data Characteristics, Successive Pattern Mining.

----------X----------

1.1 INTRODUCTION

Web Usage Mining knowledge discovery through web mining passed through web fully or partially and its various algorithms. There exists enoroumous amount of substantial data in website pages and furthermore their hyperlinks. These pages are gotten by the clients or server or proxy and thus another arrangement of data by name web logs is created. These logs contain the get to pattern of the clients. The strategies utilized for mining these logs unexpectedly find and distinguish leaving Data from the logs. Web mining originated from databases, recovered data, data acquisition and instinctive speech litigating network procedures can be characterized into three types particular.

√ Web substance mining

√ Web structure mining

√ Web usage mining

1.1.1 Web Substance Mining:

Web substance is a blend a few sorts of data like organized data, semi structures data, unstructured data encourage this data again could be, pictures, sound or video. The class of calculations that reveal valuable Data from these data sorts or reports is called web mining. The fundamental objectives of WSM incorporates helping Data discovering, (example Search Engine), shifting Data to clients on client profiles, database see in WSM mimics the Data on the network and fuse them for huge number of convoluted inquiries. Numerous astute instruments to be specific web operators were created by the analysts for Data preparing and recovery and a more elevated amount of deliberation is given to the semi-organized data on the web utilizing the data mining procedures. WCM and sight and sound data mining proficiencies are valuable exhuming the subject in network paginates. Some of these approaches are as follows:

I Operator –Based Approach By and Large, specialist based Web mining frameworks can be arranged as:

√ **Insightful Search Operator:** Different very much educated network agentive parts are developed for turning upward for relevant entropy using learning base elements and customer visibilities to get ready and render the found out data. A portion of the web specialists are Harvest (Brown, et. al., 1994), FAQ-Finder (Hammond, et. al., 1995), Data Manifold (Kirk, 1995), OCCAM (Kwok and Weld, 1996) and Para Site (Spertus, 1997).

√ **Data Filtering/Categorization** The network agentive parts utilize distinctive data recuperation techniques (Frakes and Baeza-Yates, 1992) and components of real to life machine-comprehensible network papered to mechanically recoup and evaluate them.

√ **Personalized Web Agents** Many web specialists learn client premiums as indicated by their web use and find the examples in light of their inclinations and premiums. Cases of such customized web operators are the Web Watcher (Armstrong, et. al., 1995), PAINT (Oostendorp, et. al., 1994), Syskill & Webert (Pazzani, et. al., 1996), Group Lens (Resnik, et. al., 1994), Firefly (Shardanand and Maes, 1995) and others (Balabanovic, et. al., 1995). For example, Syskill and Webert utilized Bayesian classifier to rate website page of client's interests in light of client's profile.

II Database Approach The semi-organized data is sorted out to organized data utilizing different database approaches. Different database question preparing components and data mining networks are utilized to dissect the organized data accessible on web. The database methodologies are recorded as:

√ **Multilevel databases:** The principle thought behind this approach is that the most reduced level of the database contains semi-organized Data put away in different web vaults, for example, hypertext archives.

√ **Web query frameworks:** A substantial number of networks grounded cross examination frameworks and dialects utilize standard vault cross examinations dialects like organized question dialect, morphological data about network content records, and customary natural dialect treating for the cross examinations which are used in web look ups (Konopnicki and Shmueli, 1995).

1.1.2 Web Structure Mining:

Organize structure removal is related on summoning the model or examples or structures which develop the basic portrayal of the web through connections. It is utilized to concentrate the progressive structure of the hyperlinks. The connections might be with or without portrayal about them. This structure is valuable to arrange organize varlets and accommodating to inspire data for instance same kind and family relationship among different web locales. NSE can be used to reveal approved destinations. More noteworthy are the development of the network varlets and the nature of the pecking request of web connections in the site of a particular field.

Couple of algorithmic tenets have been recommended to reproduce the network topology for instance HITS (Chakrabarti, 1997) Page Rank (Brin and Lawrence, 1998) and advancements of HITS by counting topic to the connections structure and by using inhabitant stressing.

1.1.3 Web Usage Mining:

Organize use exhuming focuses on methodologies which suspect customer direct when the customer proceeds onward the network. WUM plans to reveal energizing WUM means to uncover energizing intermittent customer get to designs delivered while the surfing the web which is kept up in the web server logs, middle of the road server logs or client logs. WUM is about discovering patterns of site hits by Web clients or to discover the use of a specific Website. There are numerous utilizations of Web use mining, for example, focusing on notices whose goal is to locate the arrangement of clients who are well on the way to react to a notice.

Web Usage Mining' would analyze the studies about: web log characteristics, data pre-processing: data cleaning, user identification, session identification, transaction identification, issues in data pre-processing, pattern discovery and analysis. Finally a summarization of this section would be presented for crisp understanding of future research and similarly the difficulties in data pre-processing would also be discussed for better understanding.

1.2 WEB USAGE DATA CHARACTERISTICS:

The web usage data fundamentally keeps up logs of get to patterns of the guests on a site. It can likewise incorporate customer visibilities, bookmarks, treats, change data, customer questions and some other associations of the customer while on the site. For simple reasonability and accommodation the data is assembled into three divisions: Network Host Logs, Gateway Host Logs and Client Browser Logs.

The web server keeps up vital Data for network used uncovering these logs are when all is said in done access of sites by various clients. For each of the records contain the IP address of the client, Petition time, Uniform Resource Locator, HTTP status figure and so on. Obviously the Data accumulated are in a few standard organizations like log document design, extended log record arrange and so forth.

An entryway like server known as web server intermediary server goes about as a door for the clients and the servers. To diminish the data time of a site page the intermediary getting is valuable and the customers visit these site pages intermittently and alongside this the intermediary accepting is likewise helpful to have the total perspective of the heap movement at the server and the customer. The intermediary server can make sense of the entire solicitations made utilizing the hypertext exchange

convention from various clients to various web servers. Utilizing this intermediary server the surfing exercises of a bunch of comparative and conspicuous customers who share a similar intermediary server is broke down and in this manner examined. The specialist accessible at the customer side is useful to accumulate the use Data of the client at the customer side. This operator can likewise be viewed as the web program having the capacities to decide the errands completed by the customers. These logs gather Data of a specific customer from different sites. The Data from the customer side catches basic Data when contrasted with web or these door logs for instance for reloading the page snaps of the mouse is utilized or back key is additionally utilized. The present part gives a synopsis of web server logs and a hefty portion of the web mining methodologies are extremely helpful for the web usage mining.

The authors Grace et. al., (2011) had studied about the basic features of the web log uses in the web mining consists of varied information, such as: time stamp, access request, user name, IP address, result status, referred URL, user agent and the bytes transferred. These logs are basically stored and maintained in the main web servers. It can be done through web server for process analysis like data pre-processing. The key features of weblog content are user name, time stamp, page last visited, success rate, user agent, URL and request type visiting path, path traversed. These log file are found at three sources mainly;

a) Web servers,

b) Web proxy servers

c) Client browsers.

According to the above scenario we analyze that a log files can be classified in two types, such as:

1) Former Log are categories into transfer/access log and agent log are basic and standard,

2) The latter logs are categories error log and referrer log which are normally found in the “extended” log file format.

Dhawan and Goel, (2013) studied the usage patterns from web logs in web usage mining. The web log file data may be web document content and hyperlink structure. Through which the Web usage mining discovered the knowledge. In fact, it has provided the vital information about the user access patterns from the web log files that represent the user activities accessing a web site. Some of server log files associated with the web server logs are Access log, error log, agent log file and referrer log.

Talakokkula, (2015) has focused on the web usage mining applications and tools used for extracting useful inform ation from various web logs. He also analyze, the web log mining has been considered as the process of extracting interesting patterns from the web access logs. The various web logs associated with the web usage mining are Server logs, client logs and network logs are maintained by server, web browsers and other networks.

1.3 PRE-PROCESSING:

This technique is utilized to handle the genuine web logs before the genuine mining process and the principle aim of this pre-processing strategy is to perceive entire web sessions or occasions. At the point when the web server logs are utilized the web server stores the total Data of the whole customer's get to conduct. During this procedure the customers are considered as the entire and as the individual web convention deliver is not coordinated to any known profile in the vault.

The greater part of the web logs are considered as the bunch of progressive chains of the get to occasions from a one of a kind customer or stage in the era in the expanding request. The present strategy is appropriate to all log records to discover the Data on sessions of the web. The strategies which are incorporated into the pre-processing are Data cleaning, customer acknowledgment and stage's acknowledgment.

Cooley et. al. (1997, 1999, and 2000) presented networks for customer conspicuous confirmation, session ID, site hit recognizing verification, way fulfillment and scene ID. In any case, a segment of the heuristics proposed is not fitting for greater and more personality boggling Web goals. This is tended to under the term 'customer session'. The proposed heuristic hopes to perceive customers with a comparable IP address by checking each page requested in a consecutive demand. If a page requested is not implied by any past page, then it has a place with another customer session.

Bonchi et. al., (2001) also developed an data circulation community for securing Web log reports. Their model does not contain sorted out data about the usage(sessions, visits, et cetera.), customers or gathered elements. The goals of their work were Web holding. For Web holding applications, each one of the requesting show in the Web logs are essential and must be secured. Concerning customer and the session conspicuous confirmation, the makers use the prospect of customer, recognized using the IP heuristic.

Berendt et. al. (2002) used the administration based sensible request for exhibiting the question limits of an online stock. This Web website page offers

examining and filtering administrations for schools far and wide. The makers expected to survey the request instead of the substance of the Web pages created, in this way, the use of administration orchestrated sensible pecking request. Concerning the pre-planning step, the makers propose three heuristics including one like the Browsing Speed. The other two heuristics insinuate the amount of sales without the referrer field and to the requesting reiterated for a comparative resource from a comparative host. In any case, the later one should be attempted against various heuristics to affirm its practicality. Huysmans et. al., (2004), dismembered the Web log data from an online wine shop (eshop). They developed an instrument for pre-processing. They determined that the pre-processing of this data and especially clearing of the photos decreased its fundamental size. This was plainly a direct result of the kind of the Web site separated, as e-shops and online business Web goals have many pictures to demonstrate the things sold. Another intriguing remark made by the makers is that by using a customary log analyzer program, the degree of the log record was diminished just by 15%, which exhibits again that fitting gadgets are essential for a convincing pre-dealing with step.

Marquardt et. al. (2004) used the usage of WUM in the e-learning zone with an accentuation on the pre-taking care of stage. In this one of a kind condition, they reconsidered visit from the e-learning point of view. In their approach, a learning session (LS, can navigate more than a couple days if this period identifies with a given learning period. A learning session may in like manner identify with the course of action of get the opportunity to make to satisfy a given task. The designers also perceived scenes like (Cooley 2000), by requesting Web pages into three page sorts (aide, substance and resource) in light of the present finding out about the Web site.

Spiliopoulou et. al., (1999) developed a Web Usage Miner (WUM) device, It means to discover progressive illustrations which are considered as captivating from a quantifiable point of view. WUM proposes to evacuate back to back illustrations having a base support and organizing a customer portrayed plan. For this, the sessions are changed into a totalled tree. Each center point in this tree is associated with a page from the session way.

Masseglia et. al., (1999) proposed another network for WUM, a Web Tool. This structure considers each one of the methods for a WUM methodology, from the data assurance to the results appears, by methods for the data change and illustrations extraction. The Web Tool relies on upon a prefix tree (the PSP, proposed by the makers) to evacuate progressive cases. The goal is to get all the persistent cases relying upon the "delivering pruning" method (Apriori standard). PSP, as various methods in perspective of this rule, ends up being less beneficial when the base support or the successive illustration depiction is low. The WebTool uses the PSP calculation. It relies on upon an indistinct general calculation from GSP, yet it uses an improved tree-like data structure for securing confident courses of action. The prefix-tree data structure used as a piece of PSP contains each one of the rivals in the going with way: any branch going from the root to a leaf stays for a candidate progression, and considering a single branch, each center point at significance i gets the ith individual from the gathering. Also, close by everything, the support of the gathering from the root to that leaf center point is similarly secured.

Pei et. al. (2000), portrayed WAP-Mine, as a network that allows the extraction of persistent cases from the customer sessions, refered as Web get the chance to outline (WAP). The calculation creates a totaled structure called WAP-tree. By using this structure, the makers connect the prerequisite for contender time as in other Apriori-like calculations. Thusly, it makes the approach more successful. They stood out the proposed procedure from the WAP-mine. It showed that for progressive illustration extraction for low support. The proposed method is prepared for removing progressive cases. El-Sayed et. al. (2004) proposed the FS-Miner calculation and it relies on upon the FS-Tree which is a compacted tree. It is used to address groupings. Not in any manner like the summed up postfix tree list, the progressions are pressed and quite recently mostly consolidated into the FS-Tree to diminish the memory space required by the data structure. The database is first checked to recognize visit joins (visit game plans of size 2) and a short time later the FS-Tree is collect incrementally. Mining the FS-Tree incorporates a significance first traversal, which is less dull than in the other calculation, as the FS-Tree is smaller. In any case, the FS-Miner does not unmistakably beat the Apriori-based calculations.

Yang et. al. (2007) proposed a calculation LAPIN (Last position Induction) uses an itemlast-position summary and prefix edge position set as opposed to the tree projection or contender make and-test techniques displayed as of not long ago. The major refinement among LAPIN and past basic calculations is the degree of the interest space. Prefix Span channels the whole foreseen database to find the consistent illustrations. SPADE transitorily joins the whole idlist of the likelihood to get the constant cases of next layer LAPIN can get comparable results by checking simply bit of the request space of Prefix Span and SPADE, which are to make certain the spots of the things.

Lizhi Liu and Jun Liu. (2009) proposed a viable approach for using the BFWAP-tree to mine general game plans, which reflects ancestor relative relationship of center points in BFWAP tree clearly and capably.

The proposed calculation fabricates the perpetual header center point associations of the principal WAP-tree in a Breadth-First outline and uses the

layer code of each center to perceive the forerunner relative associations between center points of the tree. It then, finds each nonstop progressive case, through powerful Breadth-First gathering look for, starting with its first Breadth-First subsequence event. Tests show huge execution increment over the WAP-tree technique. The space to store most noteworthy general progressions is much lower than to store complete set, and Web mining applications for the most part simply depend on upon most noteworthy unremitting courses of action instead of the aggregate game plan of persistent groupings so that mining most noteworthy consistent get to game plans is of basic practicability.

Losarwar and Joshi (2012) has mentioned data pre-processing is the data mining ground work and the time spent is 80% for this project in the real world of data mining. The information gathered from www involves pattern discovery is web crawling. He also determine that any organizations value of the product is carried out by the data pre-processing and can be analyzed by the specific customers rating, promotional products campaigns, products strategies across the marketing network, etc. The various steps of the data pre-processing are data cleaning, session identification and user identification.

C. E. Dinucă (2012) analyzed one of the data mining techniques, association to extract useful knowledge from web usage data. He used Java programming language for identification of association of web pages from sessions. The author also fruitfully analyzes to affirm on accurate information and high-quality of data. For analysis of data preparation of information required time somewhere around 60% and 90% and the whole process of extracting knowledge contributes to a success rate of 75-90%.

Shaily Langhnoja, Mehul Barot, Darshak Mehta (2012) proposed algorithm for data cleaning for cleansing web log file by which web log file records reduced 411 from 1217 records, They also propose User and Session identification algorithm marked every record in the database with respective client. Session identified groups that later can be utilized for further course of action of web usage mining process. The resulted group of records can be embedded into database and later results of which can be highly useful like aggregate number of clients, aggregate number of sessions, difference between aggregate number of records before preprocessing and post-pre-processing, etc.

S. Prince Mary and E. Baburaj (2013) studied that the steps of pre-processing involving data cleaning, user identification, session identification and path completion. He also analyze that once after preprocessing is done, various patterns are discovered by applying various techniques like statistical analysis, association and clustering. These outcomes patterns are studied for web personalization, site improvement, site modification, business intelligence, and so on for different applications.Thus accordingly data pre-processing phase is a ground work for Web data mining which spends 80% of time for any real time data mining project.

Mitali Srivastava, Rakhi Garg, P. K. Mishra (2014) discussed in detail that data pre-processing techniques of a log files is an important part, takes the almost 80% time of whole WUM process with their advantage and disadvantage.

Makwana and Rathod,(2014) analyzed the users behavior of data present in website log files He also focused on data pre-processing technique for discovering various usage patterns from log files with different tasks like extraction, user identification, path completion and transaction identification. The association rule, clustering, and classification are uses for the deriving data by methods of pattern discovery.

Tomar and Agarwal (2014) states in a survey paper that data mining steps are a like pre-processing and post processing techniques. In knowledge discovery phase the voluminous data of the information retrieved from the websites are explored and processed. The irrelevant data in log files, unrelated features, and several other inconsistencies if occurs make the processing difficult.

The prediction model of data pre-processing feature algorithm are clustering, text categorization and rule induction. The characteristics of the information of datasets are data set size, multi-dimensional supporting, revealing pattern ability, clusters and background noise amount. The major steps in the visualization of the pre-processing undergo major changes with the time requirements, detailed description of different data, dimensional reduction and cleaning approaches.

V. Vidya and Priya S. Kalaivani (2015) focused on clustering techniques of web log data to identify pattern of user access. In this technique firstly pre-processing phase is done. After data cleaning, their after 25% records are left. In second phase they used K-means and farthest first clustering techniques for the web pages that can easily identify the user interest and the access pattern.

Vijayashri Losarwar and Dr. Madhuri Joshi (2015-2016) proposed data cleaning algorithm, for web log that removes near about 50-60% irrelevant records and filter algorithm that discards the disinterested attributes from a log file. They also explained preprocessing methods to identify User and Sessions of web log for accessing the user patterns .In this paper author also mentioned to carry out real

world data mining project on data pre-processing, 80 % time is spent.

Abraham and Puthiyidam (2016) expressed the data pre-processing used in many remotes sites and complex scenarios that rely on integration of the time in collection of large amount of databases. The various steps in data pre-processing are normalization, transformation, selection, cleaning and feature extraction. The processing time is in much higher amount in the preparation of data. The operation of the data mining is potentially helpful for identifying patterns from the existing data.

Mitharam (2012) has studied Pre- processing task of a web log files or server log files in web usage mining. Accordingly Web usage mining process would be incomplete without using the various stages of pre-processing. Data fusion, cleaning and user identification are significantly associated with the data pre-processing.

According to Kherwa and Nigam (2015), the data pre-processing is a milestone of web usage mining where the raw log data was pre-processed to attain the reliable session for efficient mining. The pros of data pre-processing is to improve the quality and capability of data and it helps to enhance the mining for successfully accuracy.

1.3.1 Data cleaning

The progression contains taking out every one of the Data pursued in network logs that are unusable for uncovering plans e.g.: request of for realistic varlet subject (e.g., jpg and gif pictures) petitions for some other document which may be incorporated into a network varlet; or even route session did by robots and web insects. When appealing to for graphical substance and hotels are agreeable to annihilate robot and web bug's route blue prints must be blue print must be expressly. This is regularly practiced for instance by referring to the far off host, by referring to the specialist, or by guaranteeing in the entrance to the robots.txt document. However, couple of robots truly send a false customer specialist in HTTP asks. For these situation, a heuristic in view of navigational lead can be utilized to partitions robot sessions from exacting customer's sessions is showed that look into motor route tracks are qualified by width first route in the tree symbolizing the site structure and by unassigned referrer.(The referrer gives the site that the client reports having been related from). The heuristic proposed is grounded on the previous assumption and classifications of marine. The web logs recorded amid the clients' cooperation's can't be specifically mined. Henceforth the asked for HTML archives are dealt with as get to occasions. The document sort comprises of the records, for example, Uniform Resource Locators picture records. The picture document might be in any of these configurations like gif, jpg or bmp organize. Hypertext Transfer Protocol has an uncommon code demonstrating the statuses which are helpful for speaking to the accessibility or inaccessibility of the required thing. The occasions that have the status codes from 200 to 299 are viewed as productive occasions, and the remaining are expelled when the web logs are utilized. Some other arrangements like URLs of HTML, ASP, JSP and so forth are expelled from the logs.

Hellerstein (2008) in his paper mentioned sort based algorithm, one pass algorithm, median to other robust estimator's algorithms for data cleaning. He has also mentioned visualization techniques and data detection techniques used for data cleaning types. However, Erhard Rahm and Hong Hai Do. K. Sudheer G. Reddy, adopt the statistical matching techniques of data cleaning types.

Ganti and Sarma (2013), discuss about data cleaning with various activities in the warehouse of business and decisions in the business supporting reports. By the process of data cleaning, the high quality of data can be maintained in any organization. A small error can ruin the reports of the data, in the process of data cleaning. so clean data platform need high challenges with critical analysis.

Hongzhou Sha,Qingyun Liu (2013) proposed the EP Log-Cleaner algorithm to filter out lot of unrelated items based on the common prefix of their URLs. He also reviewed EP Log-Cleaner with a real network traffic trace captured from one enterprise proxy. The experimental outcome of this algorithm enhances the data quality of logs by filtering more than 30% URL requests as compared to traditional data cleaning techniques.

Patel P and et. al., (2014) proposed data cleaning technique to remove unwanted click streams from the log file in web log pre-processing. They reduced the original file size by 50-55%. For User Identification and Session identification, Heuristics approach is used. Session identification is done by using time frame between the page requests usually 25.5 or 20 minutes.

According to author Sagar and Nimavat (2015) the data cleaning techniques have been used to remove the irrelevant data from the log and it also has the tendency to follow all the hyperlinks from web pages appropriately. The unnecessary and irrelevant fields from the raw log files have been removed like gif and jpg because these file types are not requested by the users.

Krishnan, et. al., (2016) discuss in his paper about the data cleaning in a reliable interaction with frequency iteration process. According to him the three main schemes of the data cleaning are data cleaning iterative nature, data cleaning correctness with lack of evaluation and querying of data.

In the year 2016, (Muskan and Garg, 2016) emphasized on data cleaning algorithm that removes accessorial entries like multimedia files, status code other than 200 and entries made by spider or bots. After applying proposed algorithm, entries remained 462 from log table that had initially 1545 entries. The size of the log file decreased up to 71%.

1.3.2 Customer Recognition:

From the portraying perspective the clients' conduct initially the clients' should be recognized subsequently they are dealt with as mysterious as said before. One method for recognizing client is their customer IP address. In this way the solicitations from same IP can be dealt with as same client. Extra Data with respect to the customer could help us pick up knowledge into the clients' behavioural examples. Many clients' get to site utilizing same intermediary then the IP is same yet the sort could be distinctive. Along these lines we could acknowledge that each agentive part sort for same Internet Protocol address symbolizes a customer.

According to Singh and Badhe (2014), the customer recognition in web mining is the user centric approach for the best process of recognizing the users with their data in any website logs. The three categories of the web mining are the substance mining, usage mining and structure mining of the web. The methods used for customer recognition process are: IP address, user generated informative data, cookies collected from the users. The user can be recognise in data by specific person, name, working status, age group, hobbies, nature of person and temporal strategies.

Singh and Badhe (2014) has surveyed customer recognition on web usage mining would be considered as one of the steps of pre-processing techniques. A isolated customer recognition referred as one of the important step in web usage mining. Assigning different customer id to different IP address is one of the simplest methods in the pre-processing techniques. Some of the methods of customer recognition are: IP address, user registered data and cookies. Various methods have been implemented to follow the customer recognition.

Priyanka and Naveen (2015) has studied customer recognition, classification and recommendation in web usage mining is an approach for personalized web mining. The customer recognition through IP address and cookies has been carried out after the method of data cleaning in the log file has completed.

Jiadi and Hai (2016) have used the heuristic rule based algorithm for customer recognition .

Rejena and Malika (2016) used unique customer recognition algorithm and dynamic hashing technique for customer recognition. Customer recognition algorithm, based on patterns using clustering and classification to focused on separating potential customers from others, as compared to other algorithm literature survey by Singh and Badhe (2014) decision tree classification using C4.5 algorithm to identify interested customers for customer recognition.

1.3.3 Session Recognition:

It is comprehended that the customer has navigated the site more than one time at whatever point the session length is farseeing. The point of session acknowledgment is to isolate out network logs of soul client to their get to sessions. The session is viewed as another session if the distinction among the appeal to time of two bordering records from a client is more than the timeout limit. In this work, we have set the default timeout limit as 30 minutes.

Dinuca and Ciobanu (2011) described about session recognition in data pre-processing steps used for recognition of session for each users. The sessions are usage of web sites at different times by individual users. The session should be correctly identified to maintain the user's history of web site access. The three categories can be the determination of average time on single page, visiting of single web sites, duration of each visits and data used for session identification in the web log for perfect analysis of the session.

Kamat, Bakal and et. al. (2013) in his paper focus on the improved data preparation technique in web usage mining where Sessions is considered as time between the sign in and sign out. It also finds the click stream sequence to trace the user effectively. In his paper Session recognition algorithm has taken user list as input. In this way by considering link analysis, the session recognition is done using AHL (Access History List). The Authors have deducted that the sequences of the user session has generated with less time and greater precision effectively but the system is more complex.

Padala et. al. (2013) mentioned data cleaning algorithm, hierarchical session Recognition algorithm and customer recognition algorithm for session Recognition. They also discussed and use data pre-processing technique and pattern discovery technique in their study.

Patel and Parmar (2014) proposed about the customer recognition through the websites, log repositories and servers sites. They analyzed that the pattern of the users can be obtained by the knowledge source to acquire the behavior of the customer. The advance process of customer recognition is based only on the single piece of information known as the time of the users logging

and suing the web servers. The approach used for acquiring the cluster of sessions used by the customers with time analysis. The time analysis is the time per each session, can also observed the initial time degree of each page and statistical results of page.

Halfaker, et. al. (2015) focus on the identical strategies utility to develop web analytics and customer behavior analysis. In their work they demonstrate a strong regularity for online activities like page viewing, video gaming and searching web pages in the temporal rhythms across various web domains. The session recognition is navigational oriented heuristics while its time dependent is known as the time oriented heuristics.

Nigrel et. al. (2015) analyzed the web log pre-processing for web usage mining and proposed the system to identify the customer using two-way hash structure. According to author backward referencing has taken much lesser time. Thus it recognize the session with higher precision. Creation of new session is based on:

1. If there is a new customer, there is a new session.

2. If the refer page is null for one user session, there is a new session.

3. If the time between page requests cross a limit of 30 minutes (default timeout for session).

1.3.4 Transaction Recognition

Cleve, et. al. (n.d) describes the data transaction in the web mining process which analyze customer data transactions. This transaction recognition is used to clearly get the needed results of the transactions of the users or customers logging the web servers. The stages of the web mining process for characterizing the transaction recognition are business understanding, data understanding, data preparation, modeling, evaluation and deployment. Data preparation with preprocessing of the transactions includes data selection, normalization and transformation. Various algorithms are utilized for clustering the web data acquired from the customer transaction like artificial neural networks approach and clustering algorithms.

Joslyn, et. al. (n.d) in his paper presents the potentiality of the graphs used in the web mining process with exploration strategies of the customer data transaction. Modern science and technology in the web mining now make easy to use the graphs, models, directed graphs, hyper graphs, to represent the transaction recognition made by the users. The bit coin block chain is the representation of the transaction recognition which helps the potentiality of recognizing the data with graph analysis and mathematical structuring. The mathematical structuring representation of the block chain has three main regulations of graph called transaction graph, transaction hyper graph and transaction bipartite multi graph. Its mathematical representation can be further analyzed with the visualization process of customer recognition and statistical methods.

Tamrakar and Ghosh (2014) studied about the identification of frequent navigation pattern using web usage mining. The identification of transaction has varied from case to case depending on the technique of web usage mining. Usage preprocessing, content pre -processing are some of the preprocessing phase used for knowledge discovery through some of the data mining techniques such as association rules, sequential patterns and clustering. Data cleaning, user session, path completion and transaction identification are some of the steps of pre-processing phases handles the system effectively. A priority algorithm has designed for finding association rules.

Patel et. al. (2015) in his papers surveyed on Aprior - Tid algorithm, one of the web mining approaches that associated with number of entries which may be smaller when compared to number of transactions in the data base. This approach also scans transaction database and it has a tendency to reduce the cost of system and increases the efficiency in web mining. The representation of transaction data incorporates the complex semantic entities successively.

Vellingiri and Pandian (2011) has provided a techniques for better data cleaning and transaction identification of web log. They aimed to create relevant clusters of references for each customer of transactions recognition. Transaction Recognition is done by means of merge or divides techniques. To discover travel pattern and customers interests, two types of transactions are outlined i.e. travel path transactions and substance only transactions. The substance only transactions are content pages which are utilized as a part of mining to finds customers interest and cluster the customer visiting same web site. The three methods available to recognize transactions are reference length, recognition through Maximal forward reference and recognition by way of Time Window.

1.4 PATTERN DISCOVERY TECHNIQUES AND PATTERN ANALYSIS:

A number of many approaches have been investigated for summoning data shape to organize web logs. They are factual examination, alliance guideline unearthing, constellating, combination and successive blue print exhuming. The quantity of pattern discovery networks that can be connected to the usage data is practically boundless. The strategies go in many-sided quality from generally straightforward procedures, for example, measurable examination to all the more computationally far

reaching techniques. A few strategies are normally connected to an arrangement of usage data keeping in mind the end goal to frame a balanced picture of how a site is being utilized (Han, et. al., 1999).

1.4.1 Measurable /Analytical Approach:

Measurable techniques are the most regular strategy to bring out comprehension about visitants to site. By concentrate the session document, one can execute different sorts of the factual examination (recurrence, mean, middle etc.,) on factors, for example, varlet eyeshot's, watching time and a separation of nautical track. Various network dealings examination devices build up a repetitive review comprising of factual data, for example, the most over and again gotten to in varlets mean look into time of varlet or mean separation of a track done by a site. This review may contain limited low level blame investigation, for example, finding unapproved approaching subtle elements or identifying the most common incapacitates Uniform Resource Locator. Regardless of lacking in the profundity of its investigation, this unusual of cognizance can be potential utilitarian for bettering framework execution, raising framework insurance, easing site modification and rendering attest for promoting conclusions.

1.4.2 Grouping:

Group is an accumulation of data questions that are like each other and not at all like the data protests in different bunches. Bunching is frequently the primary data mining errand connected on a given accumulation data and is utilized to investigate if any hidden examples exist in the data. The nearness of thick all around –separated bunches demonstrates that there are structures and examples to be investigated in the data. On other hand, bunching of pages can find gatherings of pages having related substance. This Data is valuable for web indexes and web help suppliers in both applications, changeless or dynamic HTML pages can be made that recommend related hyperlinks to the client concurring g the clients inquiry or previous history of Data needs.

1.4.3 Order:

Order is the task of speaking to a Data token into one of different prefixed classifications. In the network learning base, one is abundantly worried in detailing a deceivability of customers having a place with an impossible to miss class or family. This needs plummet and selection of qualities that better delineate the properties of a particular class or family. Arrangement can be executed by using checked inductive securing calculations like choice tree classifiers, guileless Bayesian classifiers-closest neighbour classifiers, bolster vector machines and so forth. For example, characterization on host logs may manual for the exposure of concerning standards, for example, 40% of customers who laid an on-line arrange in item/Music are in the 19-26 age class and live on the west drift.

1.4.4 Successive Pattern Mining:

The procedure of successive pattern discovery endeavours to discover between session examples to such an extent that the nearness of an arrangement of things is trailed by another thing in a period requested arrangement of sessions or scenes. By utilizing this approach, web advertisers can foresee future visit designs which will be useful in setting promotions gone for certain client gatherings. Different sorts of transient investigation that can be performed on consecutive examples incorporates incline examination or change point recognition. Slant investigation can be utilized to recognize changes in the use patterns of a site after some time and change point location distinguishes when particular changes happen.

Agrawal and Bhatia (2015) researched about the methodology called the pattern discovery used for tailoring the various websites and their contents for the users gain information. The information is found in large collections in the web which can be navigated to customer's services by sending appropriate data that the users might want to discover. The pattern mining is responsible to the additional information acquired from the websites according to the requirements of data for the users. The three categories of the web mining are web usage mining web structure mining and web content mining. All the categories of the mining are personalized and recommended by the performances of the users.

B. Santhosh Kumar, K. V. Rukmani (2010) focused on web usage mining and specifically on finding the web usage patterns of websites from the server logs. The comparison of memory utilization and time utilization is compared using Apriori algorithm and Frequent Pattern Growth algorithm. The fundamental disadvantage of Apriori algorithm is that the candidate set generation is expensive, particularly if a huge number of patterns and/or long patterns exist. The fundamental disadvantage of FP-growth algorithm is the absences of good candidate generation method.

Saxena and Shukla (2010) explained clearly about the web log data and their pattern discovery in the significant interval of time in the process of sequence mining. This pattern discovery can be analysed by the significant use of algorithms with the traditional pattern detection approaches. Pattern discovery can be discovered by the Significant Interval Discovery algorithm (SID) which has the following intervals: unit significant interval, disjoint

significant interval and overlapping significant interval.

Zhong, et. al., (2012) proposed effective techniques in the text mining with the effective discovery of patterns in the text documents. The data mining approaches of data has attracted the attention of the digital world which is rapid in its growth and can be benefitted by the knowledge and data gained in the business field. The effective pattern discovery has two main fundamental issues called the misinterpretation of pattern and low frequency. Patterns can be effectively studied with the text mining by the variety of algorithms like FP tree, SPADE (Sequential Pattern Discovery using Equivalence classes), Apriori algorithms.

Kayuturk, et. al., (2005) expresses about the discrete attributes of the pattern discovery with compression and clustering. The data analysis is threatened with challenges when the data sets have high frequency in various applications. The two key technologies of analyzing the pattern discovery datasets are compression and sampling method. While comparing with the data analysis and pattern discovery the reduction in data are the main scope of latent structures discovery and matrix decomposition. Instead of considering the expensive traditional algorithms the patterns can be presented by the technique called the PROXIMUS which is accuracy oriented. The main properties of the PROXIMUS are interpretation in patterns, interesting discovery, highly capable in compact performance, large datasets scalability, and efficient at runtime for the data pattern discovery.

According to Charjan and Pund, (2013) the algorithms for the pattern discover in text mining are used in various pattern mining techniques which are used in search of interesting patterns. The patterns can be determined by the use of the effective, interesting and relevant information being deployed from the process of pattern innovation and improvising pattern evolution. The process of knowledge development model, like data selection, data processing, data transacting and pattern discovery are always being extracted from the documents of text. So the evolution of the discovered patterns is correlated with the knowledge patterns which are interesting and associated with mining of rule. The knowledge discovery has the ability to find the worthiness of information in pattern discovery of the uses text documents.

Sundari, R. and et. al., (2014) stated about the techniques for pattern discovery by using web mining process with great advancement in the revolutionary technology. The information retrieval can be made easy with the use of pattern discovery. The pattern discovery can be included with the knowledge gain of web log files and recognized the web based useful information for the user interests. Pre-processing of data with the knowledge extraction is ensure to process the data loading, accuracy checking , gaining data transformation and structuring data ,which is based on the data mining algorithms and analysis.

B. Uma Maheswari and Dr. P.Sumathi (2015), compared two standard web usage mining algorithms namely Apriori algorithm and Frequent Pattern algorithm. Particularly, they focused on discovering the web usage patterns of websites from the server log files.

Mr. Ashish Vitthalrao Galphade and Mr. Dhiraj Bhise (2016), analyzed the association rule based algorithms specifically Apriori algorithm, which address the needs of different web service providers and different viewers, clients, business analysts, and so forth. It enhances the techniques of Web Usage Mining by first finding the log files of individual users at one place. This collective knowledge can be utilized to outline business techniques to boom sales.

1.5 SEQUENTIAL PATTERN MINING

Finding consecutive patterns from immense gathering databases is a basic issue in the field of knowledge discovery and data mining. Agrawal and Srikant (1995) first introduced the preliminary idea of sequential pattern mining and can be immediately communicated as follows. A course of action of groupings called data progressions are given as the data. Each data course of action is an once-over of trades, where each trade contains a game plan of literals, called things. Right when a customer showed minimum support edge is given back to back pattern mining finds most of the consecutive sub sequences in the plan database, i.e. the subsequence's whose extents of appearance outperform the base support constrain. Of late, successive pattern mining has been broadly associated with a couple application spaces, for example, grandstand case data examination, pharmaceutical, Web log examination, media interchanges, et cetera. In the retailing business, successive patterns can be mined from the trade records of customers.

In Web log examination, the researching behaviour of a customer can be removed from part records or log archives, For example, having obtained thing A on a dealing webpage, more than 80% of customers to buy thing B. For another representation, having seen a site page on "Data Mining", customers returns to research "Business Intelligence" for new data next time. These consecutive patterns yield huge focal points and when followed up on, augmentation customer ways.

1.5.1 Efficiency of Sequential Pattern Mining:

A ref. and et. al., (2004) author's focuses on the capable mining of sequential patterns in Web utilize data .Since the measure of the readied data in mining successive pattern tends to be gigantic, it is basic to devise profitable networks to mine such data. On a

very basic level, it can be subdivided into two essential procedures:

(1) Improve the viability by sketching out novel calculations

(2) Improve the viability by giving bolster segment.

The essential approach is to layout compelling calculations. Generally speaking, these calculations can be sorted into three classes:

(1) Apriori-based, level sorting out methodologies, for example, GSP proposed by authors Srikant and Agrawal (1996).

(2) Apriori-based, vertical sorting out networks, for example, SPADE proposed by Zaki et. al., (2001)

(3) Projection-based example improvement procedures, for example, Pre-fix Span given by authors Pei et. al., (2001)

The second approach is to create bolster segment of back to back examples. In reality, applications on continuous example mining, customers constantly face the going with two conditions:

(1) Sequence database will be invigorated with new trades; by author Robert cooley (1997, 1999).

(2) User can't find fitting support constrain immediately and reliably tune its reinforce regard dependably.

In any case, the two may provoke the change of continuous examples. It pushes to plot the upkeep segments to beneficially uncover outlines under these conditions without rescanning the plan database and rerunning the whole mining process.

1.5.2 Apriori-Based Mining Algorithms:

This conventional calculation includes three stages for mining the progressive chain designs (Agrawal and Srikant, 1994). At first it tries to find all the repetitive thing sets that followed the token sets with confirm more prominent than the negligible attest. In the following stride it replaces the real exchanges with the all repetitive thing sets set comprised by the exchange. Finally, the progressive chain examples are resolved. This is expensive calculation as it needs to supplant the exchanges at every last stride and to handle the conditions for the eras and the scientific categorizations it is a perplexing errand. The meaning of progressive chain designs that were made general so that comprises of day and age conditions, moving windows and customer said scientific classification are proposed (Agrawal and Srikant, 1995). A more broad progressive chain design mining calculation known as GSP (Generalized Sequential Pattern mining calculation) is additionally proposed. Same as to that of the traditional Apriori calculation, GSP follows out the database commonly is discussed in our next section.

1.5.3 GSP Algorithm

Srikant and Agrawal (1996) in his paper discuss the GSP calculation, grasps a various pass, confident time and-test approach in consecutive pattern mining. This is laid out as takes after. In the essential pass, find each and every normal thing to outline the game plan of single thing unremitting progressions. Each ensuing pass starts with a seed set of back to back examples, which is found in the past pass. This seed set is used to deliver new potential examples called confident progressions. Each contender gathering contains one more thing than a seed back to back example, where each segment in the example may contain a certain something or distinctive things. The amount of things in a gathering is known as the length of the game plan. In this manner, all the candidate progressions in a pass will have a comparable length. The scope of the database in one pass finds the support for each contender gathering. Each one of the hopefuls with minimum support in the database outlines the course of action of the as of late found successive pattern. This set is then used as the seed set for the accompanying pass. The generate- and-test process is repeated until no new successive pattern is found in a pass or no more new hopefuls are made.

Basically, the GSP calculation bears three inborn and non-minor costs. A monster course of action of confident groupings could be delivered in a broad progression database. Since k-cheerful courses of action (candidate progressions with k things) are made from each and every possible phase of the (k-1) gigantic examples, it may create a really far reaching game plan of contender groupings despite for an immediate seed set. For example, If there are 1000 ceaseless progressions of length-1, for example, <a1>, <a2>, … , <a1000>, an Apriori-like calculation will create 1000×1000 + (1000×999)/2 = 1,499,500 contender courses of action, where the important term is gotten from the set <a1a1>, <a1a2>, … , <a1a1000>, <a2a1>, <a2a2>, … , <a1000a1000>, and the second term is gotten from the set <(a1a2)>, <(a1a3)>, … , <(a999a1000)>.

Multiple database scans are required. Since the length of each applicant grouping creates by one at each database look at, the Apriori-based procedure must range the database in any occasion k times. Inconveniences at mining long progressive illustrations, this is by virtue of a long sequential

case must grow up from a massive number of short progressive cases, however the amount of such competitor courses of action delivered is exponential to the length of the successive cases to be mined.

1.5.4 Spade Algorithm

The SPADE Algorithm Besides the level masterminding network (GSP), the succession database can be changed into a vertical course of action containing things' id-list. The id-once-over of a thing is an once-over of (game plan id, timestamp) set exhibiting the incident timestamps of the thing in that progression. Looking in the network formed by id-list crossing focuses, the SPADE (Sequential Pattern Discovery using Equivalence classes) calculation by authors Zaki and et. al. (2001) got done with mining in three goes of database analyzing. Incidentally, additional calculation time is required to change a database of level plan to vertical design, which in like manner requires additional storage space a couple times greater than that of the main progression database.

1.5.5 Pre Fix-Span Algorithm

The Prefix-Span (Prefix-foreseen Sequential illustration mining) calculation by authors Pei et. al. (2001), addressing the case advancement approach, finds the nonstop things in the wake of sifting the grouping database once. The database is then expected, according to the relentless things, into a couple of smaller databases. Finally, the whole course of action of progressive cases is found by recursively creating subsequence pieces in each expected database.

Disregarding the way that the Prefix-Span calculation adequately discovered cases using the partition and-defeat network, the cost of memory space might be high a direct result of the creation and treatment of huge number of foreseen sub databases.

1.5.6 WAP-Tree Based Mining Algorithms:

A powerful decent approach with an expansive data structure with firmly coupled data is called Web Access Pattern Tree (or WAP-tree), which is FP-tree arranged is talked about (Pei, et. al., 2000). The WAP-tree structure gives the era of novel calculations for mining access designs conceivably utilizing an immense arrangement of web log parts. Particularly, the WAP-mining calculation has been recommended for mining web get to designs from WAP-tree. The present strategy annihilates the issue of creating gigantic number of competitors as observed in Apriori-sort of calculations. Aside from this, the results of the test are speaking to that the WAP-uncovering calculation is in detail a request of number significantly snappier than the regular progressive chain design mining approaches. This can be given the credit to the firmly coupled development of WAP-tree and the new speculative discovering techniques agreeable in WAP-mining.

WAP-tree is a decent firmly coupled to keep up Data from network logs. WAP-mine is the main removal calculation grounded on WAP-tree that does not render expansive number of hopeful sets as were delivered in the traditional Apriori-based calculations. Be that as it may, the production of the in the middle of limitations of the WAP-tree while the mining is in process is expensive. At present, certain future works are under process for the WAP-tree and the related mining calculations.

1.5.7 Pre-Order Linked WAP-Tree:

Mining (PLWAP) calculation which does not create the WAP-mine in the middle of level limitation WAP-trees by organizing the double place figures to all tree hubs [64]. The PLWAP calculation follows out quickly the postfix trees or woodlands of any append token of rehashed blue prints by adapting to the arranged double place figures of hubs. The Binary Cipher Formatting (Tre BCF) strategy is later used to introduction of alone parallel place figures to handles of any widespread tree, by at first trading the tree into its paired tree of same kind and using a standard comparative that is used in Huffman coding to portray an alone figure for every handle.

The RSC-tree (Recurrent Successive Chain Tree) develops the WAP-tree skeleton for constantly developing and easy to understand mining (Lu and Ezeife, 2003). The mining calculation RSC-digging is useful for breaking down the RSC-tree for separating intermittent progressive chains. The proposed RSC-Miner framework can utilize the new info progressive chains and give the reaction for continually developing without doing complete count. The framework likewise gives an opportunity to the customers to change include groupings (e.g. least support and required example length) easy to understand without the fundamentally total recalculation of most the cases. The constantly developing altering limit of the framework gives extremely potential execution points of interest over total recalculation notwithstanding for most immense adjusting lengths. Numerous methods are used to remove successive example mining from web logs. The standard frameworks can be ordered into Apriori-based, plan advancement and early-pruning techniques. Apriori-based calculations are regarded direct and have tremendous interest space, while outline improvement calculations have been attempted broadly on mining the web log and seen to be speedy Early-pruning methodologies have examples of beating misfortune with web get to progressions secure in thick databases.

1.6 EXTENDED SPM WITH OTHER TIME RELATED PATTERNS

Influenced by the potential applications for the back to back cases, different developments of the hidden definition have been proposed which may be related to various sorts of time-related illustrations or to the extension of time prerequisites. Mannila et. al., (1997) [66]Basic research in this class consolidates finding progressive scenes in event game plans, finding nonstop traversal outlines in a web log (Chen et. al., 1998), finding irregular cases and rehashing outlines in a period stamped trade database (Han et. al., 1999), finding different dimensional back to back cases in data stockrooms (Pinto et. al., 2001), mining cross breed successive cases from grouping data (Chen et. al. 2002) and discovering time-between time sequential cases in arrangement databases (Chen et. al., 2005).

Finding consistent scenes in event courses of action: The possibility of normal scene mining was first proposed by Mannila et. al., (1997) A scene is a collection of events in a lone long arrangement that happen by and large close to each other in a partial demand. This audit portrays two sorts of scenes: serial scene and parallel scene. A serial scene evades to a case in which a user specified window width wins, things a, b and c happen in a particular demand, however confined by cloud time intervals. A parallel scene evades to a case in which things a, b and c happen in win, however in a dark demand and disengaged by cloud between times. A level-wise calculation is moreover prepared to locate each and every progressive scene in light of moving a period window over the data arrangements. (Mannila et. al., 1997) furthermore extended their work to discover summed up and fascinated scenes, which empowered one to communicate self-assured unary conditions on individual game plan events or parallel conditions once in a while joins. The calculations researched the oversee space particularly instead of the succession space.

Finding way traversal outlines: (Chen et. al., 1998) In World Wide Web (WWW), Web pages are regularly associated together to urge smart get to. Customers research the data they are enthusiastic about beginning with one page then onto the following by methods for the relating hyperlinks or URL addresses. An unmistakable perception of customer get to plans in web won't simply help upgrading the Web server diagram (e.g., giving capable access between significantly associated s, better making arrangement for pages, et cetera.) moreover incite better exhibiting decisions (e.g., putting advancements in suitable spots, better customer/customer portrayal and direct examination, et cetera.). Getting customer get to outlines in such conditions is suggested as mining way traversal plans (Chen et. al., 1998).

Finding incidental cases/cyclic illustrations: Periodic case exposure Han et. al., (1999) can be communicated as issues of finding cases occurring at standard between times. Basically, it focuses on two sections of the issue – case and between times. With a game plan of events given, the researchers might need to find the cases which go over after some time and their rehashing periods, For instance, given a business database which records bargains data of an association over a period of ten year. A yearly arrangements plan in these ten years in light of the month to month compacted data, After some examination, can be found the wage of particular things reaches their yearly most noteworthy each September, This kind of illustration can be seen as a discontinuous case. Regardless, as a rule outlines don't go over in an ordinarily isolated time between times, for instance, hourly, step by step, month to month, et cetera. A man's heart does not much of the time beat in a period describable by between times in minute hour, or something to that effect. In this way, the reiterating cases of a progression and the break which thinks about to the illustration time span are similarly to be tended to when the business database is given.

Multi-dimensional progressive illustration mining: The possibility of multidimensional back to back case mining was first exhibited by Helen Pinto et. al., (2001). In Mining progressive cases with single estimation, consider the trademark close by time stamps in illustration disclosure get ready while in mining back to back cases with different estimations, consider various characteristics meanwhile. Instead of continuous illustration mining in single estimation, mining in various dimensional progressive cases can give us more instructive and profitable cases. For example, get an ordinary progressive case from the store database. In the wake of buying thing "a" most by far similarly buy thing "b" in a described time between times.

Nevertheless, to utilize different dimensional progressive illustration mining, find assorted social events of people with different purchase outlines. For example, understudies constantly buy thing "b" after they buy thing 'a', while this progressive control weakens for various social affairs of understudies. Along these lines, we can see that diverse dimensional back to back case mining can give more exact data to further decision bolster.

A back to back case mining finds periodically happening outlines asked for by time Agrawal and Srikant (1995). An instance of a progressive case in perspective of Apriori property communicates that: "All nonempty subsets of a general item set ought to in like manner be visit". It is in like manner portrayed as being ant monotonic (or plunging closed). If a game plan can't easily get through the base support test, most of its super progressions will similarly fail

the test. Calculations that depend generally on the Apriori property, without taking further exercises to restrict the chase space, have the hindrance of keeping up the support mean each subsequence. This is being mined and testing this property in the midst of each accentuation of the calculation makes them computationally expensive.

To beat this issue, calculation need to make sense of how to find out support and prune applicant arrangements without numbering backing and keeping up the check in each cycle. Given a database of successions, where every course of action is an once-over of trades asked for by trade time, and each trade is a game plan of things. The issue was penniless around Srikant and Agrawal (1996). The issue of mining progressive cases is to locate each sequential case with a customer showed minimum support, where the support of a case is the amount of data groupings that contain the illustration. They showed GSP calculation that finds summed up continuous cases. Savasere et. al., (1995) proposed a methodology for database separating. The course of action database is allocated a couple disjoint parts that can fit into rule memory, each part is mined freely, in conclusion every single progressive gathering found merged from all parts.

Masseglia et. al., (1999) proposed a PSP calculation. A prefix tree is used to hold candidate groupings close by reinforce mean every course of action toward the complete of each branch that addresses it. This network (nearby GSP from which it is gotten) ends up being to a great degree inefficient when support utmost is low, settling on it a poor choice for web log mining. Zaki (2001) exhibited SPADE calculation for snappy disclosure of Sequential Patterns. The present responses for this issue make repeated database yields and use complex hash structures which have poor region. SPADE utilizes combinatorial properties to break down the main issue into humbler sub-issues that can be unreservedly handled in central memory using profitable matrix look techniques and using fundamental join operations. All groupings are found in three database channels in a manner of speaking. FreeSpan calculation by Han et al (2000)[73] uses expected databases to create database remarks that guide the mining methodology to find visit plans faster. Its foreseen database has a contracting variable altogether not as much as Prefix-Span.

1.7 ISSUES IN DATA PRE-PROCESSING

The following literature review will exclusively and intensively explain about the studies that dealt with issues in data pre-processing techniques in web usage mining. The issues/ difficulties in data cleaning have not been discussed in detail by the authors in their study. The issue in the identification of user has been denoted as a significant issue since it is necessary to differentiate the IP address of each individual. The main issue the author stated in this study is about 'personal information' login details that many users have been ignoring for accessing data (i.e. without registration) henceforth finding the user accessing relevant information or session is a tedious task and making the process difficult. Hence the authors had proposed DUI (Distinct User Identification) algorithm to retrieve user identification.

Authors had mentioned that the data pre-processing is a time consuming and hence it poses a huge issues (Raiyani and Jain, 2012). According to the authors, data cleaning under their proposed algorithm would be more effective than the traditional algorithms. Difficulties in User and Session identification have been studied by the authors and they found that web log data is essential to identify the user's information. KNN (K-nearest neighbour) and PCA (Principal Component Analysis) algorithm was proposed by the authors to differentiate data and to map filter information for faster pattern discovery purposes.

Authors had mentioned that the data pre-processing process consumes lot of time and hence altering algorithm would be an effective measure (Vishwakarma and Singh, 2014). In this paper the authors have studied about the issues in data cleaning process. According to them in order to remove the errors (data cleaning), basically the algorithm has to check for HTTP status code and the records found under the status codes of 200 or over 299 will be removed. The user identification along with the session identification has a different issue where the users sometimes access information without logging in with their registered information; hence identifying a user with the IP address will be difficult task. The web log consists of data accessed for each session as per the web pages time-oriented or sometimes the structure -oriented limitations. The authors had studied about the time consumption for overall process of data pre-processing and they had found that: accessing raw web log, cleaning the data, finding the users or the unique session users, are the factors that are to be considered for each sub-process (Raiyani, Jain and Raiyani, 2012). The authors studied about data cleaning and they had stated that it is quite time consuming process where the data has to be cleaned especially if the data is in the form of pictures, videos, audios, and so on. On contrary the authors studied about the user identification and session identification. According to them web log mining for identifying users will consume longer time. Overall as per their view the data processing consumes more time since it has to overcome: data cleaning, user identification, session identification, path completion and transaction identification. Hence to resolve this issue they formulated an algorithm where the traditional algorithm and a dynamic algorithm are collaborated into one to refine the session identification time-out issue to identify the user through web log mining (Patel and Parmar, 2014).

Hanane Ezzikouri, Mohammed Erritali and Mohamed Oukessou (2015), mentioned web data pre-processing required most time, due to the lack of structuring and the large amount of noise present in the raw data. The first stage of web usage mining process, which is obviously Pre-processing, occupies about 60% to 80% of the time involved in the whole process. Pre-treatment of Web log files is to clean and organize the data contained in these files to prepare them for future analysis. Data Cleaning shown that significant reduction (up to 70%) of the initial number of requests and offers richer structured logs for the next step of data mining.

Khushbu Patel, Anurag Punde, Kavita Namdev, Rudra Gupta and Mohit Vyas (2015) worked on survey of web mining. The aim of the paper is to provide past and current techniques in Web Mining. They discussed about research work done by different researchers and important research issues related to it .Moreover, they expressed Web cleaning is the important process and it becomes difficult when it comes to heterogeneous data. The accuracy of data needs to be concentrated. According to researchers 70% of the time is spent on data pre-processing. As per the author's view the web log for GIF, CSS, JPEG in the URI field will consume more time for data cleaning where the algorithm has to examine from HTTP status codes. In the status code field, if found status error is under 200 or over 299, then the errors are removed through the structured algorithm. The time for identifying the user and session varies according to the authors. The user identification could be done through session identification however the session identification has to be found through examining the IP address, use of an operating system and the browser. Hence session identification consumes more time than user identification. Hence they used Distinct User Identification algorithm to improve the overall designing and performance of upcoming access of pre-processing results.

(Raiyani and Pandya, 2012). The authors mentioned data cleaning, user identification and session identification techniques. In the data cleaning phase, unnecessary records including graphics files, robots are removed. The records resulted after cleaning phase is 1476 from 9464 records. After the data cleaning process is performed, users are identified by using IP address and User-Agent fields. The next process is the identification of sessions which is derived by forming the user behaviour in matrix format (Chitraa and Thanamani, 2011).

1.8 CONCLUSION:

The paper focus on the similar developments and research works on web usage mining that consists of web usage data, preprocessing tasks, and the several pattern extraction approaches are discussed. The numerous strategies in successive patterns mining have been proposed. At the point when all is said in done, there are two basic research issues in sequential pattern mining. The first is to improve the profitability in back to back pattern mining process while the other one is the extension of mining of sequential pattern to other time-related patterns .To preprocess the network utility information, the operation comprises of information stripping, client recognition and session recognition .

Analytical approaches are more specifically used to extract analytical judgment using the web logs. This type of judgment is vastly used to analyze network dealings of a website.

Grouping approach is very helpful and useful to extract page groups and client groups of network logs. Varlet radicals are utilized to improve look up locomotive engine and to supply network classification while the client groups are useful to infer the client demographics so as to give individualized web subject to browsers.

Ordering is the task to anticipate an information token into one of the various already defined groups. These groups generally used for representing several client profiles, and classification is carried out depending upon the characteristics which better depict the lineaments of a particular group or class.

Consecutive blue print channeling is a hereafter exercise of affiliation principle excavation by considering the consecutive of encountering of tokens in events.

Consecutive blue prints are consecutive network varlets accessed often by users. These kinds of the patterns are helpful to extract the client behavior and expecting the next pages to be browsed by the client.

REFERENCES

Abraham M. and Puthiyidam. J. (2016). "A Survey on Wind Data Pre-Processing in Electricity Generation", International Journal on Cybernetics & Informatics", International Journal on Cybernetics & Informatics, Vol-5, Issue-2, pp. 407- 415.

Agrawal, R. and R. Srikant (1994). "Fast algorithms for mining association rules, Proceedings of 1994 International Conference Very Large Data Bases", pp. 487-499.

Agrawal, R. and R. Srikant (1995). "Mining sequential patterns", Proceedings of 1995 International Conference Data Engineering, pp. 3-14.

Agrawal, R. and R. Srikant (1995). "Mining sequential patterns", Proceedings of 1995 International Conference Data Engineering, pp. 3-14.

Aref, W.G., Elfeky, M.G. and Elmagarmid, A. K. (2004). "Incremental, online, and merge mining of partial periodic patterns in time-series databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 3,pp. 335-345.

Bonchi, F., Giannotti, C., Gozzi, G., Manco, M., Nanni, D., Pedreschi, C., Berendt, B., Mobasher, B., Nakagawa, M. and Spiliopoulou, M. (2002). "The Impact of Site Structure and User Environment on Session reconstruction in Web Usage Analysis," In Proceedings of the Forth WebKDD 2002 Workshop, At the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2002), Edmonton, Alberta, Canada, pp.1-13.

C. E. Dinucă (2012). "An Application for Clickstream Analysis", International Journal of Computer and Communication, ISSN: Vol-6, Issue-1, pp. 68-75.

C. Kwok and D. Weld (1996). Planning together information. In Proc. 14th National Conference on AI.

C.M. Brown, B.B. Danzing, D. Hardy, U. Manber, and M.F. Schwartz (1994). The harvest information discovery and access system. In Proc. 2nd International world Wide Web Conference.

Chen, M.S., Park, J.S. and Yu, P.S. (1998). "Efficient data mining for path traversal patterns", IEEE Transactions on Knowledge and Data Engineering, Vol. 10, No.2, pp.209-221.

Cleve. J. et. al. (n.d)." Data Mining on Transaction Data", Wismar University, Germany, pp. 1-8.

Cooley, R. (2000). "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data", PhD thesis, University of Minnesota.

Cooley, R., Mobasher, B. and Srivastava, J. (1997). "Web mining: Data and pattern discovery on the World Wide Web", International Conference on Tools with Artificial Intelligence, Newport Beach, IEEE, pp. 558-567.

Cooley, R., Mobasher, B. and Srivastava, J. (1999). "Data Preparation for Mining World Wide Web Browsing Patterns", In Journal of Knowledge and Data Networks, Vol. 1, No. 1, pp. 5-32.

D. Konopnicki and O. Shmueli (1995). W3qs:A query System for world wide web. In Proc. of the 21st VLDB Conference, Pages 54-65, Zurich.

Dhawan S. and Goel S. (2013). "Web Usage Mining: Finding Usage Patterns from Web Logs", American International Journal of Research in Science, Technology, Engineering and Mathematics, ISSN (Online): 2328-3580, Vol-2, Issue-2, pp. 203-207.

Dinuca. E. and Ciobanu D. (2011). "Improving the Session Identification Using the Mean Time", International Journal of Mathematical Models and Methods in Applied Sciences, Vol- 6, Issue-2, pp. 265-272.

E. Spertus (1997). Parasite :Mining structural information on the web. In Proceeding of 6th International World Wide Web Conference.

Ezeife, C.I., Lu, Y. and Liu, Y. (2005). "PLWAP Sequential Mining: Open Source Code", Open Source Data Mining Workshop on Frequent Pattern Mining Implementations, in conjunction with ACM SIGKDD, Chicago, IL., U.S.A, pp. 26-29.

Ganti V. and Sarma A. (2013). Book on "Data Cleaning, a Practical Perspective", Morgan and Claypool Publishers, pp. 1-63.

Grace et. al., (2011). "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security & Its Applications (IJNSA), DOI: 10.5121/ijnsa.2011.3107, Vol-3, Issue-1, pp. 99-110.

Halfaker A., et. al. (2015). "User Identification Based on Strong Regularities in Inter Activity Time", International World Wide Web Conference Committee, ACM 978-1-4503-3469-3/15/05, pp. 410-418.

Han, J., Dong, G. and Yin, Y. (1999). "Efficient mining of partial periodic patterns in time series database", Proceedings of 1999 International Conference on Data Engineering, Sydney, Australia, pp. 106-115.

Han, J., Dong, G. and Yin, Y. (1999). "Efficient mining of partial periodic patterns in time series database", Proceedings of 1999 International Conference on Data Engineering, Sydney, Australia, pp. 106-115.

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U. and Hsu. M.C. (2000). "FreeSpan: Frequent Pattern-Projected Sequential Pattern Mining",

In Proc. 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD'00), Boston, MA, pp. 355-359.

Helen Pinto, Han, J., Pei, J., Wang, K., Chen, Q. and Umeshwar Dayal (2001). "Multi-dimensional sequential pattern mining", Proceedings of the 10th International Conference on Data and Knowledge Management (CIKM 2001), Atlanta, Georgia, pp. 81-88.

Hellerstein M. (2008). "Quantitative Data Cleaning for Large Databases", United Nations Economic Commission for Europe (UNECE), pp. 1-41.

Hongzhou Sha, Qingyun Liu (2013). "EP Log Cleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining", Information Technology and Quantitative Management, Elsevier Itqm Procedia Computer Science,ISSN:812 – 818 ,pp.812-818.

Huysmans, J., Baesens, B. and Vanthienen, J. (2004). "Web Usage Mining: A Practical Study", In Proceedings of the Twelfth Conference on Knowledge Acquisition and Management (KAM2004), Kule (Poland), May 13-15.

Jiadi. Z. and Hai G. (2016). "Research on User Identification Algorithm Based on Rewriting URL", International Journal of Security and Its Applications, Vol.-10, Issue-3, pp. 215 -222.

Joslyn C., et. al. (n.d), "Transaction Hyper Graph Models for Pattern Identification in Bitcoin Blockchain", WA 98109, pp. 1-4.

K. Hammond, R. Burke, C. Martin, and S. Lytinen (1995). Faq-Finder: A case-based approach to knowledge navigation. In working notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environment. AAAI Press.

K.A. Oostendorp, W.F. Punch and R.W. Wiggins (1994). A tool for individualizing the web. In Proceeding of 2nd International World Wide Web Conference.

Kamat M. S., Bakal J. W. and Nashipudi M. (2013). "Improved Data Preparation Technique in Web Usage Mining", International Journal of Computer Networks and Communication Security, ISSN 2308-9830), Vol-1, Issue-7, pp. 284-291.

Kherwa P. and Nigam J. (2015). "Data Preprocessing: A Milestone of Web Usage Mining", International Journal of Engineering Science and Innovative Technology, ISSN: 2319-5967, Vol-4, Issue-2.

Krishnan. S, et. al. (2016). "Towards Reliable Interactive Data Cleaning: a User Survey and Recommendations", ACM ISBN: 978-1-4503-4207-0, pp. 1-5.

Losarwar. V. and Joshi M. (2012). "Data pre-processing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems, pp. 1-5.

Lu, Y., and Ezeife, C. I. (2003). "Position Coded Pre-Order Linked WAP-tree for Web Log Sequential Pattern Mining", In proceedings of the seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining, Seoul, Korea, April 30-May 2, 2003, published in LNCS by Springer Verlag, pp. 337-349.

M. Balabanovic, Yoav Shoham, Y. Yun (1995). An adaptive agent for automated web browsing. Journal of Visual Communication and Image Representation 6(4).

M. Pazzani, J. Muramatsu, and D. Billsus, Syskill and Webert (1996). Identifying interesting web sites. In Proc. AAAI Spring Symposium on Machine Learning in Information Access, Portland, Oregon.

Makwana. C and Rathod K. (2014). "An Efficient Technique for Web Log Pre-processing Using Microsoft Excel", International Journal of Computer Applications, ISSN 0975-8887, Vol-90, Issue-12, pp. 25-28.

Mannila, H., Toivonen, H. and Inkeri Verkamo, A. (1997). "Discovery of frequent episodes in event sequences", Data Mining and Knowledge Discovery, Vol. 1, No. 3, pp. 259-289.

Marquardt, C., Becker, K. and Ruiz, D. (2004). "A Pre-processing Tool for Web usage Mining in the Distance Education Domain", In Proceedings of the International Database Engineering and Application Symposium (IDEAS' 04), pp. 78-87.

Masseglia, F., Poncelet, P. and Cicchetti, R. (1999). "An efficient algorithm for web usage mining", Networking and Data Networks Journal (NIS), pp.571-603.

Masseglia, F., Poncelet, P. and Cicchetti, R. "An efficient algorithm for web usage mining",

Networking and Data Networks Journal (NIS), pp.571-603, 1999.

Mitali Srivastava, Rakhi Garg, P. K. Mishra (July 2014). "Preprocessing Techniques in Web Usage Mining: A Survey", International Journal of Computer Applications ,ISSN:0975 -8887, Vol-97, Issue-18, pp. 1-9.

Mitharam M. D. (2012). "Preprocessing in Web Usage Mining", International Journal of Scientific and Engineering Research, ISSN 2229-5518, Vol-3, Issue-2, pp. 1-7.

Muskan, Dr. Kanwal Garg (2016). "An Efficient Algorithm for Data Cleaning of Web Logs with Spider Navigation Removal", International Journal of Computer Application, ISSN-2250-1797, Vol-6, Issue-3,pp. 6-12.

Nigrel S. et. al. (2015). "Web Log Pre-processing for Web Usage Mining", International Journal for Scientific Research and Development, ISSN: 2321-0613, Vol-2, Issue-12, pp. 604-606.

P. Resnik, N. Iacovou, M. Sushak, P. Bergstrom and J. Riedl (1994). Grouplens: An Open architecture for collaborartive filtering of net news. In Proc. Of the 1994 Computer supported Cooperative work Conference, ACM.

Padala. V, et. al. (2013). "A Novel Method for Data Cleaning and User Session Identification for Web Mining", International Journal of Modern Engineering Research, ISSN: 2249-6645, Vol-3, Issue-5, pp. 2816-2819.

Patel K. et. al. (2015). "Detailed Study of Web Mining Approaches-A Survey", International Journal of Engineering, Sciences and Research Technology, ISSN: 2277-9655, Vol-4, Issue-2, pp. 23-30.

Patel P. and Parmar M. (2014). "Improve Heuristics for User Session Identification Through Web Server Log in Web Usage Mining", (IJCSIT) International Journal of Computer Science and Information Technologies, ISSN:0975-9646, Vol-5, Issue-3, pp. 3562-3565.

Patel P. and Parmar M. (2014). "Improve Heuristics for User Session Identification through Web Server Log in Web Usage Mining", (IJCSIT) International Journal of Computer Science and Information Technologies, ISSN:0975-9646, Vol-5, Issue-3, pp. 3562-3565.

Pei, J., Han, J., Mortazavi-Asl, B. and Zhu, H. (2000). "Mining access patterns efficiently from web logs", in PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer-Verlag, pp. 396-407.

Pei, J., Han, J., Pinto, H., Chen, Q., Dayal, U. and Hsu, M.C. (2001). "Prefix Span: mining sequential patterns efficiently by prefix-projected pattern growth", Proceedings of 12th International Conference on Data Engineering, Heidelberg, Germany, pp. 215-224.

Pei, J., Han, J., Pinto, H., Chen, Q., Dayal, U. and Hsu, M.C. (2001)."Prefix Span: mining sequential patterns efficiently by prefix-projected pattern growth", Proceedings of 12th International Conference on Data Engineering, Heidelberg, Germany, pp. 215-224, 2001.

Priyanka P. and Naveen N. C. (2015). "User Identification, Classification and Recommendation in Web Usage Mining- An Approach for Personalized Web Mining", International Journal of Innovative Science, Engineering and Technology, ISSN 2348 – 7968, Vol-2, Issue-4, pp. 1021-1030.

R. Armstrong, D. Frietag, T. Joachims, and T. Mitchell (1995). Webwatcher: A learning apprencite for the world wide web. In Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous Environment.

Rejena. S. and Malika R. (2016). "Innovative Pre-Processing Technique and Efficient Unique User Identification Algorithm for Web Usage Mining", International Journal of Advanced Research Computer Science and Software Engineering, ISSN: 2277 128X, Vol-6, Issue-2, pp. 85-91.

Robert Cooley, Bam Shad Mobasher, and Jaideep Srivastava (1997). "Grouping Web page references into transactions for mining World Wide Web browsing patterns", Knowledge and Data Engineering Workshop, Newport Beach, CA.IEEE, pp. 2-9.

Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava (1999). "Data preparation for mining World Wide Web browsing patterns", Journal of Knowledge and Information System, pp. 1-27.

S. Prince Mary, E. Baburaj (2013). "An Efficient Approach To Perform Pre- Processing." Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166, Vol-4, Issue-5, pp. 404-410.

Savasere, A., Omiecinski, E. and Navathe, S. (1995). "An efficient algorithm for mining association

rules in large databases", In Proceedings of 1995 International Conference on Very Large Databases (VLDB'95), Zurich, Switzerland, pp. 432-443.

Sergey Brin and Lawrence Page (1998). The anatomy of a large-scale hyper textual web search engine. pages 107– 117.

Shaily Langhnoja, Mehul Barot, Darshak Mehta (2012). "Pre-Processing: Procedure on Web Log File for Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Vol-2, Issue-12, pp. 419-423.

Singh S. and Badhe V. (2014). "An Exclusive Survey on Web Usage Mining for User Identification", International Journal of Innovative Research in Computer and Communication Engineering, ISSN (Online): 2320-9801, Vol-2, Issue-11, pp. 6582-6589.

Soumen Chakrabarti, Byron Dom, David Gibson, Jon M. Kleinberg, Prabhakar Raghavan and Sridhar Rajagopalan (1997). Automatic resource compilation by analyzing hyperlink structure and associated text. In Proc. 7th WWW, pages 65–74.

Spiliopoulou, M. (1999). "The laborious way from data mining to web mining", Journal of Computer Networks Science and Engineering, Special Issue on Semantics of the Web, Vol. 14, pp. 113-126.

Srikant, R. and Agrawal, R. (1996). "Mining Sequential Patterns: Generalizations and Performance Improvements", In Proceedings of the 5th Int'Conference on Extending Database Technology: Advances in Database Technology, LNCS 1057, pp. 3-17.

Srikant, R. And Agrawal, R. (1996). "Mining Sequential Patterns: Generalizations and Performance Improvements", In Proceedings of the 5th Int'l Conference on Extending Database Technology: Advances in Database Technology, LNCS 1057, pp. 3-17.

T. Kirk, A.Y. Levy. Y. Sagiv and D. Srivastava (1995). The information Manifold. In working notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environment. AAAI Press.

Talakokkula A. (2015). "A Survey on Web Usage Mining, Applications and Tools", Computer Engineering and Intelligent Systems, ISSN 2222-2863 (Online), Vol-6, Issue-2, pp. 22-29.

Tamrakar L. and Ghosh S. M. (2014). "Identification of Frequent Navigation Pattern Using Web Usage Mining", International Journal of Advanced Research in Computer Science & Technology, ISSN: 2347 - 8446 (Online), Vol-2, Issue-2, pp. 296-299.

Tomar D. and Agarwal S.(2014). "A Survey on Pre-processing and Post processing Techniques in Data Mining", International Journal of Data Base Theory and Application, ISSN: 2005-4270, Vol-7, Issue-4, pp. 99-188.

U. Shardanand and P. Maes (1995). Social information filtering : Algorithm for automating "word of Mouth" In Proc. Of 1995 Conference on Human Factors in Computing Systems (CHI-95), Pages 210-217.

V. Vidya, Priya S., Kalaivani (2015). "An Efficient Clustering Technique for Weblogs", IJISET-International Journal of Innovative Science, Engineering & Technology , ISSN:2348 - 7968, Vol-2, Issue-7, pp. 516-525.

Vellingiri J. and Pandian C. (2011). "A Novel Technique for Web Log Mining with Better Data Cleaning and Transaction Identification", Journal of Computer Science, DOI: 10.3844/jcssp.2011, Vol-7, Issue-5, pp. 683-689.

Vijayashri Losarwar, Dr. Madhuri Joshi (July 15-16). "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems, Singapore. pp. 1-6.

W. B. Frakes and R. Baeza-Yates (1992). Information Retrieval Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, N.J.

Yang, Z., Wang, Y. and Kitsuregawa, M. (2007). "LAPIN: effective sequential pattern mining algorithms by last position induction for dense databases", In Advances in Databases: Concepts, Networks and Applications, LNCS 4443, pp. 1020-1023.

Zaki, M.J. (2001). "SPADE: An Efficient Algorithm for Mining Frequent Sequences", Machine Learning, Vol. 42, pp. 31-60.

Corresponding Author

Aarti Pandey*

Research Scholar, Dept. of Computer Science,
A.P.S. University, Rewa, India

E-Mail – aarti.tiwari10@gmail.com