ISSN:0975-9646

A Survey on Web Log Mining Pattern Discovery

Amit Vishwakarma,
M.tech scholar, TIT science,Bhopal

Kedar Nath Singh,
Asst. professor,TIT science,Bhopal

Abstract— **web is a great source of information and knowledge, where a numerous of users find their interest. The data available is in form of structured (relational) and text data. Therefore, different kinds of data model can be implementable with web data for pattern discovery. Web mining is a data mining tool where the web related data is evaluated for pattern discovery and user navigation pattern. Additionally, according to the nature of data, the kind of mining is also changed. The given paper includes the study about web mining tools and techniques, for finding informative pattern from the web access data. In this paper, a survey on different kinds of web usage mining techniques with their basic models and concepts is provided. In addition of that, for discovering the hidden pattern from web access log files a new model based on visual clustering is also suggested.**

Keywords— **web mining, web usage mining, log analysis, data models, pattern discovery.**

I. INTRODUCTION

The data on web is frequently accessed and changed according to the time. In this context Important and knowledgeable information extraction form the World Wide Web is the main aim of data mining approaches. The application of data mining in web data is known as web mining. Therefore, according to data availability on web, web mining can be categorized into three different manners, Web usage mining, Web content mining and Web structure mining.

Web usage mining: the web usage mining allows finding patterns from Web access information. This usage data provides the paths and user access patterns leading to accessed Web pages. This information is often gathered automatically via the Web servers.

Web content mining: the web content mining is also known as text mining. In content mining applications the scanning and mining of text, pictures and graphs of a Web page is performed. That may help to determine the consequence of the content.

Web structure mining: web structure mining is a tool, which is used to recognize the connection between web pages. This organization of data is discoverable by the condition of web structure schema through database techniques for Web pages. This kind of data analysis allows a search engine to pull data concerning to a search query directly to the connecting Web page from the Web sites.

In this paper, we are working to find appropriate web access log mining techniques. That allows discovering suitable and meaningful patterns from the web access log files. For pattern discovery different kinds of data models are implemented recently. Some of them are working in supervised manner and some of them are works on the basis of unsupervised learning manner. Therefore, firstly required to discuss about the data and data formats that is used to extract the patterns from data.

II. DATA FORMATS

The web access information can be found in different places. Between origin servers to the client end, this access information is organized in different formats that are listed in this section.

1. Proxy Servers: A proxy server is a software system. That is basically implemented by an organization that is connected to the Internet. Therefore, proxy servers are acts as an intermediary between a host and the Internet connectivity. Using this application the concerning organization can ensure security, caching services and administrative control. Proxy servers can also be a valuable source of usage data. A proxy server also manages access logs, in similar format to Web servers, this access log help to record Web page requests and responses from the web servers.

2. Client Side Data: Client side data are composed from the host. That is currently accessing the Websites. To collect information directly from the client end, such as the time that the user is accessing and leaving the Web site, a list of sites visited before and after the current site a client agent may helpful.

Client side data are more reliable than server side data. On the other hand, the use of client side data acquisition methods is also challenging. The main problem is that, the different agents accumulating information. That affects the client's system performance. These processes are introducing additional computational and resources overhead when a user tries to access a Web site.

3. Cookies: In addition to the use the web access log files, a different method often used in the collection of data is the tracking of cookies. Cookies are short strings distributed by the Web server and held by the client's browser for future use. This data is mainly used to track browser visits. By using cookies, Web server can store its own information at the client's machine. Commonly this information is a unique ID that is created by a Web server, by which next time user visits can be realized. Although the maximum size of a cookie cannot be larger than 4 Kbytes therefore it can only store a small amount of information. Additionally, many different cookies may be allocated to a single user. In addition of that, the users may choose to disable the browser option for accepting cookies, Due to privacy and security concerns.

4. Server Log Files: Server side data are collected at the Web servers of a web site. The web server automatically generates the log file when a user request is made from that server. These logs store Web pages information that is accessed by the visitors of the site. Most of the Web servers support as a default option the Common Log File Format, which includes information about the IP address of the client making the request, the host name and user name, the time stamp of request, file name that is requested, and the file size. The Extended Log Format (W3C), which is supported by Web servers such as Apache and Netscape, and the similar W3SVC format, supported by Microsoft Internet Information Server, include additional information such as the address of the referring URL to this page, i.e., the Web page that carried the visitor to the site, the name and version of the browser used and the operating system of host machine.

This section provides a general overview of data formats that can be used to web usage data analysis. The next section introduces the frequently used techniques that are area of research interest.

III. RECENT STUDIES

This section presents the essential contributions on the web access log mining. The recent study demonstrate the methods by which now in these days the pattern discovery is performed using web access logs.

Before proceeding towards our research we will look into a comparative study done by George Siemens et al and Ryan S J.d. Baker [9] wherein both the authors discussed on two e- learning communities LAK(learning Analytics and Knowledge) and EDM (Educational Data mining). They argue that due to growing interest in data and analytics in education, teaching, and learning demands the priority for increased, high-quality research into the models, methods, technologies, and impact of analytics and thus two research communities – Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK) have developed separately to address this need and so there paper argues for increased and formal communication and collaboration between these communities in order to share research, methods, and tools for data mining and analysis in the service of developing both LAK and EDM fields. Further to their discussion they have done with a comparative study of both these communities and cited some subtle differences and similarities in this context. Some of the remarkable ones being

Similarities:-

- EDM and LAK both are based on data-intensive approaches to education.
- LAK and EDM share the goals of improving education by improving assessment, how problems in education are understood, and how interventions are planned and selected.
- Both communities have the goal of improving the quality of analysis of large-scale educational data, to support both basic research and practice in education.
- EDM and LAK requires similar data and researcher skill-sets.

Differences:-

- EDM has a greater focus on automated discovery, and LAK whereas considerably focuses on human judgment.
- EDM models are more often used as the basis of automated adaptation, conducted by a computer system such as an intelligent tutoring system. In contrast to that, LAK models are more often designed to inform and empower instructors and learners.
- It is much more typical in EDM research to see research which reduces phenomena to components and analyzing individual components
- EDM researchers have placed greater focus on issues of model generalizability (e.g. multi-level cross-validation, replication across data sets). By contrast, LAK researchers have placed greater focus on addressing needs of multiple stakeholders with information drawn from data.

At last in the concluding remarks both the Authors are concerned about the rapid development of analytics and data mining tools by commercial organizations that do not build off of either community's expertise, algorithms, and research results that is faced by both EDM and LAK.
However both communities would be facilitated in communicating their vision for data-driven science and practice in the field of education.
At last an open, transparent research environment is vital to driving forward this important work. As connected, but distinct, research disciplines, EDM and LAK can provide a strong voice and force for excellence in research in this area, guiding policy makers, administrators, educators, and curriculum developers, towards the deployment of best practices in the upcoming era of data-driven education.

This research demonstrates the prospective of Web Usage Mining on e-Learning domain. Nawal Sael et al [1] use educational data mining approach to analyse student's behaviour. The student's behaviour analysis helps in learning and to improve the course structure. Author focuses on the pre-processing module, which is considered as the most fundamental phase in the whole mining process. The main objective is to improve a data pre-processing method applied to Moodle logs based on SCORM content structure. They proposed a pre-processing tool to implement new methods. In this research, they define new static variables according to the SCORM content tree and apply more statistics and visualization techniques. In addition, they present multidimensional graphics in order to understand users' accesses patterns. These accumulated variables provide teachers and tutors with interesting knowledge about students' learning process according to different levels of content accessed [1].

To a good reader's understanding we will provide with more references in this context so as to understand the

basics of MOODLE e- learning platform and SCORM content structure.

Cristóbal Romero, Sebastian Ventura, Enrique Garcia in their survey "Data mining in course management systems: Moodle case study and tutorial [10] has clearly defined Moodle as well as the Data Mining techniques for mining not only Moodle logs but further research issues in a user and reader friendly manner according to them Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational context. Their survey is of the specific application of data mining in learning management systems and a case study tutorial with the Moodle system.

The main objective of this paper is to introduce Moodle system in both a theoretical and practical way to all users interested in this new research area. This paper describes the full process for mining e-learning data step by step as well as how to apply the main data mining techniques used, such as statistics, visualization, classification, clustering, association rule mining, pattern mining and text mining of Moodle data. As well as paper makes use free data mining tools so that any user can immediately begin to apply data mining without having to purchase a commercial tool or program a specific personalized tool. This paper is oriented to the specific application of data mining in computer-based and web-based educational systems. It is arranged in the following way: Section 2 describes the general process of applying data mining to e-learning data, especially to Moodle usage information. Section 3 details the preprocessing step necessary for adapting the data to the appropriate format. Section 4 describes the application of the main data mining techniques in e-learning and an example case study with Moodle data. Finally in the concluding remarks and further research are outlined.

Also in another paper titled "The role of new technologies in the learning process: Moodle as a teaching tool in Physics"
By Teresa Martin-Blas and Ana Serrano-Fernandez, In their work an overview of the undergraduate online Physics course has been implemented in the Moodle platform. This course has been developed as an enhancement of the face-to-face courses. The aim of this course is to create an online learning community which helps both teachers and students to have a virtual space where we can share knowledge through different kinds of supervised activities, chats and forums with a positive response of students. [11]

Further to their concluding remarks they cited the implementation of Moodle for physics and their concluding remarks are first of all, Moodle is a great way for teachers to organize, manage and deliver course materials. From the didactic point of view, the usage of multimedia tools to create attractive activities makes the learning process friendlier for students. As a consequence, these activities increase the interest of the students in the study of Physics. Teachers can provide students with a great amount of resources that usually they cannot show in the classroom due to the lack of time. Moodle also makes easier the interaction with the students in real-time and also allows receiving their opinions and suggestions; as a learning community, Moodle makes possible for students to share their knowledge and difficulties, so they can help each other via forums and chats. Teachers can notice in which parts of the subject they have more difficulties to understand the concepts developed in the classroom. At the beginning of the academic year, students were a bit reluctant to participate in this activity, probably because they were not used to face new tasks. Then, they gradually increased their visits to the site. We have noticed that, when we uploaded the lecture notes, they began to explore the other items previously uploaded in the platform; then they started doing the quizzes and they even suggested us some improvements. This is a key point, as it is very important that the students feel involved in their own learning process. We can also note that the number of visits to the platform is increasing over time which suggests that the students have interest in such e-learning techniques. We have evaluated the improvement of the academic results derived from the use of this e-learning platform. The students who used Moodle regularly during the semester have obtained higher scores than the students who did not. So the impact for students of these web based applications becomes apparent. Moreover, the students have transmitted us that their general feeling is that Moodle helps them to reinforce their abilities and knowledge. These results encourage us to continue with the improvement of our Moodle virtual space. Many authors have reported about the use of web based resources in connection to General Physics courses at faculty level. In some of them, the results indicate that there was not a statistically significant difference in the average scores; only the homework performance scores based on assigned homework groups were improved. Overall, the perception of students of web-based homework testing was very positive (Crippen & Earl, 2007). We have in mind to implement another Moodle course (that will be called 'Zero Physics') devoted to improve and homogenize the basic knowledge of the students who are in the freshman year at the University. This course will include some applets involving basic concepts about Mathematics, Physics and experimental techniques. One goal we would like to achieve is to avoid the 'fear' that the students sometimes feel when confronted to the animations for the first time: we have found that they are a bit reluctant to do these kinds of exercises, probably because they are not familiar with these resources. This goal could be achieved by introducing these animations during the first days of the face-to-face course. [11]

On further moving to our discussion on SCORM Félix Buendía García, Antonio Hervás Jorge in their paper "Evaluating E-Learning Platforms through SCORM Specifications" specified that E-learning platforms can be evaluated using multiple criteria and methods but in all these methods there is a lack of benchmark in the evaluation method, and thus in this paper a frame provided that is based on the use of SCORM standard specifications that allow instructors the elaboration of benchmark tests to evaluate e-learning platforms.

The proposed framework is also based on a Learning Platform Evaluation Model that assumes three main areas of functionality of any learning platform: content,

communications, and management. It has been applied to compare the functionalities of two popular LMS that support SCORM specifications. At last in the concluding remarks these works will consider the development of new benchmarks and the application of new specifications such as IMS QTI or IMS Learning Design. [14] For a better understanding of authors research work we will also use the same framework that they have used in their research and also their words to understand the implementations of their work

Their framework figure shows these main components and how they interact with e-learning platforms such as Web based LMS. The basic operation is started by a client browser that requests a LMS server which stores SCORM packages. This LMS access triggers the Runtime service that enables the SCORM package display in the user Web browser (e.g. using a table of contents or a set of navigation buttons). Once the Runtime service is enabled, the user can navigate through the SCORM contents. Additionally, user applications can get or put information on the LMS server through the API adapter function.ADL provides several test applications that allow instructors and developers to check the SCORM functionalities. The purpose of these tests is to verify that a specific SCO (content object) can be launched by an LMS or if it supports the Run-Time Environment

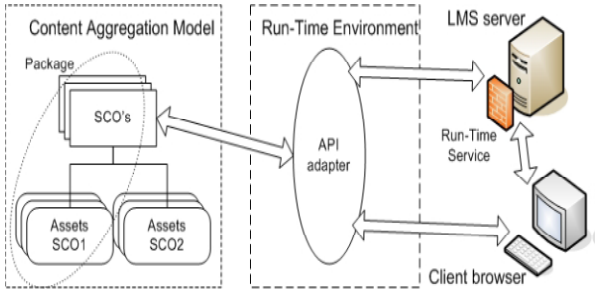

Figure 1, Content Aggregation Model

Application Program Interface (API) functions defined in the SCORM. If the tests are correct the checked LMS can be stated as SCORM-conformant. These test applications are the basis for the elaboration of benchmarks, proposed in the current evaluation framework, which intend to add a pedagogical value in the evaluation of e-learning platforms. [14]

Regarding the benchmark we will provide the study of one more paper to understand this concept clearly. In this paper since now a days there is a big need and demand for big data by academic industry and thereby a large number of systems available in the market that support big data storage and then processing and thus a need arises to evaluate the performance of these systems. In this paper the authors presented Big Bench, an end-to-end big data benchmark proposal. The underlying business model of Big Bench is a product retailer. The proposal covers a data model and synthetic data generator that addresses the variety, velocity and volume aspects of big data systems containing structured, semi-structured and unstructured data.

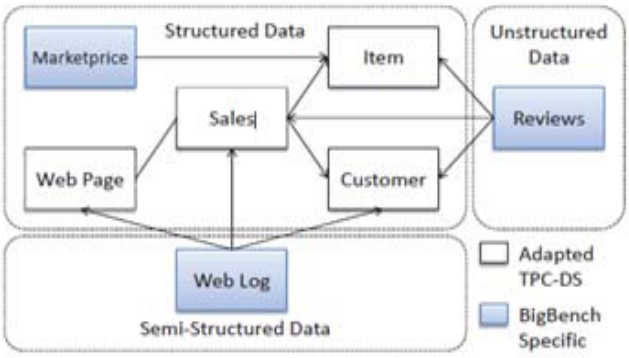

Figure 2 Big Data Benchmark Data Model[16]

The structured part of the Big Bench data model is adopted from the TPC-DS benchmark, which is enriched with semi structured and unstructured data components. The semi structured part captures registered and guest user clicks on the retailer's website. The unstructured data captures product reviews submitted online. The data generator designed for Big Bench provides scalable volumes of raw data based on a scale factor. The Big Bench workload is designed around a set of queries against the data model. [15]

Also while concluding the authors of this paper explained that presented BigBench, a proposal for an end-to-end big data benchmark. The proposal covers a data model addressing the velocity, variety and volume common in big data. Velocity is accomplished by continuous feed into the data store while variety is addressed by including structured, semi-structured and unstructured in the data model. The data model also can scale to large volumes based on as scale factor. We used PDGF as a starting point for our data generator that covers the structured part. PDGF is enhanced to produce the semi-structured and unstructured data. The unstructured component is based on a novel technique we developed leveraging the Markov chain model. The proposal also provides a comprehensive list of workload queries and sets directions for a novel metric that focuses on the different types of processing in big data. Finally, we verified the feasibility and applicability of our proposal by implementing and running it on Teradata Aster DBMS. For future work, we are planning to extend this work in three main areas. First, we would like to enhance the proposal to be a concrete specification that can lead to an industry standard benchmark. This work includes finalizing and detailing the data, workload and metric specifications. We also think system availability during failure should be addressed in the final specification. Second, we think it will be useful to provide a downloadable kit that can be used to setup and run the benchmark. This work includes finalizing the implementation of our data and query generators. Finally, we are planning to extend the benchmark proof of concept to include velocity and multi-user test. We also would like to run the benchmark on one the Hadoop eco-system like HIVE. [15]

Web usage mining is the application of data mining to the data produced by the collaborations of users with web servers. Such kind of data stored in server logs. That characterizes an important source of information. Ida Mele's [2] research emphases on two issues: enhancing search-engine performance by caching of static search results, and to find interesting web pages using recommending news articles. Regarding the caching of search results, they introduce the query covering technique. The basic idea is to occupy the cache with those documents that contribute to the results. For the recommendation of web pages, author presents a graph based method, which controls the user-browsing patterns.

Now we would like to provide with research study issues relating to web search engines so as to understand log mining. Daxin Jiang Jian Pei Hang Li in their paper "Mining Search and Browse Logs for Web Search: Survey" discussed about the two important aspects of web logs namely search logs and browse logs and also about major tasks, fundamental principles, and state-of-the-art methods.

While concluding this paper the authors explained In this paper, we presented a survey on search and browse log mining for web search, with the focus on improving the effectiveness of web search by query understanding, document understanding, document ranking, user understanding, and monitoring and feedbacks. As reviewed, many advanced techniques were developed. Those techniques were applied to huge amounts of search and browse log data available at web search engines, and were powerful in enhancing the quality of the search engines.

There are still many challenging and interesting problems for future work. We list three of them here as examples. First, it is challenging to deal with the long tail in search and browsing log effectively. Search and browse log data are user behavior data and follow the power law distributions in many aspects.

Usually it is easy to mine useful knowledge from the head part of a power law distribution (for example, [Spink et al. 2002]). How to propagate the mined knowledge from the head part to the tail part is still a challenge for most of the log mining tasks.

Second, it is important to leverage other information or knowledge in mining. Log mining mainly focuses on the use of log data. It would be helpful to leverage information or knowledge in other data sources during the mining process, such as Wikipedia. It is necessary to conduct more research on log mining in such a setting.

Last, privacy preserving log mining remains a grand challenge. In 2006, AOL released a search log dataset, but unfortunately a privacy issue arose in the data release. The identity of a user can be detected from the data, although certain data processing had been done in advance [Barbaro and Zeller Jr 2006]. How to preserve privacy in log data and in the meantime do not sacrifice the utility of the log data is a critical research issue. [12]

Akshay Kansara et al [3] offer a study by which the user navigation pattern can be obtained. With the massive quantity of information available on internet makes it exciting for relevant information delivery to the users in an efficient and personalized manner. A method to handle this issue is to use a recommendation approach. Web usage mining is a process of discovering information of user access pattern. This paper offers a hybrid technique using classification and clustering to predict user next step.

While concluding their work they have presented a usage navigation pattern prediction system. The system consists of four stages. The main objective of the proposed system is to predict user navigation patterns using knowledge from (i) a Classification process that identifies potential users from web log data and (ii) a clustering process that groups potential users with similar interest and (iii) Using the results of classification and clustering, predict future user requests. The result was then segmented to identify potential users. From the potential user, a clustering algorithm was used to discover the navigation pattern. The experimental results prove that the proposed amalgamation of techniques is efficient both in terms of clustering and classification. In future, the proposed work will be compared with existing systems to analyze its performance efficient. Plans in the direction of using association rules for prediction engine are also under consideration. [3]

In another paper titled "Event Cube: Multi-Dimensional Search and Mining of Structured and Text Data" by Fangbo Tao, Kin Hou Le, Jiawei Ha, ChengXiang Zha ,Xiao Chen, Marina Danilevsk, Nihit Desa, Bolin Din, Jing G, Heng Ji, Rucha Kanade, Anne Ka, Qi Li, Yanen Li, Cindy Xide Lin, Jialiu liu, Nikunj Oza, Ashok Srivastava, Rod Tjoelker, Chi Wang, Duo Zhang, Bo Zha a survey is presented in which a newer techniques of Event cube is introduced wherein the introductory abstract is about the classification of web data which is in the form of structured and text data and such type of data is usually interrelated. Further Event Cube is a project that provides such a general platform that can easily import any collection of free text and structured data, such as news data, aviation reports or academic papers, extract entities, construct the text-rich data cube and support powerful search and mining functions. And a demo of project has also been shown with datasets. [13]

The web has played a dynamic role to discover the information and to organize them. Now in these days numbers of web sites are increased in the same way the size of web log files also increases. Ramesh Rajamanickam et al [4] accept the challenge to reduce the log files size and classify them more accurately, In order to recognize and utilize them. The main aim is to overcome the deficiency of noisy data to web mining. The author proposed a path extraction technique by finding Euclidean Distance with a sequential pattern analysis algorithm. First, they create Relational Information System using original data sets. Then, cluster the data using Sequential Pattern Clustering Method which is use to produce Core of Information System. So it can get the same effect as original data sets, then after they construct classification model using Core data. The Sequential pattern analysis helps for Path Extraction. The experimental results demonstrate high efficiency and avoidance of abundant data in the given steps of data treatment [4].

Web has been much supportive platform for information and knowledge discovery. Data is stored in web servers and their access information on log files. Web usage mining is the process of mining useful knowledge from web server logs. Web usage analysis requires data abstraction for pattern discovery. That can be achieved by data pre-processing. This paper introduces different formats of web server log files and pre-processes steps for web usage analysis [5].

A web site is the most direct link a company has to its current and potential customers. The companies can study visitor's activities through web analysis, and find the patterns in the visitor's behaviour. These rich results yielded by web analysis, when coupled with company data warehouses, offer great opportunities for the near future. Prime concern of the proposed research is to analyse, and design an efficient and highly secured mechanism for web log mining. The proposed technique is concentrating and supporting to the basic log mining principal during analysis and design of the whole process. Expected results are showing the superiority of the proposed technique [6].

This section includes the study of recent research trends on web usage mining. The next section provides the techniques which are frequently consumed during web log analysis.

IV. PATTERN DISCOVERY

Due to analysis of the recent research on web log mining and pattern discovery techniques, there are various algorithm based methods are available. Most of them are based on decision trees, fuzzy logic, sequential pattern mining, and frequent item set discovery. In addition of that most the author suggests the advance pre-processing techniques to enhance the web log analysis.

According to the recent trend of web log analysis methodology, data mining methods are more reliable and efficient for hidden pattern discovery. But most of the authors are works on the supervised methods. There are not any unsupervised method is available for finding the knowledgeable pattern extraction form web logs.

In this paper we propose an enhanced visual clustering scheme for finding the appropriate patterns form the data. The visual clustering method involves the data transformation, feature vector calculations for the hidden pattern discovery for the web log analysis.

But before going into the further details it's mandatory that the reader should have a brief about the above discussed Data mining techniques for pattern discovery.

Meghana Deshmukh, and Prof. S. P. Akarte in their survey paper "Predictive Data Mining: A Generalized Approach" discussed about framework of data mining, that should fulfil the general data mining tasks. It should elegantly handle different types of data, different data mining tasks, and different types of patterns/models.
They also discussed data mining languages and how these data mining languages support design and implementation of data mining algorithms, as well as their composition into nontrivial multi step knowledge discovery scenarios relevant for practical application.
Their survey proceeds by laying out the basic concepts, starting with (structured) data and generalizations (e.g., patterns and models) and continuing with data mining tasks and basic components of data mining algorithms (i.e., refinement operators, distances, features and kernels).
They have also discussed how to use these concepts to formulate constraint-based data mining tasks and design generic data mining algorithms.
Finally their survey discussed about these components would fit in the overall framework and in particular into a language for data mining and knowledge discovery. [16]

In an another survey by T. Nalini, G. Sangeetha namely "A Survey of Information Retrieval in Web Mining A Survey of Information Retrieval in Web Mining" a new algorithm "A Fuzzy Self, Constructing Algorithm" is introduced for clustering feedback sessions wherein they discussed that fuzzy logic is used for clustering most of the data sets. in their the proposed algorithm the computation time required to partition the dataset will be significantly reduced . It will reduce the original data set into simplified data set. It simplifies the data set and find relevant documents based on user feedback sessions. This will automatically iterate every time and reduce the number of iterations while speeding up the calculations and improve the run time performance. [17]

As a concluding remark all of the above mentioned techniques deal with information retrieval process in web. The proposed approach enhances the efficiency of few of the techniques discussed above. A new approach called "A fuzzy self constructing algorithm" which is used for clustering the user feedbacks, based on ranking was introduced. The user feedbacks are converted into pseudo documents. After clustering, each cluster can be considered as a user search goal. This algorithm improves the speed of the calculations and reduces the computation time to enhance the efficiency. This will automatically iterate every time and reduces the running time

In another paper titled "A New Web Usage Mining Approach for Website Recommendations Using Concept Hierarchy and Website Graph", the authors added that to have a clear and well organized website have become one of the primary objectives of enterprises and organizations. Website administrators may want to know how they can attract visitors, which pages are being accessed most/least frequently, which part of website is most/least popular and need enhancement, etc.

Of late, the rapid growth of the use of Internet has made automatic knowledge extraction from server log files a necessity. Analysis of server log data can provide significant and useful information. Information provided can help to find out user intuition. This can improve the effectiveness of the Web sites by adapting the information structure to the users' behavior.

Most of the Web Usage Mining techniques use Server log files as raw data to produce the user navigation patterns. Along with the server access log file, we incorporate Website knowledge (i.e., Concept hierarchy and Website Graph) into the web usage mining phases.

This incorporation can lead to superior patterns. These patterns can be used to provide set of recommendations for the web site which can be deployed by web site administrator for website enhancement. In this paper, we have considered the server log files of the Website www.enggresources.com for overall study and analysis [20].

In their concluding remarks the authors gave a new idea of incorporating available website knowledge for better session construction which would eventually lead to better pattern during pattern discovery. By using concept based approach we can capture the actual intuition of the user which is sole purpose of any mining process. By identifying user's navigation between concepts, we have generated user profiles which will be useful for administrator to predict user behavior for a particular group of users. Recommendation models based only on usage information are inherently incomplete because they neglect domain knowledge. Better predictions can be made by modeling and incorporating context dependent information: concept hierarchy, link structure and conceptual classification allow us to do so. The results are promising and are indicative of the utility of domain knowledge. We have created the concept hierarchy from scratch. As a future work, automating this will increase the applicability of our model to a wider class of websites. Without semantic knowledge, recommender systems cannot recommend different types of complex objects based in their underlying properties and attributes. Nor can these systems possess the ability to automatically explain or reason about the user behaviors or user recommendations. The integration of semantic knowledge in terms of website ontology is, in fact, the primary challenge for the next generation of recommendation systems has been explained [20].

The overall architecture depicted below is the outcome of the authors

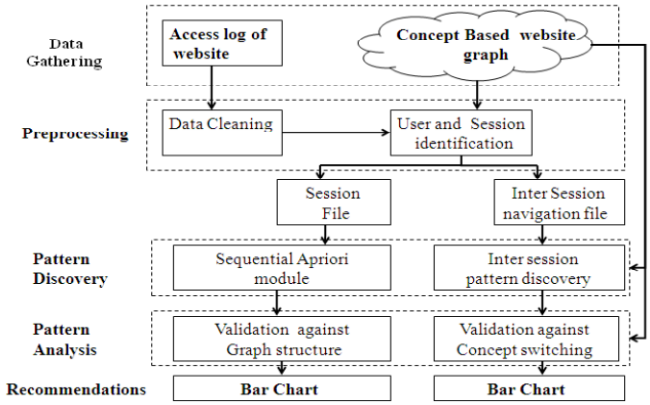

Figure 3 Pattern discovery and analysis[20]

Even more analysis regarding the temporal aspects of pattern discovery has been done by T. Lammarsch, W. Aigner, A. Bertone, S. Miksch, & A. Rind in their paper titled Rind " Mind the Time: Unleashing the Temporal Aspects in Pattern Discovery" they said Temporal Data Mining is a core concept of Knowledge Discovery in Databases handling time-oriented data. State of- the-art methods are capable of preserving the temporal order of events as well as the information in between

The temporal nature of the events themselves, however, can likely be misinterpreted by current algorithms. We present a new definition of the temporal aspects of events and extend related work for pattern finding not only by making use of intervals between events but also by utilizing temporal relations like meets, starts, or during. The result is a new algorithm for Temporal Data Mining that preserves and mines additional time-oriented information. [21]

Also in "Effect of Temporal Relationships in Associative Rule Mining for Web Log Data" by Nazli Mohd Khairudin, Aida Mustapha, and Mohd Hanif Ahmad it has been discussed that with The advent of web-based applications and services has created such diverse and voluminous web log data stored in web servers, proxy servers, client machines, or organizational databases. This paper attempts to investigate the effect of temporal attribute in relational rule mining for web log data. We incorporated the characteristics of time in the rule mining process and analyzed the effect of various temporal parameters. The rules generated from temporal relational rule mining are then compared against the rules generated from the classical rule mining approach such as the Apriori and FP-Growth algorithms. The results showed that by incorporating the temporal attribute via time, the number of rules generated is subsequently smaller but is comparable in terms of quality [22].

V. PROPOSED WORK

In this section the proposed web log analysis method is proposed, where for analysing and finding informative patterns are based on the visual clustering method based. The proposed data model is given using figure 1.

Regarding the basics of visual clustering lets quickly look in the gist of "Visual Clustering in Parallel Coordinates" by Hong Zhou, Xiaoru Yuan, Huamin Qu, Weiwei Cui, Baoquan Chen. [19] The authors argued that Parallel coordinates have been widely applied to visualize high-dimensional and multivariate data, discerning patterns within the data through visual clustering.

However, the effectiveness of this technique on large data is reduced by edge clutter. In this paper, we present a novel framework to reduce edge clutter, consequently improving the effectiveness of visual clustering. The overall visual clustering
is improved by adjusting the shape of the edges while keeping their relative order.

In the conclusion remarks they said that the overall visual clustering is achieved by geometrically bundling lines and forming pattern and further the visualization by varying color and opacity according to the local line density of the curves will be enhanced. [19]

But before going further one should not ignore the research work by Vijesh Mundokalam Nair namely "Visualization Based Sequential Pattern Text Mining" [18] in which he discussed that the mining of sequential patterns is designed to find patterns of discrete events that frequently happen in the same arrangement along a timeline. Like association and clustering, the mining of sequential patterns is among the most popular knowledge discovery

techniques that apply statistical measures to extract useful information from large datasets. So in the light of this author also concluded data mining and visualization techniques for discovery of sequential patterns are two approaches that can compensate each other's weaknesses.

Also powerful visual data mining environment that contains a data-mining engine to discover the patterns and their support values and visualization front-end to show the distribution and locality of the patterns. Our result shows that we can learn more and more quickly in such an integrated visual data-mining environment has been introduced.

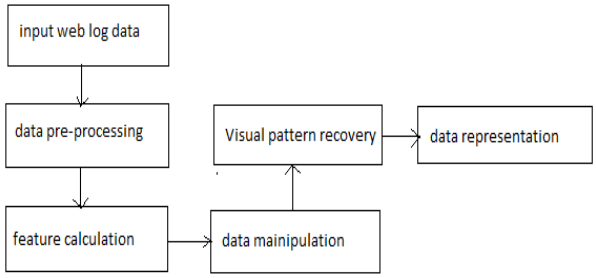

Figure 4 Proposed method [21]

The proposed technique is based on the above method, where web log is produced as input, and for removing the unwanted data pre-processing techniques are used. The pre-processing provides the clean and relational data for analysis. In next step the suitable features are computed and based on these features, in next step the available clean data is manipulated. For example a pre-processed data set can be represented using table 1.

Table 1 pre-processed data

IP	Time	Request/response	Agent
192.168.1.2	10/Oct/2000:13:55:36 -0700	Index.htm	Mozilla/4.08
192.168.1.2	10/Oct/2000:13:55:38 -0700	/apache_pb.gif	Mozilla/4.08

In order to reduce the amount of data similar log entries are removed first using KNN algorithm. After that the data is a new mapping table is constructed by finding unique values in table entries and converted into a numerical symbol. For the above table mapping is given using table 2.

Table 2 mapping table

IP	Time	Request/response	Agent
1	1	1	1
1	2	2	1

The given making table is used to extract features from the exiting data. For extracting features, in this place PCA is an efficient method. Additionally the dimension of data is also reduced. After that using a new visualization technique the data analysis is performed in order to discover meaningful patterns from the data.

In this section the basic concept of the proposed system is provided, in next section part of this section algorithms that are used for web log processing is provided.

A. KNN algorithm

The K-nearest-neighbour (KNN) algorithm measures the distance between a query scenario and a setof scenarios in the data set. We can compute the distance between two scenarios using some distance function d(x,y), where x, y are scenarios composed of features, such that

X={x1, x2, x3…..}

Y={y1,y2,y3…..}

Two distance functions are discussed here:

Absolute distance measuring:

$$\mathbf{d_A(x,y)} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{|x_i - y_i|}$$

Euclidean distance measuring: $\mathbf{d_A(x,y)} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \sqrt{\mathbf{x_i^2 - y_i^2}}$

Because the distance between two scenarios is dependant of the breaks, it is suggested that resulting distances be scaled such that the arithmetic mean across the dataset is 0 and the standard deviation is 1. This can be accomplished by replacing the scalars with according to the following function:

$$\mathbf{x'} = \frac{x - \bar{x}}{\sigma(x)}$$

Where the un-scaled value is the arithmetic mean of feature across the data set, is its standard deviation, and is the resulting scaled value.

The arithmetic mean is defined as:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

We can then compute the standard deviation as follows:

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

KNN can be run in these steps:

1. Store the output values of the M nearest neighbours to query scenario Q in vector
r = {r_1…...r_m} by repeating the following loop M times:
a. Go to the next scenario S_i in the data set, where I is the current iteration within the domain {1……P}
b. If Q is not set or q < d (q, S_i): q ←d (q, S_i), t ←O_i
c. Loop until we reach the end of the data set.
d. Store q into vector c and t into vector r.
2. Calculate the arithmetic mean output across r as follows:

$$\bar{r} = \frac{1}{M}\sum_{i=i}^{M} r_i$$

3. Return r as the output value for the query scenario q

B. *Principal Component Analysis (PCA)*

It is a traditional method of face recognition which is based on the Karhunen-Loeve Transform (KLT), works on dimensionality reduction in face recognition. Turk and Pentland used PCA exclusively for face recognition [7]. PCA computes a set of subspace basis vectors for a database. These basis vectors are representation of an object which is correspond to a class – like structures named Eigen-faces. The projection of data in this compressed subspace allows for easy comparison of data [8].

Mathematically, it can be explained as given below. Assume $(X_1, X_2, X_3, \dots, X_m)$ is a set of M train set from N classes arranged as column vector. Average data can be defined as:

$$\varphi = \frac{1}{M}\sum_{n=1}^{M} X_n$$

Each object differs from the average by vector

$$\emptyset_i = X_i - \varphi$$

When applied to PCA, this large set of vectors seeks a set of M orthogonal vectors U_n, which describes the distribution of data.

The K^{th} vector U_k is chosen such that

$$\vartheta_k = \frac{1}{M}\sum_{n=1}^{M} [U_k^T * \emptyset_n]^2$$

is maximum, applied to

$$U_k{}^T U_k = \delta_{lk} = f(x) = \begin{cases} 1, if\ l = k \\ 0, otherwise \end{cases}$$

The vector U_k and scalar ϑ_k are the eigenvectors and eigenvalues respectively of the covariance matrix

$$C = \frac{1}{M}\sum_{n=1}^{M} \emptyset_n \emptyset_n^T = AA^T$$

Where

$$A = [\emptyset_1, \emptyset_2, \dots, \emptyset_M]$$

In this section of the paper proposed work and the technique is discussed for visual pattern analysis from web log data.

VI. CONCLUSION

This paper, introduces the analysis of different methods of web usage mining. In addition of that a new approach for visual cluster analysis is proposed. The proposed method consumes, PCA and KNN algorithm for data differentiae, mapping and filtering. That helps to understand the user's navigational patterns of web usages. In near future the proposed system is implemented using visual studio environment.

REFERENCES

[1] Nawal Sael, Abdelaziz Marzak, Hicham Behja, "Web Usage Mining Data Preprocessing and Multi Level Analysis on Moodle", 978-1-4799-0792-2/13/$31.00 ©2013 IEEE

[2] Ida Mele, "Web Usage Mining for Enhancing Search-Result Delivery and Helping Users to Find Interesting Web Content",WSDM'13, February 4–8, 2013, Rome, Italy. Copyright 2013 ACM 978-1-4503-1869-3/13/02 ...$15.00.

[3] AkshayKansara, Swati Patel, "Improved Approach to Predict user Future Sessions using Classification and Clustering", International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064, Volume 2 Issue 5, May 2013

[4] Ramesh Rajamanickam and C. Kavitha, "FAST REAL TIME ANALYSIS OF WEB SERVER MASSIVE LOG FILES USING AN IMPROVED WEB MINING ARCHITECTURE", Journal of Computer Science 9 (6): 771-779, 2013 ISSN: 1549-3636©2013SciencePublications doi:10.3844/jcssp.2013.771.779 Published Online 9 (6) 2013

[5] Naga Lakshmi, Raja SekharaRao ,SaiSatyanarayana Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis", International Journal of Innovativ Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-4, March 2013

[6] Gajendra Singh, Priyanka Dixit, "A New Data Mining Technique for Web Log", International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 11, November – 2013

[7] Vapnik N.: The Nature of Statistical Learning Theory, Springer.

[8] Ajeet Singh, BK Singh and Manish Verma, "Comparison of HGPP, PCA, LDA, ICA and SVM", VSRD-IJEECE, Vol. 2 (4), 2012, 179-188

[9] George Siemens, Ryan S J.d. Baker "Learning Analytics and Educational Data Mining: Towards Communication and Collaboration" *Conference'10*, Month 1–2, 2010, City, State, Country. Copyright 2010 ACM 1-58113-000-0/00/0010…$10.00

[10] Cristóbal Romero , Sebastián Ventura, Enrique García "Data mining in course management systems: Moodle case study and tutorial" © 2007 Elsevier Science. All rights reserved.

[11] Teresa Martin-Blas and Ana Serrano-Fernandez "The role of new technologies in the learning process: Moodle as a teaching tool in Physics" _ 2008 Elsevier Ltd. All rights reserved. doi:10.1016/j.compedu.2008.06.005

[12] Daxin Jiang Jian Pei Hang Li "Mining Search and Browse Logs for Web Search: Survey" ACM Transactions on Computational Logic, Vol. V, No. N, April 2013,

[13] Fangbo Tao, Kin Hou Le, Jiawei Ha, ChengXiang Zha ,Xiao Chen, Marina Danilevsk, Nihit Desa, Bolin Din, Jing G, Heng Ji, Rucha Kanade, Anne Ka, Qi Li, Yanen Li, Cindy Xide Lin, Jialiu liu, Nikunj Oza,Ashok Srivastava, Rod Tjoelker, Chi Wang, Duo Zhang, Bo Zha "EventCube: Multi-Dimensional Search and Mining of Structured and Text Data" KDD'13, August 11–14, 2013, Chicago, Illinois, USA. Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...$15.00

[14] Félix Buendía García, Antonio Hervás Jorge "Evaluating E-Learning Platforms through SCORM Specifications" TIN2005-08788-C04-02.Education Ministerial Paper of Spain Educational Ministry.

[15] Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, Hans-Arno Jacobsen "BigBench: Towards an Industry Standard Benchmark for Big Data Analytics" SIGMOD'13, June 22–27, 2013, New York, New York, USA. Copyright 2013 ACM 978-1-4503-2037-5/13/06 ...$15.00.

[16] Meghana Deshmukh, Prof. S. P. Akarte "Predictive Data Mining: A Generalized Approach" Meghana Deshmukh et al, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.1, January- 2014.

[17] T. Nalini, G. Sangeetha "A Survey of Information Retrieval in Web Mining A Survey of Information Retrieval in Web Mining" Middle-East Journal of Scientific Research 19 (8): 1047-1052, 2014

[18] Vijesh Mundokalam Nair namely "Visualization Based Sequential Pattern Text Mining" International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 12, December 2013.

[19] Hong Zhou, Xiaoru Yuan, Huamin Qu, Weiwei Cui, Baoquan Chen "Visual Clustering in Parallel Coordinates" Eurographics/ IEEE-VGTC Symposium on Visualization 2008

[20] T. Vijaya Kumar, H. S. Guruprasad, Bharath Kumar K. M., Irfan Baig, and Kiran Babu S. "A New Web Usage Mining Approach for Website Recommendations Using Concept Hierarchy and Website Graph" Manuscript received April 20, 2013; revised July 2, 2013. DOI: 10.7763/IJCEE.2014.V6.796

[21] T. Lammarsch, W. Aigner, A. Bertone, S. Miksch, & A. Rind " Mind the Time: Unleashing the Temporal Aspects in Pattern Discovery" The Eurographics Association 2013.

[22] Nazli Mohd Khairudin, Aida Mustapha, and Mohd Hanif Ahmad "Effect of Temporal Relationships in Associative Rule Mining for Web Log Data" Hindawi Publishing Corporation Scientific World Journal Volume 2014, Article ID 813983, http://dx.doi.org/10.1155/2014/813983