CHAPTER-1

INTRODUCTION

1.1 OVERVIEW

Now, the use of hypertext linked displays goes on to elaborate at a power order as a medium for carrying business and diffusing data through Cyberspace. Even with developing measures and technology, there are many challenges that must overcome in order to exhaustively examine the utilization of a website. Due to enormous knowledge explosion and rapid growth in Research and development information over the internet is paving the way for World Wide Web (WWW) to become strong and powerful program to store, circulate and recover data as well as excavate utilitarian cognition. Due to the attributes of the vast, various, active and unstructured nature of web information. The research in network information faced many hurdles such as measurability, multimedia and secular consequences etc. and many other information oriented services are achieved. [1].

Invention of a internet site centers around arranging the data on every varlet and the machine-readable links among the varlets in a direction which appears most innate to the site users so as to help their surfing (and also buying). For modest sites and soul network designer’s suspicion along with few straight utilitarian, statistics may be enough for anticipating and asserting the client’s surfing conduct. The existing web intelligent services are mostly utilized to fulfill the necessitates of several
network browsers peculiarly with the quick aid of the semantic network usage Mining. [2]. So, future research is required recognize fresh well-informed approaches and proficiencies for web users. As a result, Web users are always drowning in an ocean of information and facing the problem of overloaded information when interacting with the web. The web related research applications faced the following problems quite often [3].

1.1.1 Finding Relevant Information

We surf the web to find specific information on the web. We usually specify a simple keyword query and the response from a web search engine is a list of pages, rank based on their similarity to the query. But with low precision which is due to irrelevance of many searches may get irrelevant information to our query and low recall which is due to the inability to index all web information. Due to this, some relevant pages are not indexed.

1.1.2 Discovering New Knowledge from the Web

We can term the above problem as a query-triggered process (retrieval oriented). On the other hand, we can have a data-triggered process that presumes that we already have a collection of web data and we want to extract potentially useful knowledge out of it (data mining-oriented).
1.1.3 Individualization of Web Page Content Process

Individualization is the client’s web experience based on many web-dependent systems for instance personalized marketing for e-commerce. Constructing the runtime suggestions to a web client dependent on the profile of the client along with the behavioral usage which is most interested to most of the applications, for instance marketing sales in e-commerce. Web usage mining in an efficient technique to attain the aim as described in the present suggestion system, so that the present suggestions available for data mining may not used. The web watcher and site helper letizia and clustering a proposal by mobasher et al. and yan et. al. have all focused on issuing the web site Individualization depending on the data usageness. The suggested system contains an offline program which does the clustering process and analysis and a online program which carries out the work for the production of the runtime references of the web pages. The browser using a website is given to a unique group depending upon its existing navigation pattern. The references which are provided to a specific browser selected at runtime depending upon the pages which were given to some other browsers what who belong to the same group.

1.1.4 Learning about Individual Users

It is close to experiencing the concerns of clients and their necessitates. In this job, there are sub-tasks such as mass customizing
the data to the meant users or even individualizing it to single client jobs concerned to good internet site design and direction troubles concerned to marketing etc.,

Network Excavation Approaches provide a set of techniques that can be used to solve the above problems. However web mining techniques are not the only tools to handle these troubles. But, network excavation approaches are not just the tools to deal these troubles. Network excavation research integrates from several fields consisting Innate Terminology Treating (ITT), Machine Memorizing, Data Recovery (DR) and Databases (DB).

1.1.5 Evolution of Web Mining

- **Data Collection (1960):** is characterized by its retrospective nature & static data delivery.

- **Data Access (1980):** is characterized by its retrospective nature and dynamic data delivery at record level.

- **Knowledge Discovery Database (1989):** is the non-trivial extraction of implicit, previously unknown and potentially useful information from databases.

- **Data Warehousing and Decision Support Systems (1990):** is characterized by its retrospective nature and dynamic data delivery at multiple levels.

- **Data Mining (2000):** is characterized by its prospective nature and proactive information delivery at any level.
• **Web Mining (2010)**: is characterized by its prospective nature and hyperactive information delivery at any level.

Web Mining in view of Data mining have 3 operations:

1. **Clustering**: finding natural grouping of users, pages etc.

2. **Sequential access pattern Analysis**: The approach of consecutive blue print disclosure efforts to determine inter-session blue print so that the comportment of a group of tokens is pursued by another token in a time-governed group of sessions or installments. By utilizing this method, network marketers can expect next impose blue prints which will be useful in laying adverts directed at particular client groups. Other types of secular analysis that can be done in consecutive blue prints consist of course analysis; alter point spotting, or similarity analysis.

3. **Associations Rule**: which URLs tend to be requested together. Association rule generation can be used to relate pages that are most often referenced together in a single server session.

Web Mining can be broken down into following sub tasks:


2. **Information Selection and Preprocessing**: Automatically selecting and preprocessing specific information from resources retrieved from the web.

3. **Generalization**: Discover general patterns at individual sites as well as across multiple sites.

4. **Analysis**: Validation and/or interpretation of mined patterns.
1.2 WEB USAGE MINING TECHNOLOGY

WUM is the application which uses different data mining techniques to analyze and extract interesting patterns of user’s usage and interests over data on web. The usage data consists of user’s behavior while browsing on web. This activity involves discovering the patterns automatically from one or more network hosts. Systems that utilize this application render and gather huge bulks of information, normally rendered mechanically by network hosts and maintained in host logs. Systems examine this information that serves to find out the customer’s interests, cross marketing plans and promotional campaigning strategies etc.

Web mining is an intelligent analysis of Web data [32]. WUM is the process which extracts “interested” blue prints from the network information. The network information consists of network host access log, gateway host log, browser log, client registration data, and client’s session. In this we mainly use web log as data source. So we use the concept of web log mining instead of WUM. The process of web log mining is as follows:

**Data preprocessing**

Data preprocessing or data preparation is the first stage of web log mining. The raw data is converted into the data with which pattern discovery could deal. It includes data cleaning, user recognition, session recognition, path supplement, transaction recognition and so on. Web
log data preprocessing has a direct impact on the correctness or models and pattern rules which are discovered in the next stage.

**Patterns discovery**

In this stage, using various methods, we attempt to find models and pattern rules of user’s access behavior. Common technologies are sequential patterns, association rules, clustering, and classification and so on.

**Pattern analysis**

In most of the cases, web usage mining can find all the models and rules. Pattern analysis is used to extract valuable interesting patterns from all these models and rules as shown in Fig: 1.1.

**Preprocessing**

The main intention of the preprocessing process is to preprocess the literal network logs to find out finish network access sessions. When utilizing the network host logs, each client’s access tasks and works carried out by the client of a website are remarked down by the network host of the website. Each client access information includes the client internet protocol address, petition time, needed Uniform Resource Locator, Hyper Text Transfer Protocol status code, etc. Clients are considered as a whole group as the internet protocol addresses are not related to each client-recognizable visibility information. Normally, network logs may be regarded as a group of consecutive of access tokens from one user or phase in a time period increasing order. Preprocessing
process [33] includes the methods like cleaning the data, client recognition and phase recognition. These methods are used to the actual web log files to acquire complete web access sessions. Data cleaning is considered as website specific process which includes necessary tasks like combining the logs from several servers and making chunks of the logs into data items. But the graphics file requests are removed from the log files after preprocessing.

**Data cleaning**

Data cleaning is a site specific step that involves mundane tasks such as merging logs from multiple servers and parsing the log into data fields. Typically graphics file requests are stripped out at this stage. This is easily done by checking for file names suffixes such as "GIF" or "JPG". Graphics files can be left in the data set and rolled up into page views in a later preprocessing step without any loss of generality or request for any other file which may be admitted into a network page: or even seafaring session performed by robots and network spiders. While petition for graphical subjects and files are gentle to eradicate robot and network spiders seafaring blue prints must be blue print must be explicitly. This is normally exercised for example by citing to the distant host, by citing to the agent, or by assuring in the access to the rpbpts .txt file. HTTP status ciphers are utilized to represent the win or lose of the called for issue. The records with ciphers among 200 and 299 are
regarded as productive records, and remaining are removed from the networks logs.

**Client and session recognition**

For examining client access conduct, unparalleled clients must be recognized. As mentioned before, clients are regarded as anonymous in most network hosts. We can alter the user recognition procedure to client Internet Protocol recognition. In a different discussion, petitions from the same Internet Protocol address can be considered as from the same client and kept in the same cluster under that client. To recognize clients more correctly, some other data from the network logs may be useful. The agentive role registered in network logs captures data on the user browser on Formal Based Concept Analysis (FBCA) [34&35]. Then application of our suggested WUL-excavation algorithm to excavate is the most potential and utilitarian group of affiliation access blue prints from the Network Utilization Lattice. The vantage of the suggested WUL-grounded method is that it can produce much less number of affiliation access blue print principles without compromising much on lineament for network individualization applications when equated with the Apriori-grounded algorithms [36]. As such, the decreased group of non surplus WUL-grounded affiliation access blue prints can greatly better the efficiency on producing good principles for individualized online recommendations as shown in fig: 1.2
Web Usage Mining (WUM) methods have been applied to many real-time practical applications [30] including the followings:

**Personalization**

[37&38] Web utilization excavation approaches can be utilized to supply individualized network browser experience. For example it is possible to predict the browser conduct in literal time by equating the present seafaring blue print with distinctive blue prints that were evoked from previous network log. In this field, suggestion systems require the most usual application; its main objective suggest concerning links to products that could be concerning to many of the clients. Individualized Site Maps are an instance of suggestion system for links suggested an adaptative method rearrange the product catalog consorting to the estimated client visibility. A method to incorporate field ontology’s into the individualization procedure grounded on web utilization excavation is suggested in admitting an algorithm to build knowledge base level mass visibilities from a accumulation.

**System Improvement**

[39&40], rate of quality and another measurable quality property are very essential to client gratification from services like databases, networks and many more and same kind of the qualities are predicted from the clients of web services. Web usage mining gives the idea to understand the web load behavior that may be useful to construct rules for web caching
network transmission, load balancing or data distribution. The actual concern may the provision for security for web based services particularly as e-commerce continues to grow at an exponential rate. Web usage mining is also useful patterns that are helpful to detect intrusion, crimes, attempted break-ins, etc., Some models are suggested to predict the truthness, to the two temporal as well spatial, in the web pages which are needed from a specific client or a cluster of clients who access from the similar proxy server. The parameters are also dependent on the location of the server which are required to decide perfecting and catching approaches for the proxy server. But the when using the more of the ever changing content has decreased the advantages of storing at the client level and server level.

**Site Modification**

[41&42], The measurable quality of the a web site which includes the attracting power, in the context of both content and skeleton is very essential to most of the applications, for instance, a catalog for the products catalog fore-commerce. WUM renders elaborated resubmit on client conduct, supplying the network site architect with data on which to ground reproject conclusions. Network utilization information supplies a chance to become each site into an ongoing useableness trail. As their data is not as accomplished as the data that can be collected from a formal useableness analysis with videos and civilized perciipients,
Network utilization information are inexpensive and ample. Access times and measures of lostness can be calculated automatically instead of manually. When the outcomes of any of the proposals could lead to re-planning the structure and subject of a site concentrates on mechanically modifying the structure of a site grounded on utilization blue prints disclosed host logs. Constellating varlets is applied to find out which varlets should instantly relate.

**E-Business Intelligence**

[43, 44&45] Information on how customers are using a web site is critical for marketers of ecommerce businesses. Buchner et. Al. have presented a knowledge process in order to discover marketing intelligence from web data. They define a web log data hypercube that consolidates web usage data along with marketing data for e-commerce applications. Four distinct steps are identified in customer relationship life cycle that can supported by the their knowledge discovery techniques: customer attraction, customer rention, crosssales and customer departure. Padmanabham et.al use web server logs to generate beliefs about the access pattern of web pages at given web site. Algorithms for finding interesting rules based on the unexpectedness of the rule are also developed.
1.3 OBJECTIVES OF THE THESIS

The main objective of the present research is to determine several new web usage mining approaches which are useful for extracting the knowledge which is hidden in the web usage logs. To be more specific the research work focuses to mine the web host logs and the client side logs. The evoked cognition is used by few real-time advanced network diligences such as Web Testimonial, network individuation and data excerption and recuperation from Network utilization information. To accomplish this accusative, investigation is carried out as mentioned:

- Introduce new approaches to extract knowledge from web logs. To be more specific the primary intention is to develop new approaches and investigation is carried out on data mining approaches for successive chain access pattern mining and related access patterns using the web logs. The main aim is to suggest effective and beneficial excavation algorithms to find out access blue prints utilizing the network utilization logs. However in addition to the general successive chain patterns and conventional Apriori-based association rules are developed.

- A special focus is made on mining some special access patterns such as Recurrent Successive Chain Access Patterns (RSAP) and FBCA-based association access patterns (FBCA) (i.e., using the
Formal Based Concept Analysis approach the access patterns are mined).

- For real-time advanced web application required approaches are to be investigated to implement the extracted knowledge. The web applications may also used for evaluating the quality and usefulness of the extracted access patterns. The new web applications to provide effective and potential real-time web services will be developed. Especially, a new web applications like web suggestion and web browser individualization and profiling are developed and are focused.

- With the introduction of modern approaches for semantic web usage mining. Investigation is made on few effective modern approaches to extract the semantics from web logs and based on the Semantic Web, the web usage data is mined. More specifically, emphasis is laid on learning Ontology and discovery.

The proposed research framework is shown below in Fig: 1.3

In this research, following investigations have been made.

**Consecutive access Blue Prints Excavation** – The orthodox Apriori-based consecutive blue prints excavation algorithms find with the similar problems that many of the other Apriori-based algorithms which need large number of scans over the database in order to mention which of the candidates are really recurrent clients. The Network Access Blue Print -
tree reckons on successive chain pattern mining algorithms [48&49] may overcome the problems faced by the many other conventional approaches. A potential Web Access Blue Print-tree grounded excavation algorithm is called as CCEA (Contingent Consecutive Excavation Algorithm) is developed.

In CSMA, deletion is necessary to re-build the in between conditional Web Access Pattern-trees to enhance the potential of the mining method. Along with this, to enhance more the potential of the CSM Algorithm, the research work has been extended and Web Logs are directly mined without building a Web Access Pattern-tree. The new approach is known as Temporal Conditional Sequence Mining Algorithm (TCSMA) [50].

In this Thesis, the discussion is not made on the CSMA. Instead, I have only discussed the TCSMA which is an improved version of CSMA (which is a part of the suggested TCSMA).

**Association Access patterns Mining** – All most all conventional Apriori-based association rule mining algorithms provides a support-confidence structure which suffers from the similar type of anomalies in which huge rules are returned generally. To the fact, the obtained association rules have the redundant association rules. The threshold value is very much helpful for navigating and thus discarding the uninterested rules. This threshold value is used in the place of minimum support and confidence. Even after using the threshold value many rules
will be resulted as uninteresting. The Formal Based Concept Analysis [51] is used here as a procedure for excavating affiliation access blue prints. The suggested approach [52] can decrease the number of principles excavated without any loss in character.

- **Network testimonial diligence** – A significant network testimonial system called as Consecutive Network Access-grounded Testimonial System (SWABRS) is introduced which is depending on successive chain access patterns. A new suggestion system called as Affiliation Network Access-grounded Testimonial System (ANAGTS) is also introduced which is Association Access Patterns oriented. AWABRS is primarily useful for serving the purpose to increase intensity of the character of the evoked FBCA -grounded affiliation access blue print and orthodox affiliation access blue prints.

1.4 **ORGANIZATION OF THE DISSERTATION**

The present Chapter initially focuses on the introduction of the underlying concepts of the web mining, web usage mining and Semantic Web Usage Mining. The detailed discussion is carried out on the objectives of the research. The remaining part of the dissertation is arranged as stated below:

Chapter 2 introduces the similar works and the background of Web Usage Mining. The chapter begins with the introduction of the web usage data. A detailed discussion is carried out on the data preprocessing
techniques for web usage data and also several approaches for the pattern extraction are discussed.

Chapter 3 introduces the suggested approach for successive chain access pattern Mining. The suggested technique is called as TCSMA (Temporal Conditional Successive Chain Mining Algorithm). The Common and Recurrent Successive Chain Access Patterns are mined using the TCSMA. A deep discussion is carried out on the performance outcomes of the TCSMA.

In Chapter 4 focus is shifted to the usage of an better data analysis method, also called as Formal Based Concept Analysis for mining association access patterns from web logs. The performances of the suggested approaches are evaluated by comparing the results of the suggested approach with the results of the conventional Apriori-based association rule mining.

Chapter 5 concentrates on the study of exploring the SWABSS method (Successive Chain Web Access-based Suggestion System) and the AWABSS (Association Web Access-based Suggestion System) for web suggestion. A detailed discussion is also carried out on the performance evaluation of the suggested approaches and the results of the suggested approaches are compared with the results of the traditional approaches.

At last, Chapter 6 summarizes the research work. Also the suggestions are given to further carry out the research works in this domain including client-side data mining for user profiling using the
client-side logs, web usage data recovery and Semantic web usage mining.