LITERATURE REVIEW

2.1 Preprocessing Methods
2.2 Pattern Discovery Methods
2.3 Pattern Analysis - Ranking Methods
2.4 Pattern Analysis - Clustering Methods
2.5 Chapter Summary
LITERATURE REVIEW

This chapter surveys the state of the art techniques which have been reviewed to develop the overall framework of this research work. Web usage mining involves three phases, namely, preprocessing, pattern discovery and pattern analysis. There are different techniques available for web usage mining with its own advantages and disadvantages. This chapter provides discussions about several web usage mining techniques that are available today.

2.1 PREPROCESSING METHODS

In preprocessing phase, user identification is an important issue that deals with how exactly the users have to be distinguished. Identifying users based on web log files is not a straightforward problem, thus various methods have been developed. To overcome the difficulties of client side caching and shared IP addresses, Renáta & Sándor (2007) have introduced three different methods for web user identification. Two of them are the most commonly used methods that are based on IP addresses and cookies respectively in web log mining systems, whereas the third one is a novel approach that uses complex cookie based method to identify web users. Furthermore, they also take steps towards identifying the individuals behind the impersonal web users. To demonstrate the efficiency of the new method, Web Activity Tracking (WAT) system that aims at a more precise distinction of web users based on log data, was implemented. The WAT produced statistical analysis on real data about the behavior of the Hungarian web users and comparisons of the three methods were presented. The system hold for creating different statistics about the Hungarian web usage alone and when applied for the different web user identification methods, it becomes an unsolvable issue.

Many heuristics are used for better identification of the users in many related works (Spilipoulou et al., 2003) and the methods are grouped into two classes, namely the proactive methods and the reactive methods. The former differentiates the users before or during the page request, while the latter relates the users with the web log entries after the web log is written. In proactive strategies, browsers invoked simple user authentication with forms, using cookies or using dynamic web pages associated with
them. Reactive strategies work only with the recorded log files, and the user navigational patterns helps to distinguish different users, download timing sequence or some other heuristics based on some assumption regarding their behavior.

A study to assess heuristics for session identification from web log data has been presented by Spilipoulou et al. (2003). To evaluate the performance of heuristics, a session identification method was employed to reconstruct sessions from the server log data. Such heuristics are called to partition activities first by user and then by visit of the user in the website, where user identification mechanisms, such as cookies, may or may not be available. A set of performance measures was proposed that are sensitive to two types of reconstruction errors and which are appropriate for different applications in Knowledge Discovery in Databases (KDD). The framework was tested on the web server data of a frame based website. The first experiment concerned with a specific KDD application and has shown the sensitivity of the heuristics to particularities of the site’s structure and traffic. The second experiment is not bound to a specific application but rather compares the performance of the heuristics for different measures and thus for different application types. The performance results for each measure can serve as an aid to the analyst for the appropriateness of the heuristics for each type of application. One subject of future work is the quantification of the impact of dynamic web log session analysis in KDD applications.

Dell et al. (2008) presented a novel approach for sessionization based on an integer program. The results of their approach are compared with the timeout heuristic on web logs from an academic website. Results revealed that the integer program provides sessions that better match an expected empirical distribution with about half of the standard error of the heuristic.

Raiyani (2012) introduced Distinct User Identification (DUI) technique based on IP address, agent and session time, and referred pages on desired session time. In addition, cleaning steps such as removal of maintenance redundant pages and grouping of sessions with similar session lengths are presented. The preprocessed results are used in counter terrorism, fraud detection and detection of unusual access of secure data, as well as through detection of regular user access behavior to improve the overall designing and performance of upcoming access.
Tyagi et al. (2010) introduced two preprocessing algorithms, one for data cleaning and the other for data reduction. In the data cleaning algorithm, the records with extension .JPG, GIF, .CSS are removed, but records with irrelevant status code are not removed. So, the status code was removed in the improved data reduction algorithm, which identified the sessions and removed the incomplete session entries. After preprocessing, the data is converted into structured form and then algorithms are applied for mining the information from it.

In the recent work on web log session identification, a discussion on an application of a new session identification method based on statistical language modeling to database trace logs is presented. Several problems like selection of values for the parameters of the language model, evaluating the accuracy of the session identification result and learning a language model without well labeled training data are revealed. All of these issues are important in the successful application of the language modeling based method for session identification.

Huang et al. (2006) presented a new method for determining an entropy threshold and the order of the language model. New performance measures are presented to better evaluate the accuracy of the identified sessions. Furthermore, three types of learning methods, namely, learning from labeled data, learning from semi-labeled data and learning from unlabeled data, are introduced to learn language models from different types of training data. Final report of experimental results shows how effectively the language model based method is used for identifying sessions from the OnLine Transaction Processing (OLTP) database application and the TPC-C Benchmark trace logs. The main drawback of statistical n-gram model is its inapplicability for dynamic query type.

Murata et al. (2006) proposed a method for examining the searching behavior of the web users and extracting the valuable information based on the user interest, which becomes an important task in web usage mining. In order to perform this task, the continuous monitoring of the web user search behaviors is represented in the form of graph. The most visited websites is represented in the form of graph along with user given keyword, so it is named as site-keyword graph. The site-keyword graph identifies the user’s interest for each keyword in their websites.
Agosti and Di Nunzio (2007) development is based on database management methods which manages and maintains the necessary data. The application enables the separation of different entities recorded, demands querying of the log data and facilitates data mining. Session reconstruction is used in order to map the list of activities performed by every single user who visits the site. Every user is identified with the pair, IP address and user agent and only a fixed gap of time are permitted between two successive requests. To identify users and sessions, authentication is recommended, so that it would allow web servers to identify users, create profiles to tailor specific needs, and more importantly to track their requests.

Ciobanu and Dinucă (2008) introduced a methodology based on the average time spent by the visitors on the website pages to identify web user sessions. Results showed that it is necessary to take into account the time spent by users on a web page to identify sessions with high accuracy. Two session identification algorithms are implemented in Java programming language, using NetBeans IDE. The first algorithm uses a fixed value of 30 minutes (1800 seconds) to indicate the end of a session and the second uses the average time spent on the pages of the website by users.

Arumugam and Suguna (2009) presented new techniques to identify user session boundaries by considering IP address, browsing agent, inter-session and intra-session timeouts, immediate link analysis between referred pages and backward reference analysis without searching the whole tree representing the server pages. A complete set of user session sequences and the learning graph, based on which user session sequences is generated. Using this graph predictive prefetching is done. Comparison on the performance of the given approach with the existing reference length method and maximal reference method was done. The analysis with different server's logs shows that proposed approach provides better results in terms of time complexity and precision to identify user session boundaries and also used to generate all the relevant user session sequences.

Xinhua and Oqiong (2011) in the preprocessing methods, the dynamic timeout method is used to investigate conventional session identification methods. Initially, for each web log user, based on the searched web page with their degree value, session timeout is computed. Based on the computed session timeout, user sessions are identified through dynamically adjusting the session timeout. The results revealed that the proposed algorithm can obtain better performance on session identification.
Chapter - 2

Fang, Y. and Huang, Z., (2010) proposed a method considering website structure and its content, reached page access time threshold by collecting page access time, which could be used to divide sessions into various sets. Then, the session sets will be optimized further with the help of session reconstruction, namely union and rupture. It has been proved through experiment that the session set which is attained by the above method is more faithful.

Peng and Zhao (2010) proposed an improved algorithm based on average time threshold value, by calculating the average intervals dynamically among request records in the session, and adjusting the time threshold value individually. Performance was evaluated by comparing with the traditional algorithm that defines a uniform threshold value for all users' web pages. It is observed that the algorithm identified the long session more accurately. Finally, the algorithm reidentified the generated sets of candidate session, which makes the identified session more reasonable and effective. Experimental results shows that the quality of session identification is improved when compared to other methods.

Kapusta and Munk (2012) assumed that the user navigates several pages during the visit until they find the content page with required information. The content page is a page where the user spends considerably more time in comparison with navigation pages. The content page is considered as the end of the session. The division of pages into content and navigation pages is based on the calculation of cut-off time C. When the cut-off time C is known, the session can be created in such a manner that we compare the time of particular web page visit with the cut-off time C. The session is then defined as a path, from the navigation page to the content page (the user spends more time than C), they claimed the content page to be the last page of the session. The cut-off time C is calculated on the basis of exponential distribution of variable RLength (Time spent by user on individual page), here the assumption is that the variance of the time spent on the auxiliary pages is small than the content page. The proposed time threshold method is compared with rule based session identification for web usage mining.

Dinuca and Ciobanu (2012) developed a modified algorithm, in which conditions can be chosen to determine with increased precision i.e., the separation between sessions. In modified algorithm, the number of sessions is greater when compared with classic algorithm, and the average visiting time depends on the page. So, it is used to
separate session’s that better maps the reality than using a single constant value for the algorithm implementation. Also, for sites running in different areas, the usage of the mean time is recommended because it depends directly on the site structure and content pages. The modified algorithm has the same running time as a classical algorithm which is another reason that recommends its use, so its complexity is not modified which is focused as future work.

Kapusta et al. (2014) focused on finding the suitable value of the time threshold, which is then used in the method of user session identification based on the time. To determine its value, the length variable is used to represent the time a user spent on a particular website. Two values of time threshold are compared with experimental methods of user session identification based on the structure of the web namely Reference Length and H-ref. On comparison of all four methods, it was proved that the use of the time threshold calculated from the quartile range is the most suitable method for session identification in web usage mining.

Sengottuvelan et al. (2015) introduced the basic procedure of data preprocessing and the traditional session identification algorithms has been fully analyzed, based on which a session identification algorithm on page threshold and dynamic timeout are presented. To conclude, the initial timeout is computed for each page according to sessions formed, combining with the degree of importance, improved dynamic threshold algorithm which discards the uninterested attributes from the log file. Experiment results shows that the proposed algorithm can obtain a better performance on session identification and user interests which is the key for web personalization.

The findings of the literature survey on preprocessing methods are summarized in Table 2.1.

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Technique</th>
<th>Observations</th>
</tr>
</thead>
</table>
| 2003 | Spiliopoulou et al | Timeout Heuristic Metrics | • Proactive and reactive strategies used for better user identification.  
• Performance evaluation of heuristics quantifies the impact of session identification. |
<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Technique</th>
<th>Observations</th>
</tr>
</thead>
</table>
| 2008 | Dell et al | Sessionization Based On An Integer Program | - Provides sessions that better match an expected empirical distribution.  
- Minimizes Error when compared to standard heuristic without seeking maximum number of specific sessions. |
| 2012 | Raiyani | Distinct User Identification(Dui) | - Detection of regular access behavior of users improves the overall performance of upcoming access of preprocessing results.  
- Sessions with similar session lengths are grouped and applicable for dynamic session identification. |
| 2010 | Tyagi et al | Data Cleaning and Data Reduction Algorithm | - Data cleaned without removal of irrelevant status code.  
- Improved reduction algorithm identified the sessions and removed the incomplete session entries but dynamic sessions not addressed. |
- Major issue is its inapplicability for dynamic query type. |
| 2007 | Agosti and Di Nunzio | Database Management Methods | - User and session identification done through authentication.  
- Session reconstruction is carried out to track user activities. |
| 2009 | Arumugam and Suguna | Session Boundary Identification Using Graph Predictive Prefetching | - Provides better complexity and precision results time to identify user session boundaries  
- Noisy samples remain a hurdle. |
| 2011 | Xinhua and Qiong | Dynamic Timeout Method | - User sessions are identified by dynamically adjusting the session timeout.  
- Obtained a better performance. |
<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Technique</th>
<th>Observations</th>
</tr>
</thead>
</table>
| 2010 | Peng and Zhao | An Improved Algorithm Based On Average Time Threshold | • Quality of session identification is improved by adjusting the time threshold value.  
• Capable of identifying long sessions accurately. |
| 2012 | Dinuca and Ciobanu | Modified Session Identification | • Uses single constant value to separate sessions. Mean time is used for sites running in different areas.  
• Improvement in complexity is directed as future work. |
| 2015 | Sengottuvelan et al | Session Identification Based on Page Threshold and Dynamic Timeout | • Discards the uninterested attributes from log file.  
• Obtain a better performance on session identification.  
• Addresses only static queries. |
|      | Summary of Findings From Literature Review | | • Session identification is done based on the intra and inter-time, but web log users’ completion time, which is not identified.  
• The n-gram model issue is that it assumes the entire database query to be static, so dynamic query type is not applicable. |

### 2.2 PATTERN DISCOVERY METHODS

In pattern discovery phase, the Apriori algorithm uses prior knowledge of frequent patterns to mine frequent itemset. It states all nonempty subsets of a frequent pattern must also be frequent. Although various interestingness measures and rule pruning methods have been applied to association rule mining of web usage data, extracting useful information from the set of generated association rules still remains a difficult task. A survey on different methods and algorithms used to find frequent patterns are discussed.

Park et al. (1997) developed an effective algorithm for the candidate set generation, which is hash based and is especially useful for the generation of a candidate set for large 2-itemsets. Explicitly, the number of candidate 2-itemsets generated by the proposed algorithm, in orders of magnitude, is smaller when compared with previous
methods, thus resolving the performance bottleneck. Transaction database size can be effectively trimmed at an early stage through generation of smaller candidate sets, and thereby significant reduction of computational cost for later iterations and reduction of the amount of disk I/O is achieved. An extensive simulation study is conducted to evaluate the performance of the proposed algorithm.

The discovery of association rules problem comes from the need to discover patterns in transaction data in a supermarket. Ale & Rossi (2000) introduced time in the problem of association rules discovery, giving place to temporal association rules. In database transactions, with the explicitly defined time, each item, itemset and the rule has now an associated lifespan. The concept of temporal support is introduced and that gives way to the discovery of new rules. Due to the lack of necessary support, association rules were not discovered with the traditional viewpoint. Now, with the concept of time, consider the rules that have enough support in their lifespan, as long as they also have temporal support. One of the problems related to the discovery of association rules is the generation of large number of rules. When the user ask for dates that are old enough, since the rules with lifespan previous to those dates would be considered obsolete, and is not presented to the user. Furthermore, if the algorithm used to generate the frequent itemset finds old items or itemset, it may eliminate them directly, but it would be an additional pruning.

Han et al. (2004) proposed a novel method for storing compressed, crucial information about frequent patterns, which is an extended prefix-tree structure named frequent pattern tree (FP-tree). An efficient FP-tree based mining method and a pattern fragment growth based FP-growth method for mining the complete set of frequent patterns was developed. Three techniques used for efficient mining were: (1) FP-tree avoids costly, repeated database scans by compressing a large database into a condensed, smaller data structure, (2) To avoid the costly generation of a large number of candidate sets, the FP-tree based mining adopts a pattern fragment growth method, and (3) the partitioning based on divide and conquer method that decomposes the mining task into a set of smaller tasks, thereby reducing the search space. FP-growth proves to be an efficient and scalable for both long and short frequent patterns and in the order of magnitude, it is faster than the Apriori and other recently reported frequent pattern mining methods. In spite of applying various interesting measures to association rule mining of web usage data, challenge remains in extraction of useful information.
Pi-lian (2005) proposed an improvised AprioriAll algorithm based on the original AprioriAll algorithm, which has been widely used in web log mining. In this, UserID property is added at every step of generating candidate set and the database scan, thereby to decide whether an item in the candidate set could be placed into the large set and subsequently used to produce next candidate set. An attempt is made to reduce the number of database scan that limits the size of the candidate set whenever generated. Results show that the algorithm exhibits less time and space complexity, and effectively fits to the memory capacity.

Velu et al. (2007) found interesting patterns from databases (DBs), such as Association Rules (ARs), correlations, classifiers, clusters and many more. The motivation for searching ARs to examine customers buying behavior is discussed in this work. ARs describe how often items are dependent on each other when purchase together. For instance, an AR (bread 100%, peanut butter 80%) states that four out of five customers who bought bread also purchased peanut butter. Those rules proved to be useful for making effective decisions concerning product pricing, promotions, store layout and etc. A detailed explanation of analysis and descriptions of algorithms like AprioriHybrid, AprioriTid, Eclat algorithm, Continuous Association Rule Mining Algorithm (CARMA), and Direct Hashing and Pruning (DHP) algorithm are done.

Sun et al. (2007) first discussed the limitations of the Apriori algorithm and then proposed an enhancement for improving its efficiency. The improved algorithm is based on the combination of forward scan and reverse scan of a given database. It can greatly reduce the scan time required for the discovery of candidate itemset when certain conditions are satisfied. To rationalize the algorithm theoretical proof and analysis are given. To show the advantages of this algorithm compared with Apriori, a simulation instance is specified.

Xie et al. (2008) proposed an IApriori algorithm that can reduce the scan times of database, and optimized the join procedure of frequent item sets generated in order to reduce the size of the candidate item sets. The result shows that the algorithm is better than Apriori algorithm.

Yang et al. (2009) found unsuspected relationships and summarized the data which is both useful and understandable to the owner of the data, by analyzing large observational data sets. Association rules are very popular data mining technique to
show attributes value conditions that occur frequently together in a given dataset. The Apriori is an efficient association rule mining algorithm. The algorithm presented, scans the database only once to generate frequent item sets, thereby saving time and increasing efficiency. These methods even though focused on reducing time and space, in real time still needs improvement.

Ezeife and Liu (2009) proposed two algorithms, RePL4UP (Revised preorder linked WAP(PLWAP) For UPdate), and PL4UP (PLWAP For UPdate), which uses the PLWAP tree structure to incrementally update web sequential patterns efficiently without scanning the whole database, even when previous small items become frequent. The RePL4UP concisely stores the position codes of small items in the database sequences in its metadata during tree construction. Also during mining, only the new additional database sequences are scanned, the old PLWAP tree is revised to restore information on previous small items that are frequent. Previous frequent items that have become small using the small item position codes are deleted. PL4UP initially builds a bigger PLWAP tree that includes all sequences in the database using a tolerance support, t, which is lower than the regular minimum support, s. The position code features of the PLWAP tree are used to efficiently mine these trees to extract the current frequent patterns when the database is updated. These approaches more quickly updated the old frequent patterns without the need to rescan the entire updated database.

Goswami et al. (2010) introduced a three different frequent pattern mining approaches (Record filter, Intersection and Proposed algorithm) based on classical Apriori algorithm. Among them, Intersection approach proved to be better than Record filter approach and classical Apriori algorithm, and the proposed algorithm proved to be much better than other frequent pattern mining algorithms.

Mabroukeh and Ezeife (2010) presented taxonomy of sequential pattern mining techniques in the literature with web usage mining applications. Investigation was performed on these algorithms by introducing taxonomy for classifying sequential pattern mining algorithms based on important key features. Aim of classification is to enhance the understanding of sequential pattern mining problems, the current status of supplied solutions, and the direction of research in this area. An attempt is made to provide a comparative performance analysis of many of the key techniques and theoretical aspects of the categories in the taxonomy are discussed.
Finding Frequent Sequential Pattern (FSP) remains a challenge in web usage mining. Vijayalakshmi et al. (2010) explored a new frequent sequence pattern technique called AWAPT (Adaptive Web Access Pattern Tree), for FSP mining. An AWAPT combines Suffix tree and Prefix tree for efficient storage of all the sequences that contain a given item. During mining, by assigning the binary codes to each node in the WAP Tree, the recursive reconstruction of intermediate WAP tree was eliminated. Web Access Pattern-Tree (WAP-Tree) mining is a sequential pattern mining technique for web log access sequences, in which the Web Access Sequence Database (WASD) is stored on a prefix tree, like the frequent pattern tree (FP-tree) used for storing non sequential data. It also mines the frequent sequences from the WAP-tree by recursively reconstructing intermediate trees, starting with suffix sequences and ending with prefix sequences. An attempt has been made to AWAPT approach for improving efficiency. AWAPT totally eliminates the need to engage in numerous reconstructions of intermediate WAP-trees during mining and considerably reduces execution time.

Difallah et al. (2011) defined and proposed the notion of action table as the ideal search domain for actions, and then proposed a strategy based on the FP-Tree structure to achieve high performance in rules extraction. The major idea is to reformulate the action rules mining problem into the association mining problem framework using the action table as the new search domain. A particularly suited approach is to use an FP-Tree structure to store the action table and the FP-Growth algorithm to extract association action rules.

Wang and Liu (2011) pointed out the bottleneck of classical Apriori’s algorithm, and presented an improved association rule mining algorithm. The new algorithm is based on reducing the time of scanning candidate sets and using hash tree to store candidate itemset. According to the running result of the algorithm, the processing time of mining is decreased and the efficiency of algorithm is found to be high.

Information about the navigational behavior of the users could be obtained by discovering frequent sequential access (usage) patterns in the web log. It could be applied in the areas of advertising, for creating dynamic user profiles etc. Singh et al. (2012) introduced a new approach for web usage mining, in which graphs are created using web access sequence by sorting web log and useful extraction of sequential access patterns. Given a set of sequences, where each sequence consists of a list of
elements and each element consists of a set of items, and given a threshold for user
specified minimum support, the aim of sequential pattern mining is to generate
subsequences that are frequent. Through the application of mining approaches frequent
sequential web access patterns are generated.

Mishra and Choubey (2012) during pattern discovery phase, frequent pattern
discovery algorithms are applied on the preprocessed data. The result of pattern
analysis phase is to extract interesting knowledge from the generated frequent patterns
and the same is applied for website modification. FP-growth algorithm is used to supply
valuable information about the user's interest by obtaining frequent access patterns
from the web log data.

Kumar (2012) developed and proposed a novel method called K-Apriori
algorithm, to find the frequently accessed web pages from the very large binary web
log databases. The Apriori algorithm extracts a set of frequent itemsets from the data, and
then pulls out the rules with the highest information content. Experimental results reveal
that the proposed method has shown higher performance in terms of objectivity and
subjectivity.

Sudhamathy and Jothi Venkateswaran (2012) an efficient approach for frequent
pattern mining using web logs for the web usage mining is proposed and this approach
is called as Hierarchical Frequent Pattern Analysis (HFPA). In this approach HFPA, the
proposed technique is applied to mine association rules from web logs using normal
Apriori algorithm, but with few adaptations for improving the interestingness of the rules
produced and for applicability for the web usage mining. This technique is applied and
its performance is compared with that of classical Apriori mined rules. The results
indicate that the proposed approach HFPA not only generates far fewer rules than
Apriori based algorithms (FPA), the generated rules are also of comparable quality with
respect to three objective performance measures, confidence, lift and conviction. Large
collection of association rules generated through association mining methodologies, are
found difficult to understand when set into action. Proposed effective pruning techniques
are characterized by the natural web link structures, and the experiments reveals that
the discovered association rules can successfully be sorted using interestingness
measures. From the web administrator's viewpoint, most of the highly ranked rules
according to the interestingness measures proved to be truly valuable.
Yadav et al. (2012) proposed an efficient Web Miner or E-Web Miner algorithm known for its valid results and computational comparative performance analysis verification. Drastic reduction in the number of data base scans and the smaller candidate sets in stage wise comparison with Improved AprioriAll Algorithm, E-Web Miner are found to be successful in any web log analysis.

Data mining is a field, which computationally processes the data collected and is able to help the analyst for proposing the ideas for some betterment of the company. The user access is recorded in log files. The web server logs provide important information. In the field of web mining, the analysis of the web logs is done to identify the user search patterns. Sharma and Bala (2014) contradicts from the usual approaches of finding the patterns, that is creating the pattern tree and then the analysis is done, but in this proposed algorithm there is no need of tree creation and the analysis is done based on the website architecture, which will increase the efficiency of the other pattern matching algorithms and needs only one database scan.

Jalgaon (2015) proposed an architecture that contains various modules which includes preprocessing, Web Access Sequence (WAS) generation and user profile creation, discovering interesting usage patterns using proposed efficient sequential access pattern mining algorithm and finally module for personalized recommendations. The proposed algorithm does not generate costlier WAP tree at any stage and it also eliminated the need for projected database by saving space and time. The new approach for sessionization results in the generation of accurate frequent patterns. When the same user issues same/similar query, the system generates improved recommendations.

Santhosh Kumar and Rukmani (2010) discovered the web usage patterns of websites from the server log files using Apriori algorithm and frequent pattern growth algorithm. The main problem associated with association rule mining is the frequent itemset problem where the items that occur together with a high frequency will also appear together in many of the resulting rules and thus, resulting in inconsistent predictions. As a consequence, a system cannot give recommendations when the data set is large. In addition to this, AR algorithms using multiple support thresholds results in better coverage but did not improve accuracy.

Peng (2010) researched how to dig interesting association rules effectively from the web logs after been preprocessed. The FP-growth algorithm is used to obtain a set
of frequent access patterns from the preprocessed web log records. A new pattern is discovered to provide valuable data for the site construction in web mining by combining browser interestingness and site topology interestingness of association rules.

The survey findings reveal that more attention is necessary to address the issues related to reduction in the number of database scan and memory space with less execution speed. These limitations and other related issues motivated us to continue the research work in this area. Even though different versions of Apriori are available, the problem with Apriori is that it generates too many 2-item sets that are not frequent. Extracting useful information from the set of generated association rules remains a difficult task, though various interestingness measures have been applied to association rule mining of web usage data.

The findings of the literature survey on pattern discovery methods are summarized in Table 2.2.

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Technique</th>
<th>Observations</th>
</tr>
</thead>
</table>
| 1997 | Park et al     | Hash Based Method With Transaction Trimming For ARM | • Number of candidate 2-itemsets generated is comparatively less, thereby resolves the performance bottleneck.  
• The amount of disk I/O required is reduced.  
• Transaction database trimming is done at early stage thereby reducing the computational cost. |
| 2000 | Ale & Rossi    | Temporal Association Rules                    | • Direct elimination of old items in frequent itemsets.  
• Pruning is carried out as an additional task. |
| 2004 | Han et al      | Frequent- Pattern Tree (FP-Tree) Approach     | • A repeated scan is avoided by compressing database into a condensed data structure.  
• Efficient and scalable for mining both long and short frequent patterns.  
• In the order of magnitude it is faster than the Apriori and other recent methods. |
<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Technique</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Pi-lian</td>
<td>Improved Apriori</td>
<td>• Exhibits less time and space complexity and better memory utilization.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AprioriAll</td>
<td>• The number of database scan is reduced.</td>
</tr>
<tr>
<td>2009</td>
<td>Yang et al</td>
<td>Improved Apriori</td>
<td>• Database scan is done only once.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Focus on reducing time and space, is required.</td>
</tr>
<tr>
<td>2009</td>
<td>Ezeife and Liu</td>
<td>RePL4UP and PL4UP</td>
<td>• PLWAP tree structure is used incrementally thereby patterns updated without need for rescan.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Tolerance support t is used which is lower when compared to regular minimum support.</td>
</tr>
<tr>
<td>2011</td>
<td>Wang and Liu</td>
<td>Improved Association Rule Mining</td>
<td>• Uses hash tree to store candidate itemsets thereby reducing the time of scan.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Execution time is reduced thereby increasing the efficiency of algorithm.</td>
</tr>
<tr>
<td>2012</td>
<td>Kumar</td>
<td>K-Apriori Algorithm</td>
<td>• Rules with highest information content are extracted.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Shows higher performance in terms of objectivity and subjectivity.</td>
</tr>
<tr>
<td>2015</td>
<td>Jalgaon</td>
<td>Web Access Sequence (WAS) Generation</td>
<td>• Avoids generating costlier WAP tree and eliminates the need for projected database.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Generates accurate frequent patterns and when the same user issues same/similar query, the system generates improved recommendations.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Attention is required to address the issues like reduction in the number of database scan, memory space, and execution time.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• The problem with Apriori is that it generates too many 2-item sets that are not frequent.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Need for automatic support calculation.</td>
</tr>
</tbody>
</table>

Summary of Findings From Literature Review
2.3 PATTERN ANALYSIS - RANKING METHODS

On completion of pattern discovery phase, the usage patterns are generated. To maximize the benefits for analyst from the web usage patterns, tools and techniques are needed to make them more understandable. Being the last stage of web usage mining technique, pattern analysis is used to extract interesting patterns from the results obtained during the pattern discovery phase. Some of the techniques comprise graphics and visualization, database querying, statistics and usability analysis.

Song et al. (2013) proposed a new ranking approach called FP-Rank, which adopt frequent pattern mining algorithms to mine frequent patterns, and then a new pattern selection algorithm is implemented to select a set of patterns with high overall significance and low redundancy. The experiments on the real datasets confirm that, by incorporating effective frequent patterns to train a ranking model, such as RankSVM, the performance of the ranking model can be substantially improved. Finally, the extended dataset is used to train the ranking model and the selected patterns are used to extend the original feature space of training dataset.

Song et al. (2015) proposed a methodology that sets the support value when performing frequent pattern mining. Since Learning To Rank (LTR), frequent patterns are not equally effective, and to further provide coverage based on pattern generation algorithm for discovering effective patterns. A ranking approach called Significant Frequent Pattern based Ranking (SFP-Rank), in which the original features as well as the significant frequent patterns are used to built the ranking model.

Sun et al. (2008) proposed feedback providing model that accurately personalized ranking services tailored to individual information retrieval systems. Comparison was done with several non personalized ranking methods including ranking Support Vector Machine (SVM) light as well as several ranking functions specific to the academic document domain. The results illustrated that proposed ranking algorithm can reach better accuracy in comparison to expected for ranking SVM light and less for all other single feature ranking methods. Also revealed how the derived personalized ranking vectors can be employed for other ranking related purposes such as recommendation system. Experimental analysis reveals that, when user preferences are stable over time, proposed method is more accurate in predicting user preferences than any other non personalized ranking methods.
Dali et al. (2010) in their work, by using demographic information, personalized by some news portals, developed and tested a personalized ranking model based on click stream web logs. Reranking of results is done based on the information that is fetched during user registration. Furthermore, the investigation is carried on how much the user’s demographic data determines the nature of news articles that user reads.

Zhuang and Cucerzan (2006) proposed a Q-Rank that is used to refine the ranking of the search results by constructing the query context from search query logs. Definitions of the query context are extracted from the query logs in order to extract the context of the new query. Using the extracted context the results are reranked. The proposed Q-Rank method effectively refines the ranking of search results for any given query by constructing the query context from search query logs. Through the evaluation results, a Q-Rank gain a considerable advantage over the current ranking system and improvises the significance of search results for 82% of the queries for a large scale commercial web search engine.

Aktas et al. (2004) introduced a new method for personalization of PageRank vectors by assigning weights for links by matching hyperlinks with user profiles. In particular, here a profile representation is described using internet domain features extracted from URLs. Users specify interest profiles as binary vectors where each feature corresponds to a set of one or more DNS tree nodes. In a given profile vector, a weighted PageRank is computed by assigning weight to each URL by matching the URL with the profile. Results were studied from an experiment in which users were allowed to select among nine URL features, combining the top two levels of the Domain Name Service (DNS) tree, leading to $2^9$ precomputed PageRank vectors from a Yahoo crawl. Personalized PageRank performed favorably compared to pure similarity based ranking and traditional PageRank.

Al-Saffar and Heileman (2007) compared personalized and topic sensitive local PageRanks to the global PageRank showing experimentally how similar or dissimilar results of personalization can be to the original global rank results and to other personalization. The proposed approach examined a snapshot of the web and determined how advantageous personalization can be in the best and worst cases, and how it performs at various values of the damping factor in the PageRank formula.
Haveliwala (2003) computed the topic sensitive PageRank scores for pages satisfying the query using the topic of the context. In context based searches (e.g., when the search query is done by highlighting words in a web page), when the query appears, the topic sensitive PageRank scores are computed using the topic of the context. More accurate rankings can be generated with a single, generic PageRank vector by using linear combinations of these (precomputed) biased PageRank vectors to generate context specific importance scores for pages at query time.

Sharma et al. (2010) learned historical query logs and from which the results are optimized so that the user intended pages are ranked higher. Queries from the logs are clustered using the similarity function and the sequential patterns from the selected web pages are captured and based on the patterns the results are reranked. A novel result optimization technique is presented based on learning from historical query logs, which predicts users’ information needs and reduces their navigation time within the result list. The method first performs query clustering in query logs based on a novel similarity function and then captures the sequential patterns of clicked web pages in each cluster using a sequential pattern mining algorithm. Finally, search result list is reranked by updating the existing PageRank values of pages using the discovered sequential patterns. The proposed work results in reduced search space as user intended pages tend to move upwards in the result list.

Barouni-Ebrahimi et al. (2008) by using frequency meaning based algorithm, the appropriate results are reranked, according to the frequent phrases from the past queries that are obtained. An online page rerank model is proposed, which relies on the users’ clickthrough feedbacks as well as the frequent phrases from the past queries. The method is compared with a similar page rerank algorithm called I-SPY. The results proved the efficiency of the proposed method, in ranking the more related pages on top of the retrieved list while monitoring a smaller number of query phrases in a hit matrix. Employing thirteen months of queries for the University of New Brunswick search engine, the hit matrix in the algorithm was on average 30 times smaller, while it showed better performance with regards to the rerank of web search results. The proposed rerank method is expandable to support user community based searches as well as specific domain web search engines.

Agichtein et al. (2006) modeled the user behaviors and by learning those models the preferred results for the users are predicted. User behavior beyond click
through are modeled so that the reranking thus obtained is far better than the one that is obtained by considering only click through methods. A real world study of modeling the behavior of web search users is presented to predict web search result preferences. Applications like ranking, click spam detection, web search personalization, and other tasks makes use of accurate modeling and interpretation of user behavior. Robustness of interpreting feedback is improved as query dependent deviations are modeled from the expected "noisy" user behavior. From the results obtained, it is understood that the higher preference prediction accuracy is improved by clickthrough methods.

Kavitha et al (2010) and Ratnakumar (2008) proposed a weighted URL ranking algorithm that is used to rank the web search results based on the features extracted from hyperlinks, anchor terms and user interested domains. The retrieved results from the search engines are weighed according to the occurrence of tokens and are again weighed in accordance with the user interested domain and the same are retained for reordering the results according to the match with the query weight.

Kumar et al. (2010) used different algorithms like Page Rank (PR), Weighted Page Rank (WPR) and Hyperlink-Induced Topic Search (HITS) for link analysis, and are discussed and compared. Focus on the Hyperlink analysis, and for link analysis, to compare those algorithms and to study the role of hyperlink analysis in web searching. In the hyperlink analysis, the number of incoming links to a page and the number of outgoing links from that page, and the reliability of the linking is analyzed. Authorities and hubs concept of web pages will be explored. The formula used by those algorithms will be explored.

Paul Alexandru et al. (2004) presented a classic algorithm such as HubFinder algorithm to find the related pages and the result is used to provide a platform for personalized ranking. HubFinder algorithm shows how to automate this hub selection process and build upon the latter algorithm to implement a platform for personalized ranking. Started from the set of bookmarks collected by a user and extending it to contain a set of hubs with high PageRank related to them. To get additional input about the user, a proxy server is implemented which tracks and analyzes user’s surfing behavior and generates a set of pages preferred by the user. This set is then enriched using our HubFinder algorithm, which finds related pages. The user’s bookmarks are used as input and the hubs with higher page rank are filtered for further processing thereby contributing for personalized ranking.
Harb et al. (2009) designed a personal search engine which provides relevant results according to user's interests. To ensure relevant and accurate results, the search engine depends on three factors. The first factor is the degree of importance of the document category to the user. The second factor is the user's interest page rank which depends on the user's browsing of the page. The third factor is the degree of relevance of the document. Three factors contributing to the accurate retrieval of results are important for the document category, user interest and the degree of relevance of the document.

Richardson et al. (2006) proposed a work that suggested that PageRank may not perform better than other simple measures on ranking web pages in general purpose search engines. From web logs, information such as IP addresses, time stamps, and requested pages can be extracted, which can then be used by the web application to infer implicit feedbacks of users such as motivations, goals and preferences. Much research has been done to apply data mining and machine learning technologies to study implicit user feedback to improve ranking.

Ranking documents effectively for large scale information retrieval systems continues to be a challenging task. In web search, more recently, machine learning technology has been applied to rank the documents in search engines, (Richardson et al. 2006; Xu and Li. Adarank 2007). Clickthrough data extracted from search engine query logs are used to train ranking models and globally optimize the search results. Machine learning methods are also used for static ranking and show an increase in relevance compared to PageRank, where a standard neural network back propagation algorithm RankNet is trained on the features of web pages. Boosting is also used in ranking documents for information retrieval systems, where explicit user feedback is used as the training data to improve the ranking performance.

Aktas et al. (2004) made use of the personalization vector of the PageRank algorithm, a great deal of focus is done on personalized ranking research. URL features are used to obtain a topical and geographically biased PageRank.

Fogaras and Racz (2004) proposed a work focusing on the scalability and performance of personalized PageRank algorithms using graph mining methods to generate personalized views of document importance. These algorithms present a categorical perspective on personalization. It is difficult to expand the idea to a large
number of categories since these algorithms require a precomputed set of ranking scores for each category, which is given a priori and is difficult to adjust dynamically.

Preference learning has been previously used to identify interesting patterns in an interactive manner. Xu et al. (2006) investigated learning a user specific ranking of frequent patterns (primarily itemsets and sequences). A clustering based method similar to information retrieval approaches is used to select patterns for feedback. However, they only considered a specific learning target based on the discrepancy between the expected and the observed supports of a pattern, and they do not use the learned functions to search for novel patterns.

Rueping (2009) demonstrated the feasibility of learning subgroup rankings and applying learned ranking functions to discover high quality subgroups. However, Rueping does not discuss active learning aspects but used a custom variant of the learner and data modifications that are specific to subgroup discovery are done.

Eyke Hullermeier and Johannes Furnkranz (2011) presented preference learning, which is a research area, encompassing several tasks related to learning preferences within the field of machine learning. Toshihiro Kamishima et al (2011) deal with an instance of the object ranking problem that is acquiring ranking functions from sample orders. Active object ranking is related to the problem of learning to rank in information retrieval, as both aims at learning a ranking from a minimum number of sample rankings. Zuobing Xu et al. (2007) developed a number of general heuristics aimed at improving top results of search engines.

Abdulmohensen Algarni et al. (2010) found web systems utilize the user relevance feedback to interpret the user’s information needs. Vector space model computes the similarity between the query and the document and is based on the terminological overlap between them. Relevance feedback requires the user to classify the documents in to relevant and irrelevant groups. Rocchio algorithm is used to expand the queries from the feedback thus obtained.

Hyoung rae and Philip K.Chan (2005) studied that users are generally reluctant to provide information on whether they are interested with a particular document or not, so relevance feedback is not satisfying mechanism to fulfill the user needs. Web personalization could be achieved by organizing the user profile as User Interest Hierarchy (UIH). UIH tracks the user interest implicitly and Density-based Hierarchical
Chapter - 2

Clustering (DHC) algorithm is used for the same in order to classify the results. Different characteristics of a term are derived and accordingly the terms are scored. UIH could be refined by specifying two new characteristics namely term and node specificity.

Safaa et al. (2015) proposed an Efficient Hybrid Usage based Ranking Algorithm called EHURA. It was applied to 1033 English Corpus to measure its performance. When compared with the content based ranking algorithm representation, EHURA shows improvement over precision while realizing approximately the same recall percentage.

These features are used to rerank the top results, but the same approach failed to handle some new user queries. There are still several challenges in applying the principles. However, such work has not modeled the preferences for each individual user to provide personalized ranking.

The findings of the literature survey on ranking methods in pattern analysis are summarized in Table 2.3.

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Technique</th>
<th>Observations</th>
</tr>
</thead>
</table>
| 2013 | Song et al        | FP-Rank                       | • A set of patterns with high overall significance and low redundancy is selected.  
• Performance of the ranking model is substantially increased. |
| 2008 | Sun et al         | Personalised Ranking Algorithm | • More accurate in predicting user preferences.  
• More efficient when user preferences are stable over time. |
| 2006 | Zhuang and Cucerzan | Q-Rank Search               | • Ranking of search results is done by constructing query context from web logs.  
• Significance of search results is improvised. |
<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Technique</th>
<th>Observations</th>
</tr>
</thead>
</table>
| 2007 | Al-Saffar and Heileman  | Personalized And Topic-Sensitive Local PageRanks | - Determines how personalization can be in the best and worst cases.  
- Analyses how performance at various values of damping factor in the PageRank formula differs. |
| 2008 | Barouni-Ebrahimi et al  | Frequency Mining Based Algorithm For Reranking | - Reranking is done based on frequent phrases from the past queries obtained.  
- Time complexity needs improvement. |
| 2006 | Agichtein et al         | Learning User Interaction Models              | - User preferences are predicted by learning user behavior models.  
- User behavior beyond clickthrough is modeled to enhance Reranking.  
- Results in higher preference prediction accuracy. |
| 2009 | Harb et al              | Personal Search Engine and Modified Page Rank | - To ensure relevant and accurate results, factors like document category, user interest and degree of relevance are considered.  
- Provides relevant results according to user’s interests. |
| 2004 | Fogaras and Racz        | Scalable Personalized PageRank                 | - Presents a categorical perspective on personalization.  
- Precomputed set of ranking scores for each category is required to improve scalability. |
| 2009 | Rueping                 | Learning Subgroup Ranking                     | - Uses a customer variant of the learner and data modification.  
- Does not discuss active learning aspects. |

Summary of Findings From Literature Review

- New features to be explored to rank user preferred queries.  
- Reviewed work has not modeled the preferences for each individual user to provide personalized ranking.
2.4 PATTERN ANALYSIS - CLUSTERING METHODS

Understanding the customer preferences and the requirements in time is a premise to optimize these web services. In this section, clustering models for generating and maintaining clusters which represent the changing web user patterns in websites are discussed. The clustering model can be fast updated to reflect the current user patterns to the web administrators. The topic of web session clustering has become popular in the field of practical application of clustering techniques in recent years.

Yan et al. (2010) introduced a two step K-means cluster algorithm to mine web usage patterns with four attributes extracted from collected data, namely page access times, category number, relative entropy and element of categories can reflect users’ activities, information scope and preference which can embody users’ requirements in different aspects. The analysis results concluded that, clusters provide some heuristic ideas to design targeting or recommending applications. Considering usage features of activities, information scope and preference, proposed algorithm searched user groups in realistic data collected from Wide Area Network (WAN).

Khasawneh and Chan (2005) proposed an approach, in which web usage patterns are represented as rules generated by the inductive learning program, BLEM2. Inputs to BLEM2 are clusters generated by a hierarchical clustering algorithm applied to preprocessed web log records. Empirical results show that the prediction accuracy of rules induced by the learning program is better than a centroid based method and also the usage of learning programs can generate shorter cluster descriptions.

Yaxiu et al. (2009) described a method for clustering similar web user, by considering the page click number and web browsing time, available in the web log. The technique proposed in this work can help web site organizations to recommend web pages, improve web structure, so that it can attract more customers, and increase customers' loyalty.

Zhang et al. (2009) made an attempt to improve the fuzzy clustering algorithm that find groups with common interests and behaviors by analysis of web log data. The experience made in this project points at the potentials web usage mining offers for financial services. The goal of increasing the number of registered customers by focalized marketing campaigns to an interesting target group has been reached. This
particular application hints at the potentials, web usage mining offers for improved Customer Relationship Management (CRM) e.g. in financial services.

Li (2009) introduced a new method for measuring similarities between the web pages that takes into account not only the URL but also the viewing time of the visited web page, and a new method to measure the similarity of web sessions using sequence alignment and the similarity of web page access is given in detail. Finally, an algorithm of web session clustering is proposed. This algorithm defines the number of clusters according to the knowledge of application fields, takes advantage of Robust Clustering Algorithm for Categorical Attributes (ROCK) to decide the initial data points of each cluster and determines the criterion function according to the contributions of overall increase in similarities made by dividing web sessions into different clusters, which not only overcomes the shortcomings of traditional clustering algorithm, which merely focus on partial similarities, but also decreases the complexities of time and space.

Mustapha et al. (2009) introduced an advanced model for mining of user's navigation pattern, based on Expectation Maximization (EM) algorithm that is used for finding maximum likelihood estimates of parameters in probabilistic models that deals with the unobserved latent variables. The results shows that the decrease in the number of clusters converge the log likelihood towards lower values and the increase in the number of clusters decrease the probability of the large clusters. It also indicates that the visit coherence (accuracy) of navigation pattern mining has improved.

Li and Cui (2007) studied the clustering algorithms based on learning models on model data. By defining the extended measure, clustering methods are studied for the abstract data objects. They presented the framework for clustering objects of model and studied the relations between the number of cluster in clustering analysis, and the size and performance of ensemble learning. To validate the effectiveness of models, the hierarchical model clustering algorithm is chosen for the experiments.

Yu et al. (2005) proposed a unifying generative framework for partitioned clustering algorithms according to a novel definition of the mean, called a General c-means Clustering (GCM) model. GCM, a unifying generative framework for partitioned clustering algorithms, is presented and studied according to a novel definition of the mean. The connection between Occam's razor and partitioned clustering is established for the first time based on the local optimality test of the GCM. For application purpose, a complete review of the existing objective function based clustering algorithms is
offered based on GCM. A theoretical guide for developing and applying clustering algorithm is discovered based on a common assumption about partitioned clustering. The conclusions are verified numerically by experimental results.

Wenchao et al. (2007) proposed a novel clustering algorithm based on k-means and hierarchical clustering, which has good computational complexity. Initially, the concept of silhouette coefficient is introduced and the optimal clustering number $K_{opt}$ included in data set of unknown class information is decided. Then the distribution of data set is obtained through hierarchical clustering and clustering center is decided. Finally, the clustering is completed through K-means clustering. The efficiency of the algorithm is validated through the test of IRIS testing data set. Experiments show that the K-means algorithm can give us more satisfactory results by escaping from the sensibility to initial value and improving the accuracy of clustering.

Suresh et al. (2011) presented that clustering method is very sensitive to the initial center values, requirements on the data set is too high, and cannot handle noisy data type. The proposed method used information entropy to initialize the cluster centers and introduce weighting parameters to adjust the location of cluster centers and noise problems. The navigation data sets are sequential, and clustering web data finds the groups which share common interests and behavior by analysis of web logs, thereby effectively improving clustering on web datasets using improved Fuzzy C-Means (FCM) clustering.

Ghosh and Dubey (2013) studied that the outcome of the clustering process and efficiency of its domain application are generally determined through algorithms. There are different algorithms which are used to solve this problem. Two main clustering algorithms, (i) centroid based K Means and (ii) representative object based FCM clustering algorithms were compared. They were experimented and evaluated on the basis of the efficiency of clustering output. The behavior patterns of both the algorithms are analyzed based on the number of dataset points and the number of clusters. The results of FCM were similar to K-Means clustering except for more computational time.

The advantage of the partition based algorithms is that they use an iterative way to create the clusters, and the drawback is that the number of clusters has to be predetermined and spherical shapes can only be determined as clusters.
Density-based algorithms (DBSCAN) initially searches for core objects, and the clusters are grown based on the core objects and the neighborhood objects within a radius of a given object are searched. The main advantage is to detect arbitrary form of clusters and to filter out the noise.

Santhisree et al. (2010) have presented a technique to cluster web transactions using rough set DBSCAN clustering algorithm. They have considered datasets of msnbc.com and found user access patterns, the order of visits of the hyperlinks of each user, and inter cluster similarity among the clusters.

Santhisree et al. (2010) presented Rough set Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm which identifies the user behavior through page visits and order of occurrence of visits. The rough set similarity upper approximations are used for forming the web data clusters. Experimental results on MSNBC web navigation dataset proved that Rough set DBSCAN clustering has better efficiency and performance in web usage mining to find the groups with common interests, when compared to rough set agglomerative clustering.

Langhnoja et al. (2013) made an attempt to find visitor group with common behavior from web logs using DBSCAN clustering algorithm.

The findings of the literature survey on clustering methods in pattern analysis are summarized in Table 2.4.

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Technique</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>Yan et al</td>
<td>Two-Step K-Means Cluster Algorithm</td>
<td>• Capable of handling categorical attributes by converting them into numeric.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Setting of parameters remains a challenging task.</td>
</tr>
<tr>
<td>2009</td>
<td>Zhang et al</td>
<td>Improved Fuzzy Clustering Algorithm</td>
<td>• Finds groups with common interest and analyses user behavior of web log data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Euclidean distance measures can unequally weight underlying factors.</td>
</tr>
<tr>
<td>Year</td>
<td>Author</td>
<td>Technique</td>
<td>Observations</td>
</tr>
<tr>
<td>------</td>
<td>------------------</td>
<td>--------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2009</td>
<td>Li</td>
<td>Web Session Clustering</td>
<td>• Initial data points of each cluster are decided using ROCK.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Criterion function is determined according to the increase in similarities.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• It also focuses on partial similarities.</td>
</tr>
<tr>
<td>2009</td>
<td>Mustapha et al</td>
<td>Expectation-Maximization (EM) Algorithm</td>
<td>• Used to find maximum likelihood estimates of parameters that deal with unobserved latent variables.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Improves the visit coherence (accuracy) of navigation pattern mining.</td>
</tr>
<tr>
<td>2005</td>
<td>Yu et al</td>
<td>General C-Means Clustering (GCM) Model</td>
<td>• It presents a unifying generated framework for partitioned clustering.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• It provides a theoretical guide for developing and applying clustering algorithm.</td>
</tr>
<tr>
<td>2007</td>
<td>Wenchao et al</td>
<td>Hierarchical And K-Means Clustering</td>
<td>• It escapes from the sensibility to initial value and improves the accuracy of clustering.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Handles unknown classes.</td>
</tr>
<tr>
<td>2011</td>
<td>Suresh et al</td>
<td>Improved FCM Clustering</td>
<td>• Information entropy is used to initialize and adjust the cluster centers and handles noise problems.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Handles sequential dataset.</td>
</tr>
<tr>
<td>2013</td>
<td>Ghosh and Dubey</td>
<td>Centroid Based K Means &amp; Object Based FCM</td>
<td>• Behavioral patterns are compared based on the number of dataset points and number of clusters.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Comparison results were found to be similar.</td>
</tr>
<tr>
<td>2010</td>
<td>Santhisree et al</td>
<td>Rough set DBSCAN clustering</td>
<td>• It identifies the user behavior through page visit and order of occurrence of visit.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• The rough set similarity upper approximations are used for forming web data clusters.</td>
</tr>
</tbody>
</table>
### Chapter 2

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Technique</th>
<th>Observations</th>
</tr>
</thead>
</table>
| 2013 | Langhnoja et al | DBSCAN clustering | • Attempts to find visitor group with common behavior from web logs.  
• DBSCAN suffers from border point’s problem. |

Summary of Findings From Literature Review

- Existing clustering algorithm require overall performance improvement.
- DBSCAN suffers from border point’s problem as it treats all border points as noisy.

#### 2.5 CHAPTER SUMMARY

From the review of the literature, it is established that lot of research is carried out till date on web usage mining framework for extracting interesting patterns from weblogs. During preprocessing session identification of web log remains a challenge, since each user maintains multiple sessions for the specific duration. To solve this problem automatic session identification is performed based on the timeout method, in which the session is differentiated based on the time interval with predefined threshold value. But, it is difficult to set the time threshold for each session identification process. During pattern discovery phase, occurrence of too many frequent itemset, and the time for generating association rules is much longer. In discovering frequent patterns the challenge lies in the selection of min support and min confidence value. There exists a need to reduce the number of database scan and memory space with less execution speed. Analysis of user preferred query pattern becomes challenging task. To overcome the limitations stated in the observations due to literature, a web usage mining framework has been proposed in this research work and is discussed in the subsequent chapter.