CHAPTER 2

REVIEW OF THE RELATED LITERATURE

2.1 INTRODUCTION

In this chapter we present the related literature surveyed throughout our study on enhancing user personalization. We organize this chapter mainly into three sections. In the first section we describe about the various works in classification and usage mining related to our proposed method. Second section is about the various link representation methods and in the last section sequence generation and pattern analysis algorithms and methods related literature is briefed. Section 2.5 summarizes the chapter.

2.2 USAGE MINING FOR PERSONALIZATION

Data mining techniques (Jiawei Han and Micheline Kamber 2006) specially usage mining and analysis of a website can serve as a source to identify user behaviour [Lara D Catledge & James E Pitkow 1995), Oren Etzioni (1996), Madria et al. (1999), Srivastava et al. (2000), Shingo Otsuka et al. (2004), Mirghani. Eltahir & Anour, Dafa-Alla (2013)]. This idea can be used for personalizing web pages for user preferences. One of the initial studies in this area, Henry Lieberman (1995), suggested the idea of browsing the web with the help of an agent. Thorsten Joachims (2002) progressed in this area by proposing an automatic optimization method for search engines and relevance ranking of a page based on click through data. His work proposed a supervised learning based on Support Vector Machine for ranking. The proposed study focus on effective personalization of site by first classifying user preferences.


2.2.1 Related Literature on Classification and User Behaviour

Since the user behaviour may change over time we need to consider a probability based statistical method for classification. We based our study on Bayes model [Maritz & Lwin (1989), Stuart Russell and Peter Norvig (2010)] for classification. It deals with probability of occurrence of data based on prior occurrence. Prior occurrence is important in personalization as it is directly related to user behaviour in a site. When need to group data based on probability of occurrence, co occurrence possibility also need to be considered.

Scott Deerwester et al. (1990) studied the need of latent semantic indexing and retrieval of document. Based on co occurrence possibility they represented documents in single space. The semantic space is formed the large array of terms and documents which are closely associated. This is based on the assumption that users may need to retrieve documents based on content rather than individual words. There document retrieval is based on continuous match identified in query or in document and a threshold of selection can be set according to the needs of the user. The method to some extend finds match but it is inefficient in terms of synonyms of a word. The method never considered the error probabilities.

Hofman (1999) proposed a statistical model to identify the hidden semantic relationship between co occurrences in a web page based on data from web usage mining. The core idea is to identify the meaning of a word under different contexts. He proposed a Probabilistic Latent Semantic Analysis (PLSA) model which is probabilistic model for text mining applications. He observed that with each co occurrence of data there is an associated latent factor. PLSA is a statistical version of Latent Semantic
Analysis and thus statistical methods can be employed for various tasks such as filtering and model selection.

Bamshad Mobasher (2000) proposed a web recommendation system for personalization. To extract knowledge for recommending a site, a different clustering technique was used. Usage clusters, which are URL clusters often referred by user are selected for personalization. Matching of frequent patterns is done by selecting repeated item sets in sessions. This system is useful for E-commerce applications and is a web recommendation system for user rather than individual user personalization. Ranking of URL may be based on the total count rather than personalized for the specific user.

Mobasher et al. (2002) pointed out that the finding of patterns from usage data by itself is not adequate for performing the personalization tasks. For effective categorization of user behaviour, Xin Jin et al. (2004), proposed a method of integrating common usage patterns and semantic information’s from web logs. They proposed a probabilistic model based on PLSA for finding the latent relationships. Their study proposed dynamic recommendations based on the personalized web segments. This method effectively made recommendations based on navigational behaviour.

For better capture of user navigational pattern extraction, Guandong Xu et al. (2005a), presented a novel usage mining based method based on PLSA. They combined information from web usage and page linkage to extract the exact user access pattern. Linkage data is the data of user extracted from two hyperlinked pages. EM algorithm, presented in the work is applied to the integrated usage and linkage data to find the latent semantic pattern. This data helps in effectively extracting access style of users. Studying a specific users access behaviour and to propose recommendation without much reconstructing the site is still a challenge.
User Profiling Algorithm, proposed by Guandong Xu (2008) in his research created better user profiles based mainly on three models, namely Latent Semantic Analysis, Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation models. Dirichlet models searches the possibility of occurrence of a word from a group of similar occurrence. Such methods work on the probability of co occurrences percentage to recommend a word. In the study user profiling is done implicitly by learning from usage data. For selection of user profiles a similarity measure between user sessions and user profile is employed. Only active user sessions are selected for this purpose. To represent session similarity measures clustering based on page weights method is adopted.

Dong (2009), proposed a query based usage mining method for extracting access patterns. He suggested that this could enhance decision making of organizations. He illustrated the power of OLAP and data mining techniques for web usage analysis.

Kenneth Wai-Ting Leung and Dik Lun Lee (2010) studied the need of considering positive and negative user preferences for user profiling. Most of the studies in this area ignored the fact of negative user preferences. They considered negative preferences which can to some extend show the dislike factor of user. For example, a user interested in laptop may not be interested in the electronics circuitry or assembling of laptop. They used these negative remarks mainly to separate between similar and dissimilar queries. They clustered similar groups of search by users by a two phase personalized clustering algorithm which concentrated both on user as well as the community. But this also is a web personalization with a consideration on the community impact of the query rather than perfect user personalization. Here similarity of a search vary when community factor is considered.
A user can extract needed information by Information Retrieval methods [Liu (2011)]. IR methods can collect data from the distribution of user information contained in web documents. The study showed the importance of probabilistic models from usage data of a website for finding the access pattern.

Cheng Lihui et al. (2012) proposed a personalized computing model based on individual needs of a user. They suggested that this personalized service can meet user requirements. They proposed a predictive model based on probability terms to set configuration parameters. They observed dynamic parameters obtained through virtual environments can provide personalized services to user.

Thi Thanh Sang Nguyen (2014) proposed a web recommendation system by considering both usage data and domain knowledge of a website. Their method based relied on domain ontology to represent domain knowledge of a website. Then based on the domain knowledge and automatically generated usage knowledge they tried to do the web personalization. Web page recommendations can be done by querying the available knowledge. It partially overcomes the new page problem, which is the difficulty in accessing the sequence on a users’ first visit to a page, since page semantics are also considered for recommendation.

The proposed work incorporates session analysis and probability based classification based on total number of hits on a page for better personalization. The work concentrates on numeric data. The authors use a method based on Bayes classification for classification and thus to recommend for automatic personalization. The authors put a threshold for number of hits on a page and tried to relate the lower and upper bounds from this threshold value. This filtering ensures better personalization.
2.3 VISUALIZATION TECHNIQUES AND RANKING METHODS

Following section briefs various related visualization techniques, link structure and ranking methods of website which formed the motivation.

2.3.1 Visualization Techniques

One of the initial studies on considering maps as a medium of communication was by Alan Hodgkiss (1981). He explained the map making process from basic concepts of scale, orientation etc. in his book “Understanding Maps- A Systematic History of their Use and Development”. He has shown in detail how data gathered can be converted to maps. This can help in representing real world data of web usage, through maps.

Chaomei Chen & Roy Rada (1996), tried to solve the problems, such as disorientation associated with hypertext documents, through graphical maps/browsers. The main challenge for them was the non connections or weak connections between hypertexts. They observed that tools or graphical maps have strong relation in improving the traversal ability of users through hypertexts.

A flexible link structure to represent hypermedia technology can improve user navigation through a website (Paul Kahn 1995). The structure, which is a representation of previously learned experience of user, is found to be having the effect of 3D models. He showed the effect of such design structures to improve navigation results through the example of hypermedia publications.

MAPA system [Durand and Kahn 1998] extracts hierarchical structure of a website to provide better navigation facility in large websites.
The hierarchical structure was extracted using shortest weighted path method. These shortest weighed paths are calculated based on the user assigned and heuristically determined weights. They showed that visualization is a better way in aiding navigation as these interactive maps derived from hierarchy achieved in faster navigation. The problem with detecting of incoming links in websites is overcome in MAPA system by referring to contextual information. The previous traversal paths as well as where to go next and peer group information are integrated through maps to facilitate better navigation.

Thorup (1999) proposed an algorithm to solve undirected SSSP with only integer weights. This algorithm considers only positive weights. This is a betterment to the shortest weight path method in comparison with time complexity but is inefficient to draw a perfect relationship with weight and user traversals.

Munzner (2000) proposed methods for the interactive visualization of large graphs. They showed how large networks with hyperlinks can be represented by graph structure. He studied how the constraint of parameter space affects graph drawing. The work revealed the algorithms for providing layout and navigation of larger graphs. They used Spanning trees for layout by learning from domain information. Their spanning trees representations were similar to Depth First Search of traversal. The larger graphs presented are not for layout alone, but also analyses the spatial structure of hyperlinks. These structures effectively do the domain specific encoding. The graphs serve as a model for development and analysis of visualization systems.

Thorup (2004), studied the factors for recommending reachability in planar digraphs. He observed that by pre processing the graph for a specified time, and the produced oracle can give solution to reachability
queries. Their method approximates distance recommendations for a planar di
graph based on weights.

Wei Yusi (2014) in his research suggested an improvement for the
Thorup (1999) SSSP algorithm. The improved Thorup algorithm is found to
be time efficient and cost effective in its two phases, of construction of
component tree and in the visiting of nodes phase. It uses the concept of a
modified component tree which excludes unvisited details. The improved
Thorup algorithm does not contain unvisited structure. Tentative distance to
components is found by the visit of component tree alone.

2.3.2 Link Structure and Ranking Methods

Page et al. (1998) in their study “The PageRank Citation Ranking:
Bringing Order to the Web”, showed how web pages can be ordered by
associating ranks. The system does a global ranking to web pages. Users’
interest on a website contribute to ranking. The method finds main application
in the area of searching. All the values of URLs are converted to integer
values, for easily ranking and identifying pages. Hyperlinks are given integer
identification values. Their method of page ranking shows a direct proportion
between the rank of a page and sum of ranks of its back links. While
calculating page ranks they avoided dangling links, the links which just points
to a page. They observed that as dangling links do not have any outgoing
links, they do not contribute to the page ranking directly.

Kleinberg (1999) proposed an algorithm named HITS (Hyperlink
Induced Topic Search) which effectively calculates the weights of webpages.
HITS classify webpages in to authoritative pages and hub pages. Authoritative
pages are authorities of topic of search and hub pages are the pages that
connect to authoritative pages. HITS algorithm constructs an adjacency graph
of the page under consideration. For calculating the weight of a page, authoritative weight as well as the hub weight is considered. HITS is an efficient algorithm for page rank calculation but this does not ensure personalization. Google follows this algorithm for ranking pages of a search. Since the total weight is considered for ranking HITS has the disadvantage of the possibility of old web pages ranked high.

David Cohn & Huan Chang (2000) put forward a probabilistic method for identifying the authoritative links from web pages. This is a modification of the HITS iterative approach for ranking pages. HITS rely on the largest eigen matrix value for fixing the authority. This methodology proposes the authority based on the conditional probability. The probabilistic factor for high citation of a page is also considered. The algorithm has the main drawback of not getting a perfect fit for authority factors between iterations.

Monika R Henzinger (2001), studied the need for hyperlink analysis in faster retrieval of data from web. Hyper links contributes much in ranking as they may be linked on the same topic. They pointed out that hyperlink analysis also can serve the purpose of finding quality of pages.

Wenpu Xing & Ali Ghorbani (2004) proposed a Weighted Page Rank algorithm which is considered to be an extension of standard Page Rank algorithm. In Page Rank and HITS algorithms all the links are considered as equal. The algorithms rank pages based on this concept. WPR algorithm enhance ranking by assigning ranks to links based on popularity. In WPR each out link is assigned a rank value based on the proportion of in links and out links. To select the relevant pages only the algorithm divides pages in to four types. Very relevant pages, relevant pages, weak relevant pages and irrelevant pages. This algorithm suffers from the problem of relevancy setting
of a page. They calculated relevancy of a page by taking total count and choosing the highest count value of relevancy. The count is obtained through responses of users of web page. User opinion of relevancy of a page can vary from person to person and this affects ranking.

Jianhan Zhu et al. (2004) proposed a new method to visualize link hierarchy of a website. They had treated the weight of a link as the number of user traversals. The link hierarchy with different conceptual levels is formed then. They ranked the pages in the order of user traversals on hyperlinks based on their proposed Page Rate Algorithm which is a modified version of the Page Rank algorithm. A random surfers’ pattern also is considered for ranking pages using this algorithm. In order to help user for easy navigation they also used clusters to show the link hierarchy. Their PageCluster algorithm clusters related pages. The algorithm considers both in link and out link similarity for grouping of related pages. Their work does not focus to the change in user behaviour over time.

Allan M Schiffman (2006) proposed an improvement of HITS algorithm by including hierarchical link structure while doing link analysis for ranking. They experimented on real web data by introducing the idea of phantom nodes. They incorporated a weighing function in HITS namely link weighing function. This function helps in finding the hierarchical relationship from URL structure. URL as such do not have structure but it is extracted by text analysis and from the obtained information about other pages. Explicitly available graphs showing relations can also contribute to the hierarchy of URL. They also studied the possibility of inter page link relationships which are common in single domain. They pointed out that a page should not get priority in ranking only based on the reference. They measured the quality of page for ranking by associating node weight functions. Their study was
limited to HTML files and the access was only through Hyper Text Transfer Protocol.

Distance Rank Algorithm, proposed by Ali Mohammad Zareh Bidoki and Nasser Yazdani (2008) is considered as an intelligent ranking algorithm. This algorithm calculates high ranks for pages which have less inter page distance. Distance is considered to contribute in inverse proportion to ranking. The logarithmic distance between pages is considered for ranking. The shortest distance value get the high rank. We can fetch pages faster with this algorithm. The main drawback of this algorithm is that it does not consider crawler possibilities.

Qingcai Chen et al. (2008) modeled a link structure based approach for representing web pages. First they represented the link hierarchy as directed graph. To find the priority of topic evaluation SSSP algorithm is used. They considered crosswise link, which is a target page in a different directory. They constructed the site map of web page to faster navigation. Their proposed future work of topic hierarchy is considered in our methodology.

Andreas Harth et al.(2009) proposed a novel approach of ranking for unstructured data. For integrated unstructured data which can be from different sources, they proposed a prioritizing data return algorithm. Their method gives authority based on naming authority graph constructed. For assigning authority page rank method is followed. The authority values are then ranked based on the number of references. The proposed method is a schema independent method. This works well in datasets of web represented in resource description framework.
Renaud Delbru et al. (2010) came up with the idea of finding the locality of data based on a two layer dataset ranking methodology. Their proposed methodology DING finds ranks based on links between datasets and also by considering values dependant on page relationships. The methodology integrates three steps. During the first step ranks are calculated based on the analysis of links. Local link analysis during the second step and finally combining of popular ranked pages with local ranks for finding global ranking is done.

Kumar & Singh (2010) studied the importance of analyzing hyperlink structure in information retrieval. They studied the convergence of various algorithms for ranking. The survey of ranking methods shows the importance of in link and out link hierarchies for ranking.

Yue He et al. (2012) proposed an improvement to the traditional hits algorithm by including the click through data factor for rank calculation. The authoritative weight is calculated in similar way as HITS calculation but the weight factor of a page also contributes to rank calculation. They have added measures to rank pages based on user behaviour by adjusting the click through rate values.

2.4 LITERATURE ON SEQUENCE GENERATION

Following literature on sequence generation and pattern analysis formed the basis of our study on future visit prediction of user.

2.4.1 Patterns for Future Prediction

We are often in need of finding relationship between customers’ transactions from database. Rakesh Agarwal et al. (1993) proposed a novel idea of finding relations based on association rules. One of the initial works in
data mining, their method effectively found relationships of items purchased in a supermarket context. For finding relationship among data sets they also considered additional constraints such as syntactic constraint and support constraint. Syntactic constraints need to be considered as they impose restrictions to items which come based on rule. Such constraints are useful in filtering only needed items for association. The support constraint deals with the effective extraction of support of a rule based on number of transactions.

They classified the item sets to large item sets and small item sets based on the support threshold. This effectively reduced the overhead of finding relationship between frequent item sets. The algorithm gets the large item sets by scanning the database a number of times. They employed a pruning procedure to get better optimization for candidate item sets. The two pruning functions used prunes item sets which need not be considered, on the generation itself. The method also has a better memory management technique, by removing items when memory is full. Such items are considered during the next scan. The work was effective in proposing the idea of finding relationship and associating item sets based on querying the database.

Rakesh Agrawal & Ramakrishnan Srikant (1995) continued the research work and came up with the idea of two famous algorithms AprioriAll and AprioriSome for mining sequential patterns in their published work “Mining Sequential Patterns” in the proceedings of the International Conference on Data Mining in March 1995. Considering transactions having customer id, items list and time of transaction AprioriAll was able to find sequential patterns in five phases.

The algorithm AprioriAll starts with a sort phase where a sorted customer sequence is produced. The sequence is based on customer id and the
time of transaction. Large item sets are then found and a transformation phase where mapping of item sets to integer values of large item set is done. The following is a sequence space where the candidate sequence generation is the main consideration. The output of this phase is a set of sequential patterns. In the last phase the algorithm extracts maximal sequences to produce the optimal sequence. It produces better sequences as it takes the time of transaction in to consideration. During the forward phase of sequencing the algorithm finds patterns up to a particular length limit. The remaining patterns are generated during the backward phase. The algorithm has a major disadvantage of time consuming number of passes over the database to find the ultimate sequence.

AprioriSome is another sequential pattern mining algorithm by the same authors. This algorithm cares about the maximality support. From the generated sequence AprioriSome selects only those ones which have maximal support. Speed of operation is high for AprioriSome compared to AprioriAll. But AprioriAll performs faster in some typical situations where a user selects an item based on the count of people purchasing a related item also. AprioriSome has to make an additional scan for getting such count values.

Srikanth & Agarwal (1996) suggested a performance improvement for sequences in their Generalized Sequential Patterns (GSP) algorithm. With base of Apriori, the algorithm adds time constraints to achieve efficient mined patterns. It follows a gap centric approach to select sequences. The max gap constraint deals the maximum time duration of reoccurrence of an element or sequence. GSP incorporates the idea of sliding window, with size as difference in maximal and minimal transaction time, to generate better sequences. In the forward phase GSP finds successive elements of the sequence in data sequence based on difference between start time, end time
and the maximum gap between any two adjacent transactions. GSP has a backward phase where previous item is pulled based on the time and gap constraints. When compared with AprioriAll and AprioriSome, GSP generates lesser candidate sequence and produces better frequent sequential patterns. But it has the limitation of performance degradation in larger data sets.

Heikki Mannila et al. (1997) proposed a novel approach for finding the repeated sequence of episodes from the given pattern. A group of events that may occur together are termed as episodes. The algorithm starts with scanning small episodes and progress towards finding larger episodes which are frequent. Episodes normally are partially ordered set. Parallel and serial episodes are considered in the process of finding larger frequent episodes. The proposed methodology can effectively handle the collecting of episodes from different domains, which depicts the user behaviour.

Zhang et al. (2001), proposed a successive refinement approach to obtain the sequential patterns in their new two stage MFS algorithm. The work is an enhancement of GSP algorithm by R.Srikanth and R.Aggarwal (1996). The sequences obtained in each previous mining is taken as the estimate for proceeding till it get the final sequence. The generation method in MFS namely MGen produces variable length sequences from a frequent sequence. The methodology works based on the BFS approach. It effectively reduces the I/O cost of GSP by generating only the needed candidates for the next sequence.
Figure 2.1 Activity diagram for Apriori Item set generation

(Source: Musa J Jafar 2010)
Observing the fact that databases often undergo insertions, deletions and updation the same authors (2001), proposed two incremental algorithms-GSP+ and MFS+. The algorithms can produce the updated sequence of frequent patterns. The MFS+ algorithm has reduced number of iterations as it does not consider the unmodified part in the updated database. Pruning techniques applied has impact over the sequences generated in this algorithm.

The works mentioned above in pattern generations focused mainly on Apriori methods for sequence generation. Jiawei Han et al. (2000), put forward the idea of combining the mined frequent sequences and the frequent patterns to form a projected sequence database. The methodology, frequent pattern projected sequential pattern mining—freespan method, follows a pattern growth method and it generates a reduced candidate subsequence. The algorithm first constructs a frequent item representation, generates length-2 sequential patterns, generates item repeating patterns and finally does recursive mining on projected database until all the patterns are obtained. When considered with large database, Apriori algorithms has the main demerit of generating large sub patterns. Less candidate sequences are generated in freespan as it generates sequence based on projected database. In freespan the subsequence mining is limited to the projected database and thus it achieves the results faster.

Roberto J Bayardo et al. (2000), proposed the idea of incorporating user suggested constraints to mining in large databases. The algorithm considers all user specified constraints such as minimum support and minimum confidence. They tried to associate additional rules for finding frequent patterns. The method can effectively be applied to dense data.
The Web Access Pattern Tree (WAP Tree) structure proposed by Jian Pei et al. (2000) is an efficient way of extracting access patterns from web log data. It follows a prefix growth approach of traversal based on suffix pattern to generate sequences.

The tree generation procedure is as follows:

- First the database is scanned to find repeated individual events.

- The second scan of the sequence database is then performed to generate tree for individual events which are frequent.

- It then proceeds with the extracted conditional suffix pattern.

- During the next steps generation of intermediate trees and recursive traversal over the tree to generate the frequent patterns are executed.

The sample access pattern database with set of items \( I = \{ p, q, r, s, t, u \} \) is shown in Table 2.1 [Jian Pei et al. (2000)]. User ID is represented as UID.

<table>
<thead>
<tr>
<th>UID</th>
<th>Access Sequence</th>
<th>Frequent sub sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>pqspr</td>
<td>pqpr</td>
</tr>
<tr>
<td>10</td>
<td>tptqpr</td>
<td>pqpr</td>
</tr>
<tr>
<td>15</td>
<td>qpqupr</td>
<td>qpqpr</td>
</tr>
<tr>
<td>20</td>
<td>puqprur</td>
<td>pqpr</td>
</tr>
</tbody>
</table>

In the WAP Tree generation process after the first scan frequent items are selected. For example in set of items \( p, q, r, s, t \) and \( u \) given in
Table 2.1, support values are taken. For individual items support as \( p=4, q=4, r=4, s=1, t=2 \) and \( u=3 \). The frequent items are selected as the ones above minimum support. Minimum support is found to be 3. Thus only \( p, q \) and \( r \) are selected. The frequent items sequence is given in the last column of Table 2.1.

The representation of item subsequences and how the linkage is made is represented through WAP Tree. The WAP tree generation process for the above data is shown in Figure 2.1. Once the tree generation process is complete, the tree is mined to get the frequent sequence.

During the regeneration process a conditional pattern list of suffix can be obtained by scan from root to a particular page represented as node. The sequence count is found to be same as the count of the node. After storing data in WAP Tree, mining starts with the lowest frequent event \( r \), for getting the frequently repeated sequence.

In 3.2 (iv), the conditional sequence base of \( r \) is calculated by \( pqp:2, pq:1, pqrp:1, qpqp:1, pqp:1 \) and \( pqp:-1 \). The traversal for generating the conditional sequence list of an event is by following the header link the event. Only events that have a minimum support count of 3 are selected as conditional frequent events. \( p(4) \) and \( q(4) \) are found to be frequent. Thus the generated conditional pattern on \( r \) will be \( pqp:2, pq:1, pqp:1, pq:-1, qpqp:1, pqp:1 \) and \( pqp:-1 \). Based on this sequence, conditional WAP Tree| \( r \) of Figure: 2.2(i) is generated.

By recursive scan of this the next subsequence for \( qr \) is built. The process is repeated and after the complete mining process, the frequent set of patterns is \( \{ r, ppr, qpr, pqpr, pr, pqr, qr, q, pq, p, pp, qp, pqp \} \).
Figure 2.2 The WAP Tree generation process

[Source: Jian Pei et al. (2000)]
Figure 2.3 WAP Tree regeneration process

[Source: Jian Pei et al. (2000)]
Mohammed J Zaki (2001) proposed a method based on lattice search to produce frequent patterns. The algorithm Sequential PAttern Discovery using Equivalence classes (SPADE) is an efficient DFS based algorithm. The algorithm first does the decomposition and then employ join operations. It uses a vertical ID list and associates sequences with generated timestamp. It generates Temporary ID list for all the candidates and represents it in a lattice space. It uses temporal joins to generate patterns from the temporary ID list. The support for a k-sequence can be obtained by looking at any two of the (k-1) subsequences intersection. SPADE is very economic in terms of computational cost and is having reduced number of database scans when compared with GSP. It has linear scalability based on the number of candidates and database parameters.

Jay Ayres (2002), suggested a betterment for SPADE algorithm by removing Vertical ID and Temporary ID pair. Their proposed DFS based methodology, Sequential PAttern Mining using A Bitmap Representation, followed a bitmap representation for the database. It then integrated support counting with the representation. For each item present on scan, bit is represented as 1 in the representation. The method begins with S-step and progress to I-step. During the DFS pattern of traversal, extended sequences are generated during the S-step. I-step phase deals with addition of item sets and check on corresponding bit map for sequence generation. Pruning is done at the S-state and I-state tree itself to minimize candidate patterns. SPAM is efficient with long patterns in databases [Da Ruan et al. 2005].

Jian Pei et al. (2004), adopted a pattern growth approach in their sequential pattern matching algorithm Prefix Span. Prefix span is found to be efficient for large data sets but suffered from the difficulty of the FP tree not fitting in to available memory.
Ben Kao et al. (2005), considered the factor of incremental updates on frequent sequences to generate maximal frequent patterns. They observed the database updations over time. MFS works well with long sequences but this methodology which is an incremental update to MFS effectively reduces the I/O cost than MFS. MFS increments gives better results on initial scans but this approach is found to be effective after each incremental refinement.

Zhenglu Yang (2008) in his research work proposed a novel idea of Last Position Induction (LAPIN) for sequence generation. In this algorithm the last position determines whether a k+1 sequence can be generated. The algorithm effectively keeps track of access patterns based on web logs. The algorithm performs well with large dense data sets.

Ezeife & Yi Lu (2005) proposed a methodology based on WAP tree to extract sequential patterns. Web Access Pattern Tree is a tree structure representation similar to Frequent Pattern Tree. It generate sequences by recursive traversal over nodes. Numerous intermediate trees are generated during this traversal. Their methodology proposes a position code approach which avoids the intermediate tree generation process. Their proposed suggestion PLWAP Tree does the pre order linking of nodes starting with the first prefix and efficiently eliminates backward traversals. To find the descendants of a particular node position codes can be used. The codes also help in finding the nodes of current tree which has relation to sub tree.

Charu C Aggarwal et al. (2009), observed the behavioural changes of Apriori and pattern growth based algorithms when uncertainty factors are considered. The algorithms considered produced unexpected patterns when dealt with uncertain items. They had shown degraded performance in connection with memory usage and efficiency. The authors suggested the
need of probabilistic approach for pattern generation in real world situations. They proposed this as extensions to such algorithms.

Willie Ng & Manoranjan Dash (2010) proposed a methodology for extracting frequently occurring sequences from streaming data. Their proposed algorithm DSS, selects the sample sequence based on histogram data from continuous streams. Selection is from single data items. DSS also takes into account the noise factors in streams of data.

Toon Calders et al. (2010) studied the possibility of occurrence of uncertain data based on statistical methods. They converted the binary mode of representation to probability based representation. This represents the uncertain data model. They considered the problem of over/under estimations. In cases where support values close to minimum threshold values over estimations results in getting high values which may be wrong. They introduced precision and recall to ensure accuracy over the possibility of real support being under estimated or over estimated.

Usef Faghihi et al. (2011) proposed an emotional considerate learning system which enhances learning during training sessions. After creating models for user groups, the system can predict user behaviour. This web based system facilitates better learning environment to user by user centric performance evaluation. Data is extracted through by employing a mixture of data mining algorithms. Our method enhances learning by automatic generation of user preferred topic on login itself.

Philippe Fournier-Viger et al. (2011) suggested a rule growth method for mining rules for multiple sequences. The method is based on the pattern growth approach. The algorithm follows a recursive growing method.
First it finds the rules between two single items. After each scan the method keeps track of left and right parts of the rule recursively.

Zhou Zhao (2012) proposed a method to find patterns from uncertain sequences based on the pattern growth approach. They tried to extract frequent sequences with high probability from uncertain databases. This work was based on the pattern growth method. They developed two models, one a sequence level and the other an element level by incorporating uncertainty. The idea is to mine probabilistically frequent sequential patterns. Their proposed algorithm U-Prefix span reduces possible word explosions in pattern generation.

Xiang Lian & Lei Chen (2013), put forward a query named top-k dominating query to retrieve k number of uncertain items from uncertain databases. The advantage of this mechanism is that it scales effectively in multi dimensions, without giving the user the overhead of selecting a ranking function. Based on the size of the application the query selects ranking function.

Yuxuan Li et al. (2013), studied the pattern generation of uncertain transaction databases. They introduced the idea of gap constraints to probabilistic sequences. They adopted a dynamic programming methodology for frequently occurring pattern identification. They then integrated this in to pattern enumeration algorithms by following breadth first search and depth first search strategies. This method effectively generates spatio-temporal frequent sequences when data considered has uncertainty factor.

Philippe Fournier-Viger & Vincent S Tseng (2013), proposed a method named TNS for mining top ‘k’ non redundant sequential rule which overcomes problems with threshold values in sequence mining. Here k is the
number of rules to be determined. The value of k is user defined. This algorithm follows a rule expansion method. It combines the non redundant rules with sequential rules to get the non redundant sequential rules.

Lei Duan et al. (2015) considered the unusual aspects termed as outlying aspects on numeric data. Based on the studies in outlier analysis [Aggarwal (2013)], they tried to extract the unusual behaviour of objects compared to available data. They compared the aspect of outlyingness in sub spaces. They ranked this on a probability based approach. Their study analyzed both real and synthetic data sets. The method is found to be more efficient in finding outlying aspect in multidimensional numeric data sets.

2.5 SUMMARY

This chapter briefed the related survey of literature on Link Analysis, Structure representation methods and various rules for association and sequence generation.

By considering various factors such as uncertainty of items in database, large database, partially ordered sequences, click rates, relationship between weight of a node and user traversal and in link out link factors it is observed that some methodologies are too complex. Some methods need more iteration to get result.

Based on the observations we came up with a better personalized model which is more user friendly. Our proposed model does automatic re arrangement of the page based on usage log data which in turn is the user preference. We elaborate our study in the following three chapters.