CHAPTER 1

INTRODUCTION

1.1 Background

We are living in an information age. Information is persistently produced from various sources and with different structures. Processing of this enormous information is possible with the assistance of an efficient and intelligent system. Organizations can't depend just on their administrations to grow; they need to utilize verifiable information for better reason. Data analysis is especially needed for the development of society and business. Data mining strategies are used to discover this helpful data. Utilizations of such procedures and techniques can give organizations productive administration, important bit of knowledge and consumer loyalty. Data and Information mining plays a vital part in data industry because of wide accessibility of gigantic information in electronic forms. There is need of using such information into valuable learning for the business administration. This profitable data is in various structures like designs, patterns, correlations, noteworthy structures and so forth.

Data mining tasks are ordered into descriptive information mining and predictive information mining. Descriptive mining gives information in concise and succinct manner but predictive mining attempts to predict the behaviour of a new data set from the available information [1]. Sequential pattern mining is a standout amongst predictive data mining with an extensive variety of applications which discovers the frequent sub-sequences from a sequential database. Sequential pattern mining includes finding frequent itemsets from large sequence dataset. Associations framed from frequent sub-sequences are called Association Rules (ARs) whereas; Sequential rules are called as an episode rule, temporal rule or prediction rule.

Association Rule Mining (AR_MINE) is a compelling method of analysing and finding strong correlations among different successive patterns in an extensive dataset. AR_MINE is prominent in presentation and comprehends capacity of information from the database. The problem of Association Rule Mining is firstly applied for market basket analysis [2, 3].The association analysis for market basket is to find what client doesn’t buy and why? This gives the details about characteristics of the client. Agrawal, Imelinki and Swami have proposed a ‘Support-Confidence’ structure for the revelation of association rule [4]. This system is widely accepted.
Mining association rules problem is fragmented into two sub-issues as discovering all successive frequent itemsets and forming rules from frequent itemsets. However, the downside of AR_MINE algorithms is that, it discovers a large set of rules from a given dataset, from which most of the rules are repetitive and uninteresting to clients. Hence it is needed to eliminate certain undesirable rules. This system which gives great heuristic to the clients with interest as constraint to confine search space is called as constraint based mining. Constraint based mining empowers interactive exploratory mining and analysis. Used constraints are of type as 'data', 'dimension','knowledge – type' and 'rule_constraint' [5].

1.2 Literature Review

Ample research has been done on sequential pattern mining and framing associations in the form of rules. This section discusses the work done on different approaches of pattern mining techniques and improving performance with the addition of various measures and constraints.

One of the frustrated MIS executive’s statement as “Computers have promised us a fountain of wisdom but delivered flood of data”. Hence a technique has been specified in the literature for knowledge discovery. It is the process of nontrivial extraction of implicit, previously unknown and potentially useful information from the dataset. This knowledge is interesting according to a user imposed interest measure and criteria. The output of a program in the form of patterns that monitors the set of facts in a database is called discovered knowledge. The paper discusses in detail on knowledge discovery. Discussions are given as novelty and utility is not enough to qualify the patterns but this discovery technique must be implemented efficiently. Detail explanation of a discovery algorithm is specified in the literature. Paper further discusses on pattern mining algorithm techniques designed to extract knowledge from data. This involves two processes as identifying interesting patterns and describing them in a concise and meaningful manner. These algorithms must deal with the issue of computational complexity. Computational requirement of these algorithms grows faster than a small polynomial in number of records and are also inefficient for the large database. For large quantity of data, empirical methods are not useful. Data sampling is one of the solutions suggested in the paper to solve the problem of larger datasets. Incorporation of domain knowledge can improve efficiency of the discovery process. [6, 87]

Many applications are suggested in the literature such as in medicine for genetic sequence analysis, prediction of the hospital cost containment, in finance for bankrupting
prediction, stock market prediction, securities fraud detection, mutual fund selection, in Agriculture for the classification of plants, in social for voting trends, election results, in marketing and sales for identification of unusual behaviour of socioeconomic subgroup, product analysis, sales prediction, in Insurance for detection of fraudulent and excessive claims, in military for intelligence analysis, in space science, search for extraterrestrial intelligence astronomy, space data analysis etc. [6]. The work related with Association Rule Mining algorithms with use of different measures is given as follows-

Strength of association rules in the form of confidence, support and expected predictability is specified in the paper [7]. Expected predictability is the difference between expected predictability and actual predictability. An association model has been proposed for the refinement of association rules. An algorithm is proposed in the literature which produces set of useful rules for prediction by evaluating importance of attributes. Importance of columns or attribute is given on entropy or the amount of information provided by an attribute. Main feature of this approach is the generation of an unexpected pattern by taking a single attribute per iteration. Refinement process for selected patterns is done with the use of knowledge of sets of attribute for classification [7].

Another literature has talked about the points of interest and drawbacks of target and subjective measures of Association Rules. Author has done the comparative analysis of dynamic association rules without life cycle, with life cycle and for weighted dynamic association rules. It discusses on objective and subjective measures. Objective measure considers statistical characteristics of an objective data as support, confidence, lift etc. While subjective measure contain knowledge field, other personality characteristics of users. A comparative analysis of objective measures are carried out which states that support can filter out most of the negative association rules but more strong rules are explored with support-confidence system. However this system cannot distinguish in between positive and negative association. With lift measure, rules which are with high frequent items are easily filtered out and validity measure reduces some of the redundant rules. The rule with higher improvement is considered as more interesting even if validity of the rule is same [97,101].

Traditional association rules with the change of support and confidence value at different time are analysed. Those rules are filtered out which are not present consistently with change in time and appear in the later stage. In traditional evaluation system, rules are considered as strong which remain consistent with higher support and confidence. A concept of dynamic association rules is presented with less support most of the time but at certain
time period its support and confidence value is high. Paper focuses on weighted dynamic association rules where each item is represented with life circle or time period. Analysis of an association rule with life circle shows that an association rule has a higher recognition [8].

Two new algorithms Confabulation-inspired Association Rule Mining (CARM) [9] and Binary Based Technique-An Efficient Association Rules Mining Algorithm (BBT) [10] are developed for associations. CARM algorithm requires only a single pass through the dataset and which utilizes cogency inspired measure for creating association rules while BBT algorithm requires only one scan of the original database, and it uses binary data representation. BBT uses the bit mask for frequent pattern generation. Both CARM and BBT algorithms are used to find association rules from the distinctive datasets like student dataset, customer transaction dataset, and medicinal dataset etc. CARM uses cogency inspired measure and mines association rules with only one pass through the file [9]. It uses a matrix, whose every element is represented by a number which strengthens the link between items. Using the link strengths of an item in the matrix we can find frequent itemsets in the datasets. BBT uses binary representation of data and overcomes the drawback of candidate generation [10]. A bit mask is set for every itemset. Logical AND operation is performed between bit mask and binary map of itemset. On the off chance that binary map is equivalent to its bit mask then the frequency of itemset is incremented.

Three rule generation algorithms are elaborated in reference [11]. It explores the concept of computational intelligence for finding relation between different diseases and patient. Three rule generation techniques are explained as ‘Apriori’, ‘Predictive Apriori’ and ‘Tertius’. In Apriori, rule discovery is done in a computationally feasible time. Accuracy of rule is measured by confidence of rule. Ranking of rule is done on the basis of confidence. Predictive Apriori technique is mainly used for the task of classification. It uses Bayesian framework to calculate predictive accuracy from support and confidence of the rule. Algorithm uses two measures such as expected probability and observed probability. UCI heart diseases dataset containing with different values such as ‘age’, ‘sex’, ‘chestpain_type’, ‘trestbps’, ‘cholesterol’, ‘fbs’, ‘sicktype’ etc. are experimented. Paper demonstrates a very purpose of rule mining algorithm to find interesting knowledge and mainly focuses on the application of computational intelligence [11].

The variation of Apriori using hash function for finding length-2 itemset, length-3 itemset and direct search method for large \( k \) itemset are composed with variation of ‘Eclat’ algorithm. Paper has elaborated a problem of Apriori algorithm for generation of large
candidate set. To reduce the number of candidate, ‘DHP’ algorithm with hash function is proposed. Algorithm like ‘PHS’ and ‘MPIP’ uses a perfect hash function. Another method has been specified to improve an efficiency of Apriori with the use of hash tree. Paper compares these techniques and proposes for less number of candidate sets. Paper has implemented ‘AprioriHashTree’, ‘MPIPTwoThreeHash’, ‘Apriori_DirectSearchHash’, ‘Éclat’ and ‘EclatHash’. Paper concludes that Apriori algorithm of direct search works better than Apriori with the tree [12].

A ’Δ – closed’ criterion to compact the resulting rule set is proposed. It is an extension of closed and maximal frequent patterns. A concept of ’Δ – closed’ sequential mining is introduced. A sequence ‘S’ is ’Δ – closed’ if and only if no super sequence ‘S1’ exists such that ’Sup(S) ≤ Sup(S1) + Δ’. Proposed ’Δ – closed’ sequence algorithm initially assigns ’Δ = 1’. Algorithm removes subsequence which offer a support less than or equal to ’Δ’. It starts with finding sequence with lowest support. Recommender was designed for predicting user’s first add-to-basket action this is called leading events. This process is divided into three different stages as mirroring the design or sequence pattern mining stage, the rule construction stage and rule application stage. In sequence pattern mining stage, sequences which are ending with an add-to-basket event is mined by sequence miner. Rule construction stage is formed by continuing each sequence into sequence rule in the form of < r1 → r2 > where ’r1’ subset of leading events and ’r2’ are add-to-basket event. The rules are considered valid only if rule’s confidence is more than minimum confidence. Product ranking is given for a leady event sequence in rule application stage. Different weights of rule such as lift, relative support, confidence decides top-K product. Rules are used to generate recommendation set and these are matched against fixed number of leading events. Paper evaluates recall, precision and harmonic mean of recommended system. Lift signifies successful recommendation. Recall is defined as proportion of a sequence in which recommendation set contained the hidden add-to-basket event while precision is proportional to size of recommendation set as ’k’. It is given by [recall = \( \frac{\sum_{\text{hits}}}{\text{no of seq}} \)] and [Precision = \( \frac{\sum_{\text{hitsno of seq}*k}}{} \)]. Experimental study of literature shows that with the use of lift metric, rules can boost recall to ‘28%’ of the actual purchase transaction for short navigational sequences [13].

Another efficient data structure is used with Enumeration Table. Frequent Closed Enumeration Table (FCET) is used to store the relevant information. Two new mining algorithms are proposed as Generating Mining Association Rules (GMAR) and Generating
Mining Frequent Itemset (GMFI). ‘FCET’ is used to reduce the memory needed for intermediate computation and also to get support count. Table is formed by taking union of all ‘TID’ sets for each item in maximal itemset. Let ‘A’ is in {3,4} and {6}, ‘C’ is in {4,5,6} and ‘D’ is in {2,7,8} and ‘ACD’ is maximal itemset then, it’s ‘TID’ is formed by union of {2,3,4,5,6,7,8}. Author has proposed an algorithm which differs from Apriori as all k-itemsets are searched based on the association graph. In ‘GMAR’ algorithm six different pruning techniques are used these are specified as - if support of rule is less than ‘MinConf’ then it is pruned, if $P \rightarrow I - P >$ condition holds then any of rules also holds condition of $Superset(P) \rightarrow I - Superset(P) >$. Thirdly if $P \rightarrow Q1 > 0$ holds and $P \rightarrow Q2 >$ doesn’t hold then a rule with an itemset ‘I’ = $(P \cup Q1 \cup Q2)$ doesn’t hold. Fourthly if $P \rightarrow Q1 >$ holds then $P \rightarrow Q2 >$ always holds where ‘Q2’ is derived from ‘Q1’. Fifthly if $P \rightarrow Q1 >$ hold and $P \rightarrow Q2 >$ doesn’t hold then $P \rightarrow Q1 \cup Q2 >$ doesn’t hold. Sixth technique is if ‘I’ is not frequent then $P \rightarrow I - P >$ doesn’t hold. With ‘GMAR’ join method is used to get rules with more length and uses different pruning techniques to ignore irrelevant rules. This expression shows that ‘GMAR’ and ‘GMFI’ performs better than that of the Basic candidate generation algorithm. Time complexity of finding maximal itemset for GMFI is $O(log_2 n)$ where ‘n’ specifies total number of maximal itemset [14].

A new approach of classification with association rule mining is given with ‘CPAR’. This approach combines the advantage of association classification and rule based classification. It uses a greedy algorithm approach to generate rules. ‘CPAR’ generates smaller set of high quality predictive rules, avoids redundant rules. For prediction, it uses the best ‘$k$’ rules. To avoid repeated calculation of rule, the algorithm uses dynamic programming and selects all close-to the best literals for generation of rules. Paper proposes a novel approach of ‘CPAR’ is proposed to overcome the problem of more time consumption by ‘CBA’ and ‘CMAR’ algorithm for larger datasets. ‘CPAR’ is based on First Order Inductive Learner (FOIL). Author proposes FOIL algorithm in detail which uses ‘Foil_Gain’ measure to find the information gain from the rules. Its runtime is given as $O(n * k * m|R|)$ where ‘$k$’ gives number of attributes, each attribute having m values on average and ‘n’ are number of transactions. |$R$| is generated number of rules. ‘CPAR’ stands in between exhaustive and greedy algorithm. It builds rules by adding literals one by one similar to Predictive Rule mining. ‘CPAR’ chooses all close-to the best literals in rule generation. With this technique algorithm builds several rules with one literal at the same time. For the evaluation of prediction using rules paper uses Laplace expected error estimate. It is given by Laplace
Accuracy as 

\[ \frac{\binom{Nc+1}{N_{tot}+k}}{} \] where 'k' be the number of classes, < N_{tot}1 > be the total number of examples of rules body, < Nc > be the example belong to class 'c' and 'c' is predicted class of rule. If a rule satisfies the highest expected accuracy in the form of Laplace accuracy then it is considered as more predictive power. These rules satisfying the highest expected accuracy are called best 'k' rules [15].

Another literature does the detail analysis of association rule. For the analysis of association rules, author has proposed two measures in the form of confidence width and blocked rules. Paper focuses on redundancy among association rules and representative rule. Redundant association rules are defined as those rules like < X \rightarrow Y > and < X \rightarrow XY' > are mutually redundant as confidence and support of both of the rule is same. According to author, representative rules form a unique smallest basis. These are considered as starting point as they give best basis size. One measure is presented as confidence width, it is given in the form of

\[ W(X \rightarrow Y) = \frac{C(X \rightarrow Y)}{C(X' \rightarrow Y')} \] for a rule < X \rightarrow Y > where < X \rightarrow Y > is equivalent to < X' \rightarrow Y' > and < Confidence(X' \rightarrow Y') \leq Confidence(X \rightarrow Y) >.

It is a highest confidence value of representative rule to the highest confidence value of different representative rule which makes a rule redundant. Confidence measure is rephrased in other way as 'squint', Confidence threshold which is associated to parameter 'squint' through intuition which clearly differs from [Support(X), MinSup(X)] and [Support(XY), MaxSup(XY)] etc. Paper proposes a concept of blocking rules with the help of confidence width and squid intuition parameter. Let a rule < X \rightarrow Y > and assume 'X\land Y = \emptyset' and 'Z \in X' blocks < X \rightarrow Y > at squint 'q' if Sup(XY) - Conf(Z \rightarrow Y)Sup(X)) / (Conf(Z \rightarrow Y)Sup(X)) \leq q where 'q > 0' and [Sup(X) - Sup(XY)/Sup(X)] > q] for a rule < X \rightarrow Y >. In this way paper states that redundancy caused due to larger consequence is better handled by confidence width and the problem of smaller ancestors leads to blocking of rule [16].

Another measure of confidence in the form of novelty is proposed in the paper [17]. It is used as an indicator for potential radical change or loss of system functionalities. A novel approach for generation of association rules with novelty with the use of mutate examiner is elaborated. The proposed algorithm compares each rule with a testing example. Each attribute of left hand side and right hand side of a rule is compared with an attribute value of testing example. If it does not satisfy the attribute value then the rule is rejected. Rejection value for
testing example is computed by \[\text{rej} = 1 - \sum_{i=1}^{CR} \frac{Vi}{CR}\] where 'CR' is number of rules; 'rej' is rejection value and 'Vi' attribute value of testing example. Paper does empirical study and shows that this method works well for the novelty detection which is based on association rule [17].

Paper [18] derives the knowledge in the form of association rule by measuring semantic distance between antecedent and consequence with user oriented novelty measure. This method performs well for all major objective interestingness measure and a novelty measure 'WordNet' for discovery of more strong association rules. It uses POCA which refers to Probability of Co-occurrence Analysis. With POCA 'X' is parent of 'Y' if \(P(X|Y) > P(Y|X)\) and \(P(X|Y) \geq N\) where '0 < N ≤ 1'.

Different novelty measures are defined as-

1. **User-Oriented Novelty measure** - It is defined as the distance between 'A' and 'C' of rule. This distance is defined as follows-

   Let rule \((A, B) \rightarrow (C, D)\) where its novelty is calculated by- \(D(A, C)D(B, C)D(A, D)D(B, D)\) where \(D(A, C)\) specifies semantic distance between 'A' and 'C'.

2. **Occurrence Distance** - This measure how distinct the occurrences of two keywords are. Larger the distinction in occurrence indicates less strength of association.

   Let for \(< X \rightarrow Y >\) rule occurrence distance given by-

   \[
   [D(X,Y) = \frac{P(X\cup Y) - P(X,Y)}{1 - \frac{P(X,Y)}{P(X\cup Y)}}]
   \]

   Where \(P(X,Y)\) is probability of 'X' and 'Y'

   \(P(X \cup Y) - P(X,Y)\) Specifies distinction between occurrence of 'X' and 'Y'

   \(P(X \cup Y)\) is the probability of joint occurrence of 'X' and 'Y'.

3. **Connection Distance** - It measures the strength of connection between two keywords in concept hierarchy. It is given in the form of length of path connecting 'X' and 'Y'. If path is larger, connection is weaker. Paper evaluates user oriented novelty measure with two perspectives as novelty prediction accuracy and usefulness indication power. Paper analyzes that user-oriented novelty has a high correlation with subjective rule novelty and with these measure more strong, user interesting patterns are discovered [18, 96].

   An extended sequential pattern mining algorithm is proposed to extract a problem
space that is more adapted for supporting tutor services. Author has illustrated Roman Tutor system. It is used to teach astronauts for operating a seven degree of freedom robotic arm deployed on International space Station Dataset. Extension of PrefixSpan algorithm is used for the problem of Generalized Sequence Pattern Mining with Time Interval (GSPM). To mine a valued sequence database as a time-extended sequence, Action/pair counting is modified to note the values of action being counted and their sequence id’s. To find the clusters, ‘$K$ – means’ algorithm is used. Third limitation encountered by them with PrefixSpan algorithm is that it doesn’t consider context of each sequence. Several limitation of Sequential Pattern Mining for extracting problem spaces is given in the paper. An extended Sequential Pattern Mining algorithm is proposed which contains time interval, multidimensional pattern mining and automatic clustering of valued actions. Dividing problems into sub-problem, states to enhance the relevance of tutor services [19].

With pushing constraints in the mining process improves an efficiency of rule mining algorithms is described with the work related to Constraint based Rule mining as-

A new algorithm is described which exploits all user specified constraints. It offers a predictive advantage of dense and large dataset. Proposed algorithm allows user to eliminate unnecessary complex rules with the specification of minimum improvement constraint. Main idea of algorithm is to mine rules whose confidence is at least $' \geq MinImp' \text{ and any of its simplification where simplification is formed by removing one or more confidence from its antecedent.}$ Positive setting of ‘$MinImp$’ would prevent unnecessarily complex rules. This feature solves rule explosion problem. The algorithm leverages consequence constraint through pruning function for enforcing confidence, support and improvement constraint during mining phase.

An algorithm Dense-Miner [20] generates rule constraint to efficiently mine consequence constrained rules from large and dense dataset. The search space pruning strategy for Dense-Miner has three components-

1. It gives function which allows algorithm to compute bounds on confidence, improvement and support.
2. Main approach is used for recursive support information which is gathered during previous database pass.
3. It ensures that, plenty of pruning opportunities are available [20].

‘TRuleGrowth’ algorithm an extension of ‘RuleGrowth’ Algorithm [22] is proposed to resolve the problems occurred by general mining sequence rules algorithm. The more
concentration is given to the generation of valid, interesting, non redundant and more predictive rules. A sliding window constraint is pushed to find rules occurring within maximum amount of time. The algorithm is applied on many real applications which wish to discover patterns occurring frequently or most of the time [21].

‘RuleGrowth’ algorithm used in the paper avoids the problem of candidate generation as it uses pattern growth approach. It first finds rules of size $1 \times 1$ and then recursively grows by scanning sequence containing them to find single items which can further expand to left or right side. This expansion is specified by left expansion and right expansion method in an algorithm. Left expansion is the process of adding an item $I'$ to left side of rule $<x \rightarrow y>$ to obtain larger rule where as $<x \cup \{I\} \rightarrow RightExpansion>$ is defined as process of adding an item $I'$ to right side of a rule $<x \rightarrow y>$ to obtain larger rule as $<x \rightarrow y \cup \{I\}$. An algorithm also searches subsequence to form rule with depth-first search method.

Constraints are added to an algorithm as ‘RuleGrowth’ generates rule one item at a time. ‘TRuleGrowth’ adds constraint of sliding window to ‘RuleGrowth’. This is very useful to many real life applications containing temporal patterns for analyzing sensor network or stock market data. With the addition of ‘sliding_window_constraint’, efficiency of an association rule mining algorithm increases as it decrease the execution time by pruning the search space. It also produces smaller set of rules, this reduces disk space requirement for strong rules and also easier to analyze results. Efficiency of an algorithm is improved with addition of ‘sliding_window_constraint’. TRuleGrowth algorithm increases prediction accuracy [22].

Two main changes are done to ‘RuleGrowth’ algorithm to form ‘TRuleGrowth’. First change is that instead of keeping first and last occurrence of each item for each sequence all occurrences are now kept for each sequence. And second change is related to sliding window constraint. Here before generation of rule it is checked that an item is within specified window size. It is checked as - if there exists occurrence of ‘$x’ of ‘$l’ and occurrence of ‘$y’ of ‘$j’ such that $y - x > 0$ and $< y - x + 1 \leq windowsize >$ then it considers that ‘$l’ occurs before ‘$j’. This modification for larger rules is done at ExpandLeft and ExpandRight procedures [23].

‘CMRules’ algorithm has been proposed in three different steps. In the first step of an algorithm it takes input as any sequential transaction database,'MinSeqSup' and 'MinSeqConf’. It ignores all temporal information and convert sequential database into corresponding transactional sequential dataset. In the second step, it uses any AR_MINE
algorithm such as ‘Apriori’ to find frequent itemsets and discover association rules with 
'\text{MinSup} = \text{MinSeqSup}' and '\text{MinConf} = \text{MinSeqConf}'. In the third step it eliminates 
those rules which don’t satisfy the constraint of minimum threshold. According to given 
time complexity and measures its performance with click-stream data from the logs of online 
news portal. An application of algorithm is presented in an intelligent tutor agent [24, 98].

Use of negative association rules is elaborated in the literature as- Unexpected patterns 
and exceptional patterns are called exceptions of rule and also called as surprising patterns. A 
nice example is given as- \(<\text{bird}(x) \rightarrow \text{flies}(x)\>\) is a positive rule but an exceptional rule 
is \(<\text{bird}(x), \text{penguin}(x) \rightarrow \sim \text{flies}(x)\>\). These exceptions are called as negative rules. Negative 
Association Rules (NARs) plays an important role in decision making. Negative 
Association Rule Mining potentially assists for automated prediction of trends and 
behaviours. The paper has introduced the good example of market surveillance team. 
Generally for market surveillance, efficient trading environment is given for all participants 
through an alert system. But NARs assist to determine which alerts can be ignored. Paper 
says, an efficiency of an algorithm is improved by pruning strategy with Piatetsky-Shapiro’s 
measure of interestingness. Increasing degree of conditional probability is relative to prior 
probability used to estimate the confidence of positive and negative association rules [102].

An algorithm is presented in paper [26] which extends support confidence framework 
with a sliding correlation coefficient threshold. Proposed negative association rule algorithm 
differs from the previously described algorithm as it uses correlation coefficient as measure 
of interestingness. Author has computed correlation coefficient for every pair \(<X, Y\>\) of an 
item ‘\(i\)’ where \([i = X \cup Y]\). Substitution rule mining algorithm is presented in the literature. 
This discovers concrete items with high Chi-Square value and more expected support. When 
itemsets are found, correlation coefficient is computed for each pair. From these pairs 
negative correlation is generated and desired interesting negative rules are extracted [26].

A Global Negative Attribute Oriented Induction (GNAOI) approach is proposed 
which generates comprehensive and multiple level negatively generated knowledge. To 
measure the strength of negative correlation a new measure is introduced called 
'\text{nim}(cl)\'. The algorithm combines attribute values with their concept hierarchies to induce 
negative generated knowledge. To speed up the computation, it stores Generalized Identifier 
(GID) code of corresponding attribute value and cover list in an information encoded table. 
To improve the efficiency of an algorithm, downward level closure and upward superset
closure properties are employed. Interesting negative generalized tuples are found with the use of three constraints. The first constraint specifies 'MinSup' and expected support constraint i.e. the negative generalized tuple must satisfy minimum expected support constraint. The second constraint gives strong negative association. Two items are having strong negative association if probability of their co-occurrence is significantly smaller than multiplication of their individual probabilities. Third constraint is of redundant measure constraint. This tries to reduce such tuples which are induced from other tuples [27].

Another technique for ordering items is given with an algorithm which computes ranking of items on ascending order of support value. Only those items with higher rank are considered. To reduce the computation time required by MINIT, the dataset is pre-processed by building linked list of TIDs for each item. All items are arranged in ascending order by Support. Paper has given the explanation of computational complexity of minimal infrequent occurring itemsets. This problem is considered as an instance of Hitting Set problem. Paper gives the detail explanation of minimal infrequent itemsets problem is NP-Complete [28].

Weighted mining is done with a new measure ‘IWI-Support’. It is defined to analyse the data. Two different IWI-support measures as ‘IWI-Support-Min’ and ‘IWI-Support-Max’ are defined which relies on least interesting and maximum cost function respectively. Two new algorithms are introduced in the paper. One is Infrequent Weighted Itemset Mining (IWI) miner and Minimal Infrequent Weighted Itemset Miner (MIWI) miner. It works like FP-growth like algorithm [99]. The main feature of algorithm is pruning of FP-tree node which is done by IWI-Support constraint. Early stopping of recursive FP tree search in MIWI Miner is used to avoid extraction of non minimal IWIs [29].

A systematic study on constraint based sequential pattern mining is conducted in literature [30]. The main focus is given to user interesting sequential patterns. A very good example is stated in the literature for characterization of a new discovery where a researcher want to find sequential patterns about symptoms as “Finding patterns with constraint of 2 – 7 days of cough followed by fever in the range of '37.5 – 39°C' for 2 – 5 days with average temperature of '38 ± 0.2°C' and all the symptoms appear within two weeks”. This type of mining query contains few constraints. Firstly author has given a classification of constraint based on their application and their role in Sequential Pattern Mining. Secondly a new framework for prefix-monotone property for the constraints like regular expression is specified.

An idea of mining sequence pattern with Regular expression is given in the paper
It uses an idea of recursive mining where if a prefix itself is a pattern which satisfies the user interesting constraint then it should be outputted. Further this prefix satisfying constraint should be grown and mined recursively. The process ends when there is no local frequent item or no legal prefix. Experimental study of paper shows that although prefix monotone property is weaker than Apriori but it still achieves better performance than Apriori based method. Paper explores the pushing method of aggregate constraint in pattern growth approach. An item is considered as a small item if its value is \( \leq v \), where \( v' \) is the user specified threshold. Otherwise it is called a big item. In the first scan of projected database, unpromising big items are removed. In \( \alpha' \) projected database, when a pattern \( \beta' \) is found following \( \alpha' \) as small item, first it is checked whether small item can be replaced by a big item \( \chi' \) and if still finds average value of satisfying constraint, then prefix \( <\alpha, x> \) is marked as promising and not checked again. If \( <\alpha, x> \) violates the constraint then projected database are pruned. This is called unpromising pattern pruning rule. Paper states that pattern growth technique is used to handle tough aggregate constraint without prefix-monotone property [30].

Sequential pattern mining algorithm’s unfocused approach is presented in paper [3]. Two major drawbacks of pattern mining algorithms are given as firstly disproportionate computation cost for selective users and overwhelming volume of potentially useless results. To overcome this problem, an algorithm is proposed which incorporates user controlled focus in the mining process.

The problem of mining sequential patterns with Regular expression constraint is presented and constraint is pushed in the pattern mining process. Algorithm exploits equivalence of regular expression to deterministic finite state automata. The experimental study in a paper shows that including regular expression into pattern mining computation, yields more improvement in performance. Algorithm ‘SPIRIT’ [31] works in different passes. Each pass results in discovery of longer patterns. In \( k^{th} \) pass, a set of candidate \( k' \) sequence \( C_k' \) is generated and pruning technique uses information from earlier pass. Support for each candidate sequence \( C_k' \) is counted with the scan of whole dataset and \( F_k' \) is formed for frequent sequence in\( C_k' \). At the end of the pass the result is \( F_k' \) which is a set of all frequent \( k' \) sequence that satisfies constraint \( C' \). SPIRIT framework handles pruning in two ways as constraint based pruning and support based pruning. In constraint based pruning appropriate constraints are pushed in candidate generation phases and in support based pruning it checks sub-sequence with satisfying minimum support constraint.
Four SPIRIT algorithms points spanning the entire spectrum of relaxation for user specified Regular expression are given as follows-

[SPIRIT(N) – N] It employs weakest relaxation of regular expression. It prunes only candidate sequence containing elements that don’t appear in Regular expression.

[SPIRIT(L) – L] It requires every candidate sequence to be legal with respect to some state of automata.

[SPIRIT(V) – V] It filters out candidate sequence which is not valid with respect to any state of automata.

[SPIRIT(R) – R] It pushes regular expression inside mining process by counting support only for valid candidates.

Experimental results of literature [31] ‘SPIRIT(V)’ provide consistently good performance over entire range of regular expression. It is faster than ‘SPIRIT(N)’ for some regular expression. For highly selective regular expressions, ‘SPIRIT(R)’ outperforms. The overhead of candidate generation for ‘SPIRIT(L)’ and ‘SPIRIT(V)’ is negligible. This results into pushing regular expression in mining process leads to improve the performance [31, 93].

Another framework given in the paper [32] focuses on discovering valuable, interesting knowledge from the large dataset. ‘CoGAR’ framework is presented in the paper. Two main constraints are specified as schema constraint and opportunistic confidence constraint. ‘CoGAR’ represents Constraint Generalized Association Rules framework. The framework carries mining process in two steps where the first step extracts frequent itemsets with support and schema constraint. In the second step rules are generated with confidence and opportunistic confidence. Opportunistic confidence constraint is based on confidence measure and it composes each rule with a subset of its descendants. This framework is also able to evaluate multiple hierarchies on same attribute. [R(r)desc] contain all descendants’ rules of ’r’ such that the body part of rule is same as ’r’ while antecedent or head part is descendent of the head of ’r’. ‘CoGAR’ framework gives CI-Miner algorithm for itemset mining. CI-Miner algorithm enforces schema constraint to prune uninteresting itemsets. Further RuleGen algorithm is presented for generation of rule. Paper gives experimentation on the effect of enforced constraint like schema and confidence opportunity which shows scalability of the mining process.

An algorithm has been proposed which computes consecutive repetition in sequence dataset by reducing the amount of data to process which speeds up the extraction time. A time window constraint and constraint of sequence length are added to reduce the size of
occurrence list. The strength of a constrained generalized occurrence list is used, which contains identifiers, timestamp of an occurrence of first event, interval of min-max and timestamp of last occurrence of last event of pattern. ‘GoSpec’ algorithm is proposed which is an instance of abstract algorithm with joint designed for generalized occurrence list. Experimentations are carried out with ‘cSpade’ algorithm. Experiments in the literature [33] show that with ‘GoSpec’ algorithm, the gain in terms of memory space and execution time is achieved [103].

New concepts of ‘recency’ and ‘compactness’ are introduced in the literature [34]. Recency causes patterns to adapt to latest behaviours in sequential database and compactness gives reasonable time spans for these patterns. These patterns are referred as CFR-patterns. Paper proposes ‘CFR-postfixSpan’ algorithm which is developed from PrefixSpan algorithm for finding frequent sequence patterns with additional two constraints as recency and $ms - length$. ‘RecencyMinSup’ for recent sequence database gives the most recently occurring subset of sequential database and compact constraint means the time span from first item to last item in a pattern must be no more than maximum span length or $ms - length$. Proposed ‘CFR-postfixSpan’ algorithm defines compact postfix, compact recent postfix, compact projection, compact prefix and projected database. Comparison of PrefixSpan and ‘CFR-postfixSpan’ algorithm is done. First difference with PrefixSpan is concerned with order of items in data sequence while CFR handles timestamp of each item in data sequence. Secondly CFR projects only frequent recent postfixes. Algorithm starts with finding all length-1 CFR patterns. For each item ‘$cf - Support$’ and ‘$cr - Support$’ value is calculated. In the second step it divides the search space. For each compact projected database respective ‘$sid$’ and ‘$Endtime$’ values are recorded and compactness constraint is satisfied. Compactness constraint is the difference between timestamp of earliest item and [Endtime $\leq ms - length$]. In the final step all CFR-patterns are found by constructing corresponding projected database and recursive mining. Experimental study shows that by adding ‘$r - MinSup$’ and ‘$ms - length$’ properly, many uninteresting patterns are pruned and CFR-postfixSpan performs well for long sequence database.

Another method proposes a user interest as constraint patterns. This deals with dependent items where attributes show group of items that have a common property. Paper introduces constraint patterns in order to extract valid sequence patterns. These patterns are further used to predict sub patterns. These constraint patterns are reformed as constraint item sequence. If length of constraint pattern is large, combination of attribute values increases
exponentially hence proposed method checks these constraint itemset step by step. The method first generates candidate itemset and finds whether it is frequent or not then repeats the process for candidate sequence patterns. Process is repeated till all frequent sequential patterns are discovered. A candidate item set with \((i+1)\) item is generated from two frequent items with 'i' item. These frequent itemsets have to satisfy the constraint item subset. In the next step \(k+1^{th}\) candidate sequence pattern is generated from two \(k^{th}\) frequent sequential patterns. It must satisfy constraint pattern. This decomposition of constraint patterns is done as per user specified constraint patterns. These patterns are decomposed into constraint sub patterns and constraint item subsets respectively. These decomposed constraint patterns are used for the generation of sequential pattern mining method. If candidate doesn’t satisfy constraint, it will not calculate its support. Effectiveness of this method is verified on the sequence data of stock price indexes [35].

Constraint based algorithms are given as - Sequential Pattern Mining algorithm with constraints has been proposed for discovering recurring structures in protein sequences. An algorithm with constraints as gap and regular expression are presented. Experimentation in the literature proves that with application of regular expression and gap constraint, more specific and relevant information is extracted. An algorithm is used for large protein database. During pushing gap constraints in sequential pattern mining, the transformation step is modified. In transformation step of an algorithm, it sets the bit position from \([1,2,3, . k]\) to zero and every bit after 'k' is set to one where 'k' is first index of one in vertical bitmap of an item. While ‘SPAM’ with gap constraints only the bits between positions 'p + mingap + 1’ to 'p + maxgap + 1’ be set to one and all other bits are set to zero where 'p' is any position with bit one in original bitmap section of a data item. For the regular expression constraint as it is not prefix-anti monotonic, use of relaxed constraint is proposed to prune the search space. The pruning strategy is given as, at each sequence-extension-step and item-extension-step, it prunes all nodes containing sequence that violates relaxed constraint. At the end of tree building process, regular expression constraint is pushed. An experimental study shows that with 'MinSup \(\geq 30\%\)' algorithm finds no much difference in runtime performance but with less 'MinSup' it shows more runtime benefit. Also gap_constraint reduce constraint coefficient and improves runtime performance for large input dataset. Experiments proved that output size is reduced with increase in 'mingap' from '0 to 5’. With the use of mining algorithm containing gap and regular expression constraint, paper has presented a three phase as pre-processing, feature extraction and feature
verification framework with software called Pex-Spam for feature extraction from protein sequences which is an ideal tool for bioinformatics problem. Experimentation proved that the whole data mining process of Sequential Pattern Mining is done within main memory which avoids I/O delay hence an algorithm is space inefficient but more suitable for extraction of features from protein sequence [36].

Discovering association rules are used for sales transaction data due to large progress in bar-code technology for storing and collecting large amount of sales data. Two algorithms as ‘Apriori’ and ‘Apriori-Tid’ [2] are proposed. Features of ‘Apriori’ and ‘Apriori-Tid’ are combined to form a hybrid algorithm called ‘Apriori-Hybrid’. ‘Apriori’ and ‘Apriori–Tid’ algorithm differs from previous AIS [1] and SETM [2] algorithm as it generates candidate itemsets from the itemsets found in previous pass. Apriori –Tid has additional feature that it doesn’t use whole dataset for counting the support of candidate itemset. If transaction doesn’t contain any candidate 𝑘 -itemset then its entry is not considered for set of candidate 𝑘 itemset with transaction id. This lowers the number of entries of set than that of the number of transaction in database and improves the performance of algorithm. Author’s experimental study shows that both ‘Apriori’ and ‘Apriori-Tid’ always outperforms AIS [1] and SETM [2]. By combining good features of both, ‘Apriori-Hybrid’ algorithm is presented which scales linearly with number of transaction. Time needed for an algorithm will decrease with decreasing the number of items in the database.

Downsides of association rules are specified as lack of user exploration, lack of focus and rigid notion of relationships. To overcome this problem, an architecture which supports constraint based human centred mining of associations is proposed. Constrained association queries are given for this. Suggestions are given in the form of user feedback, user guides system for finding intended candidates for antecedents and consequences, Adhoc mining of associations etc.

Architecture of exploratory association mining is presented which contain two phases–where user initially gives Constraint Association Query (CAQ), hence output of first phase is list of pair candidates for antecedent and consequences satisfying constraint. In the second phase user has to specify significance metric, threshold and candidates on antecedents and consequences. Hence second phase includes computation needed with user interestingness. Constraints are categorized into various types based on anti-monotonicity and succinctness property. To achieve more speedup, ‘CAP’ algorithm is presented which push constraints deep inside the algorithm. Constraints are classified into four different classes as constraints
that are anti-monotone and succinct, succinct but not anti-monotone, anti-monotone but not succinct and constraints that are neither. For succinct and anti-monotone constraint literature has proposed strategy as sets containing any element not satisfying succinct and anti-monotone constraint are not considered. For these constraints, the subsets satisfying succinct constraint that are disjoint with anti-monotone are excluded for anti-monotone but not for succinct. These sets not satisfying constraint which is anti-monotone but not succinct are excluded. For non succinct and non anti-monotone, weaker constraints are used which might be anti-monotone and or succinct. Later such constraints are found by exploiting one of the strategies as above [37].

A new ‘RI’ algorithm is propose in the literature [38] which reduces the search space for candidate rules with pruning irrelevant items during the process of building classifier. This algorithm is implemented in ‘WEKA’ (Waikato Environment for Knowledge Analysis). Problem associated with PRISM algorithm of induction is described which results into large size classifiers. This limits applicability of ‘PRISM’ as a tool for decision making; secondly ‘PRISM’ generates a rule only when its error is zero thirdly there is no search space reduction mechanism for candidate items. To overcome the drawbacks of ‘PRISM’ the paper has proposed enhanced Dynamic Rule Induction Algorithm (eDRI). The algorithm has two primary phases. One is rule production and second is class assignment of test data. This method produces a set of rules with zero error and which have large Rule Strength parameter. To minimize the search space, similar to support frequent parameter it removes items which are less frequent. In the second phase these rules are derived and used to decide the type of class. Algorithm first scans the dataset to find respective frequencies. It creates first rule by appending an item to current rule which gets best accuracy. Algorithm continues adding items to the current rule till it becomes hundred percent accurate or greater than ‘Rule _Strength’ threshold. Experimental study of an algorithm shows that classifier is derived by an algorithm of higher benefit and with high predictive accuracy. This is used as a rich knowledge base decision tool [38].

Constraint based pattern mining is elaborated with an extension of ‘GSP’ algorithm for dataset filtering. Constraints driven pattern discovery is classified into three groups as-

i. Post processing which removes or ignores pattern not satisfying user constraints after the discovery process.

ii. Candidate Filtering - Constraints are applied to reduce the number of processed candidate e.g. ‘SPIRIT’ algorithms.
iii. Dataset Filtering - This restricts the source dataset itself.

Classes of constraint are specified as database constraint, pattern constraint and time constraint. Database constraint is used to specify the source data, pattern constraint for specification of which are interesting patterns and time constraint for checking a data sequence contain given pattern or not. Time constraint is further given as 'max - gap', 'min - gap' and 'time - window'. Constraints with Boolean predicates are expressed on pattern or pattern elements as follows-

\[
\begin{align*}
J_I(SG, \alpha, \text{pattern}), J_I(LG, \alpha, \text{pattern}), J_I(C, \beta, \text{pattern}), J_I(G, \gamma, \text{pattern}), \sigma(SG, \alpha, \text{Seq}), \\
\sigma(CL, \alpha + 1, \text{Seq}, \text{max}, \text{min}, \text{win}), \sigma(C, \gamma, \text{pattern}), \sigma(C, \beta, \text{Seq}, +\infty, \text{min}, \text{win}), \sigma(C, < Y>, \text{Seq}, \text{max}, \text{min}, \text{win})
\end{align*}
\]

where \(< Y\>\) denote 1-element sequence and \(\text{max}, \text{min}, \text{win}\) as \(\text{max} - \text{gap}, \text{min} - \text{gap}, \text{window} - \text{size}\) respectively. These patterns are transformed into query. The relation between supports of pattern with filtered dataset is given as-

\[
\text{Sup}_F(p) = |D| * \text{Sup}(p) / |D_F|
\]

Where \(\text{Sup}_F(p)\)-Support of pattern ‘p’ in the filtered dataset. \(\text{Sup}(p)\)- Support of pattern ‘p’ in original dataset.

\(|D_F|\)- Number of data sequence in filtered dataset.
\(|D|\)- Number of data sequence in original dataset.

‘GSP-F’ algorithm is proposed by incorporating filters in ‘GSP’ algorithm.

As Generalised Sequential Pattern mining algorithm generates candidate sequence iteratively and computes support or occurrence of data sequence forms its source dataset. In ‘GSP-F’ filtering is done with iteration. Patterns not satisfying constraint are not included in candidate verification process. In the post processing step, all patterns are filtered which doesn’t satisfy user specified pattern constraint. Experimental study of the given research work states that lower the selectivity of dataset filtering constraint, better the performance of ‘GSP-F’ than ‘GSP’ [39]. Algorithm includes three points; (a) a capability to handle two kinds of item interval measurement, item gap and time interval, (b) a capability to handle extended sequences which are defined by inserting pseudo items based on the interval itemization function, and (c) adopting four item interval constraints. This proposed method is able to substitute all types of conventional sequential pattern mining algorithms with item intervals. Using Japanese earthquake data, paper has confirmed that algorithm is able to extract sequential patterns with item interval, defined in a flexible manner by the interval itemization function [40].

A language based on regular expression is proposed to restrict frequent sequence to user specified constraints. The language is referred as ‘RE-SPaM’ based on constraints over
atomic item. Expression in the language contains attributes, functions over attribute and variables. Moving objectdb (MOD) is used which includes trajectory aggregation in traffic analysis. Trajectory of object is given by \(< oid, t, x, y >\) where ‘t’ specifies time and \(< x, y >\) gives location co-ordinates of object ‘oid’. Constraints to these sequence patterns are given by places of interest (‘POI’). Example provided in the paper as tourist application in the city of Paris where ‘POI’ may be restaurant, hotel, tourist attraction etc. In ‘Re-SPaM’ constraints are enclosed in square brackets. These constraints also contain function over attributes such as rollup. Algorithm proceeds as - First user defines regular expression and ‘MinSup’. From the given constraint, DFA is built. Algorithm proceeds in incremental phases until it reaches to final condition which is dependent on choice of relaxation. In the main loop of algorithm all candidate sequence are generated and all paths of automaton is detected [41].

An algorithm ‘orderspan’[100] is proposed which is used to extract a set of closed partially ordered patterns from sequence database. It combines well known properties of prefixes and suffixes. Let ‘Sp’ be a sequential pattern, ‘Sp’ is closed sequential pattern if there is no other sequential pattern \(< Sp’ >\) such that length of ‘Sp’ is less than or equal to length of \(< Sp’ >\) and \(< Support(Sp) = Support(Sp’) >\). An algorithm uses pattern-growth approach which uses divide and conquer strategy. An algorithm is flexible and follows the sequential pattern paradigm. It is more efficient in the search space exploration as it skips redundant branches and outputs complete set of closed partially ordered patterns. It provides new data structures to remove redundancy in patterns. It also provides upper bound on the number of extracted patterns for a given minimum support.

Improvement in efficiency of AR_MINE algorithms is given in literature [43]. Paper gives overview of techniques that are used to improve the efficiency of the Association Rule Mining (AR_MINE) from huge databases. The efficiency of Association rule mining is improved by different methods like sampling, with reducing the number of passes, Hash-based itemset counting, transaction reduction, partitioning, addition of extra constraints, Association rule clustering system etc. Sampling can speed up the mining process by reducing I/O costs and drastically shrinks the number of transactions. By reducing the number of database scan reduces the time required to mine the association rules.

A new approach to extract non-redundant association rules from multilevel dataset without information loss is given in the literature [44]. This approach identifies and remove exact and approximate multilevel and cross level association rules which are hierarchically redundant. Paper introduces three interestingness measures from which two are diversity
based and one is distance based. Diversity measure determines diversity of rule items based on their position in hierarchy. While distance measure forms comparison between two different associations rules and finds items not containing in both of the rules. This approach is used for cold-start user problem in recommender system. With generation of multilevel and cross level association rules, frequent interested items are found. This information is further used on sorted user profiles. The approach allows association rules which are used in recommender system for improving the performance [44].

Data mining methodology is used to analyze the student’s career with the use of Sequential pattern mining algorithm. A concept of an ideal career is proposed where sequences of exams are taken by an ideal student without any delay. Paper uses CloSpan algorithm which mines frequent closed sequential patterns instead of finding all complete sequence of frequent patterns. Paper considers sequential events in the form of exams taken by the student and the semester in which exam has been taken. Delay is given in the form of difference between the semester in which student takes an exam and semester in which course has been conducted by the teacher. Paper gives case study of students undergoing Computer Science at University of Florence and Maths exam. Depending on the final grade and length of the study algorithm classifies students into good or not so good [45].

Continuous Association Rule Mining Algorithm (CARMA) is proposed in the literature [47] to find large itemsets online. This allows user to change support at anytime. Algorithm needs two scans of transaction sequence. In the first scan an algorithm forms a lattice of an itemset. It provides lower and upper bound for its support. User is free to change Support thresholds. In the second scan it determines precise support for each set in the lattice. Experimental study of the paper shows that number of scans needed for Continuous Association Rule Mining Algorithm compared to Apriori and ‘DIC’ algorithm are less. Continuous Association Rule Mining Algorithm performs better with lower Support thresholds. Continuous Association Rule Mining Algorithm readily computes large items and it is more memory efficient [46].

Student usage data from moodle to find rare Association Rules is used in paper [47]. It has presented Rare Association Rule Mining (RARM) to find rare or non frequent itemsets for generating interesting rules from it. Paper shows an experimentation done on the students of computer Science at university of Cordoba. Comparison between AR_MINE and different ‘RARM’ algorithm to find rare ‘AR’ is done. Different Apriori algorithms are compared to find number of generation of rare rules like ‘Apriori-Frequent’, ‘Apriori-Infrequent’, ‘Apriori
–Inverse’ and ‘Apriori–Rare’. It is proved that the use of ‘RARM’ algorithm is more efficient to determine infrequent student’s behaviour in e-learning environment like moodle [47].

Drawback of ‘CPAR’ algorithm is given in the literature [48] as imbalanced distribution of rules in each class, probability of wrong classification and generation of non-useful instances satisfying no rules. ‘ICPAR’ is introduced to handle these problems. ‘ICPAR’ algorithm gives class weighing adjustments, cluster vector based pre-classification and post processing with support vector machine. Experimentation on Chinese text classification Corpus TanCorp is carried out. Experimental study shows that ‘ICPAR’ achieves more improvement compare to ‘CPAR’. To balance classifying ability of each class, class weight adjustment algorithm used. With Centre Vector Based Prediction Base Classification rules with high level of concern are loaded which reduces the probability of incorrect classification. Also for classes with no rules, post processing with support vector machine is used. With all these features ‘ICPAR’ gives a good quality in extracted rules [48].

Closeness preference is proposed for generation of sequence rules with close itemset based on user time preference. This measure uses a parameterized slopping function. This is used in web analysis for predicting the page which will be visited by user. The experimental study shows that Closeness preference measure is efficient to find small and simple sequential rules [49].

A problem of finding rules from collection of large number of association rules is proposed when one is interested in finding behavior of small subgroup of original dataset. A Rule visualize tool is given for management of association rules with its visualization. It has three components as Rule Selection, Rule Browsing and Rule Graph. With a rule selection tool user can specify confidence, support threshold and commonness. It specifies the size of rule and maximum number of rules. With rule browsing tool, set of interesting rules can be found. This rule can be browsed in different modes as textual, graphical or combined mode. In a rule graphs tool, rule is visualized as a hyper graph where each edge contain all attributes appearing in a rule and each edge has distinct vertex with its attached weight as support and confidence. An attribute graph model includes technique as tree model where each attribute is represented as node and association are shown with an arc. The thickness of an arc represents confidence or support of corresponding rule. Rule visualizer also gives number of ways to reduce complexity of graphical presentation. It offers a template mechanism which limits number of displayed rules. With rule visualize, user can manipulate rule template directly [50].
Textual documents created and distributed on the Internet are ever changing in various forms. Most of existing works are devoted to topic modeling and the evolution of individual topics, while sequential relations of topics in successive documents published by a specific user are ignored. The proposed method is used to characterize and detect personalized and abnormal behaviors of Internet users. For that purpose the Sequential Topic Patterns (STPs) and formulate the problem of mining User-aware Rare Sequential Topic Patterns (URSTPs) in document streams on the Internet is needed. They are rare on the whole but relatively frequent for specific users, hence are applied in many real-life situations, such as real-time monitoring on abnormal user behaviors. Paper contains a group of algorithms to solve this innovative mining problem through three phases: preprocessing to extract probabilistic topics and identify sessions for different users, generating all the STP candidates with (expected) support values for each user by pattern-growth, and selecting URSTPs by making user-aware rarity analysis on derived STPs. Experiments on both real (Twitter) and synthetic datasets shows that the presented approach discovers special users and interpretable URSTPs effectively and efficiently, which significantly reflect users’ characteristics [51].

A literature on e-commerce data is given in paper [52], a product review in e-commerce is helpful for prospective buyers to make decision whether the product that they want to buy is good or bad and it can help the sellers to get their costumers feedback. As there are two problems exist: the number of product review increases day by day and e-commerce allows their consumers to write positive negative opinion about some product features in one review, as known as ”free format”. Hence, system is needed which can extract product feature from review and classify their opinion automatically. Class Sequential Rule (CSR) method is implemented in product feature extraction and Opinion Lexicon method is implemented in feature opinion classification. The best $F-score$ of feature extraction using ‘CSR’ in free format review is 51.26% and the best $F-score$ of opinion classification using Opinion Lexicon in free format review is 35.65%.

Problem of handling redundant rules is given in the literature [53], Mining sequential rules helps to discover useful sequences in sequence databases quickly and efficiently. Most of the proposed algorithms focus on generating all possible sequential rules. The technique produces a lot of redundant rules, affecting efficient mining. In order to solve this problem, mining non-redundant sequential rules has been presented. However, the algorithms proposed are dependent on obtained patterns of the existing frequent pattern mining algorithms. Several steps are needed to organize the data structure of these patterns before being used for
generating rules efficiently. This phase takes a lot of time and memory usage. The proposed algorithm uses a compressed data structure and adopts a prefix tree in the mining process. Moreover, the proposed algorithm uses some pruning techniques to remove unpromising candidates which enhances the efficiency of the algorithm in term of runtime and memory usage.

Although many studies have been carried out for sequence classification in the past decade, the problem is still a challenge; particularly for pattern-based methods. There are two important issues related to pattern-based sequence classification which motivate the present work: the curse of parameter tuning and the instability of common interestingness measures. To alleviate these issues, a new approach is used for mining sequential rule patterns for classification purpose. Literature [54] introduced a space of rule pattern models and a prior distribution defined on the model space. From this model space, a Bayesian criterion is used for evaluating the interest of sequential patterns. A parameter-free algorithm is used to efficiently mine sequential patterns from the model space. The new method identifies interesting and robust patterns; the direct use of the mined rules as new features in a classification process demonstrates higher inductive performance than that of sequential pattern based classifiers.

A key phrase (a multi-word unit) in a document denotes one or multiple keywords capturing a main topic of the document. Finding good key phrases of a document can quickly summarize knowledge for efficient decision making and benefit the domains involving intensive text information. Existing key phrase extraction methods cannot be customized to each specific document, mainly because their patterns used to form paraphrases are too restrictive and may not capture flexible keyword relationships inside the text. In paper [55], the proposed method is sequential pattern mining based document specific key phrase extraction method. The key innovation is to use wildcards (or gap constraints) to help extract sequential patterns, so the flexible wildcard constraints within a pattern can capture semantic relationships between words, and the system will have full flexibility to discover different types of sequential patterns as candidates for key phrase extraction. To achieve the goal, each single document is taken as a sequential dataset. The proposed algorithm is used to mine sequential patterns with wildcard and one-off conditions that allow important key phrases to be captured during the mining process. For each extracted key phrase candidate, some statistical pattern features are used to characterize it, and further all key phrases are collected from the document to form a training set. A supervised learning classifier is trained to
identify key phrases from a test document. As this pattern mining and pattern characterization processes are customized to each single document, key phases extracted from this method are highly specific for each document. Experimental results demonstrate that the proposed sequential pattern mining method outperforms existing pattern mining methods in both runtime performance and completeness. Comparisons on key phrase benchmark datasets also confirm that the proposed document-specific key phrase extraction method is effective in improving the quality of extracted key phrases.

Research proposes and implements an expert system to predict earthquakes from previous data. This is achieved by applying association rule mining on earthquake data from 1972 to 2013. These associations are polished using predicate-logic techniques to draw stimulating production-rules to be used with a rule-based expert system. The proposed expert system in the literature [56] is able to predict all earthquakes which actually occurred within 12 hour at-most. Expert systems (ES) are a branch of applied artificial intelligence. The basic idea behind ES is simply that expertise, which is the vast body of task-specific knowledge, is transferred from a human to a computer. ES provide powerful and flexible means for obtaining solutions to a variety of problems that often cannot be dealt with by other, more traditional methods. Hence, their use is proliferates many sectors of our social and technological life, where their applications are proving to be critical in the process of decision support and problem solving. Earthquake professionals for many decades have recognized the benefits to society from reliable earthquake predictions, but uncertainties regarding source initiation, rupture phenomena, and accuracy of both the timing and magnitude of the earthquake occurrence have often times seemed either very difficult or impossible to overcome [56].

Identifying irregular file system permissions in large, multi-user systems is challenging due to the complexity of gaining structural understanding from large volumes of permission information. This challenge is handled when file systems permissions are allocated in an ad-hoc manner when new access rights are required, and when access rights become redundant as users change job roles or terminate employment. These factors make it challenging to identify what can be classed as an irregular file system permission, as well as identifying if they are irregular and exposing a vulnerability. The current way of finding such irregularities is by performing an exhaustive audit of the permission distribution; however, this requires expert knowledge and a significant amount of time. Paper [57] specifies a novel method of modeling file system permissions which can be used by association rule mining.
techniques to identify irregular permissions is presented. This results in the creation of object-centric model as a by-product. This technique is then implemented and tested on Microsoft’s New Technology File System permissions (NTFS). Empirical observations are derived by making comparisons with expert knowledge to determine the effectiveness of the proposed technique on five diverse real-world directory structures extracted from different organizations. The paper results show that the technique is able to correctly identify irregularities with an average accuracy rate of 91%, minimizing the reliance on expert knowledge. Experiments are also performed on synthetic directory structures which demonstrate an accuracy rate of 95% when the number of irregular permissions constitutes 1% of the total number. This is a significant contribution as it creates the possibility of identifying vulnerabilities without prior knowledge.

Literature on utility mining is given as-High utility itemset mining is an emerging data mining task, which consists of discovering highly profitable itemsets (called high utility itemsets) in very large transactional databases. Many algorithms have been proposed to efficiently discover high utility itemsets but most of them assume that items may only have positive unit profits. However, in real-world transactional databases, products often have positive or negative unit profits. Mining high utility itemsets in a transactional database where items have positive or negative unit profits is a computationally expensive task, and it is thus desirable to design more efficient algorithms [104-105]. To address this issue, paper proposes an efficient algorithm named Faster High-Utility itemset miner with Negative unit profits (FHN). It relies on a novel ‘PNU-List’ structure (Positive-and-Negative Utility-list) structure to efficiently mine high utility itemsets, while considering both positive and negative unit profits. Moreover, several pruning strategies are introduced in ‘FHN’ to reduce the number of candidate itemsets, and thus enhance the performance of ‘FHN’. Extensive experimental results on both real-life and synthetic datasets show that the proposed ‘FHN’ algorithm is in general two to three orders of magnitude faster and can use up to 200 times less memory than that of algorithm ‘HUINIV-Mine’. Moreover, it is shown that ‘FHN’ performs especially well on dense datasets [58].

1.3 Problem Definition

Appropriate use of extracted information and knowledge with different data mining techniques is a major field of research. Hence, improvement in design and development of Association Rule Mining technique is necessary to enhance the usage of real world data. Novel and efficient algorithms are used to discover valid and user interesting association
rules from a large transactional dataset. Varieties of measures with user interestingness are assessed to ensure the quality of rules. This research is applicable to real world scenarios.

Most of the algorithms of Association Rule Mining (AR_MINE) mainly focus on finding frequent items with efficiency. Less attention is given to the quality of rules. Generation of unnecessary large set of rules reduces the effectiveness of rule mining algorithms. If more focus is given to the quality of rule then these strategies are used to solve real world problems. Set of valid rules are formed by adding constraints to the source dataset at pre-processing or at post-processing stage.

The problem of generation of association rules from sequential dataset is given as follows- A sequence is a collection of an ordered list of itemset where itemset is a collection of unordered, non empty set of items. Consider a transactional sequence database \( D \) containing set of sequences as \( S \) and set of items as \( I \). \( S \) is given as \( S = \{s_1, s_2, \ldots, s_m\} \) and the set of items \( I \) given as \( I = \{i_1, i_2, \ldots, i_n\} \) present in the set of sequences. Each transaction has transaction identifier \( TID \) and respective sequence \( S \). A sequence is represented as list of items or set of items as \( s_1 = I_1, I_2, \ldots, I_r \) where \( I_1, I_2, \ldots \subseteq I \). For a given sequence databases an association rule is generated of the form \( A \rightarrow B \) where \( A \) represents antecedent and \( B \) represents consequent. Both \( A \) and \( B \) are the frequent itemsets of sequence database \( D \).

Frequent itemsets from sequence databases are found with ‘Support-Confidence’ framework. In order to generate rules with more interestingness measures such as \('lift', 'consequence', 'leverage' and 'interest'\) are added. Restrictive class of constraints to the sequential patterns is obtained with adding constraints at the pre-processing level which are in the form of item constraint, item length constraint, regular expression and minimum gap etc.

### 1.4 Problem Statement

**Association Rule Mining on Constraint based Sequential Patterns:**

For a given transactional sequence database \( D \), 'MinSup', 'MinConf' and set of constraint as \( C = \{\text{MinGap}, \text{MaxGap}, \text{Regex}, \text{Itemconstraint}\} \) to find valid and user interesting positive and negative association rules which satisfy all constraints in \( C \). Here 'MinSup', 'MinConf' are the input thresholds as minimum Support and minimum Confidence respectively. \( C \) is a set of constraint given in the form of gap constraint, regular expression and item constraints.
1.5 Objectives

Objectives of the present research are-

a) To analyze existing association rule mining algorithms.
b) To find an efficiency of association rule mining algorithms.
c) To develop novel pruning strategies for association rule mining.
d) To find the effect of pruning by pushing constraints in the mining process.
e) To design an association rule mining algorithm with constraint for sequential patterns.

1.6 Hypothesis

This research proposal presents the detailed study and analysis of association rule mining algorithms. The algorithms are experimented on sequence dataset. Constraint based sequential pattern mining algorithms are designed to prune the search space. Constraints are chosen on user interestingness. 'Item', 'length', 'gap' constraints are added during the pre and post processing step of mining.

1.7 Methodology

Methodology adopted for the present research is given as- This includes the different steps involved in software development life cycle like Requirement Analysis, Design, Implementation, Testing and evaluation.

1. Data mining algorithms work on very large databases. Since the source data comes from different databases which may have duplications and inconsistencies. First step needed is to preprocess the source data.

2. In preprocessing the first step is data cleaning and the second step is to select the related data from integrated resources and transform them into a format which is ready to be mined.

3. Extraction of frequent patterns from preprocessed data with association rule mining algorithm.

4. Pushing constraints at post processing step to get filtered user interesting patterns.

5. Forming association rules on frequent filtered patterns.

6. The analysis of an algorithm is carried out on various datasets such as very large, average case and smaller dataset.
1.8 Organization of Thesis

Thesis elaborates the designing of association rule mining algorithms on sequential dataset and analysis of algorithms with application of constraints. This work is given in five chapters as-

Chapter One presented in the thesis gives review of literature on sequential pattern mining techniques, use of different measures, novel approaches used in the constraint based mining etc. Discussions and contributions of various authors are elaborated in the chapter. Problem statement, objectives, hypothesis and methodology are presented.

Chapter two presents proposed system architecture and theoretical background needed for frequent pattern mining and generation of association rules. Frequent pattern mining with different approaches is explained. Need of pushing constraint in mining is elaborated; further different measures involved during mining process are explained in detail. Factors affecting the improvement in efficiency of association rule mining algorithms are discussed. Comparison of different approaches for frequent pattern mining is performed. Pushing constraints for generation of association rules improves efficiency of AR_MINE algorithms has been proposed in this research work.

Chapter Three discusses on designing and Analysis of Association Rule Mining algorithms. Algorithms for finding frequent itemsets with two approaches are designed as Apriori and Pattern growth method. Further Rule generation algorithms are designed for both approaches. Chapter also discusses on the generation of negative association mining algorithms. Apriori based sequential pattern mining and Pattern growth approach for generations of association rules with constraints is analyzed. All algorithms are analyzed theoretically and with an example of real time dataset. Analysis of algorithms is carried out with respect to time and space efficiency.

Experimental work is discussed in Chapter Four. Design, analysis and experimental results are presented for various sequential datasets. Experimentation is performed with varying dataset and constraints.

Chapter Five presents summary, results and discussion. The significance of research work, analytical and graphical comparison of the work with existing work is performed. Chapter
ends with conclusion and directions for future work.

References section represents referred research papers from reputed journals, books, conference proceedings, websites, chapters, notes, presentations etc. to perform the presented work.