CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

The world revolves around the recorded events such as birth of an offspring, celebrations, occasions, incidents and other happenings. The events are often referenced in the history for further studies and information. In the field of computer technology, computers of any kind need deeply sensed data and analysis of them, framing into appropriate combinations to predict the future evolution. Anyone may question the necessity for recording the occurrences in computers with related technologies unless they understand the concepts of prediction and forecasting of future events for getting solutions from the already failed cases. Going through the observed experiences, has indeed yielded much better results in most cases. Tracing back the detailed information thus requires storage and relevant technologies to determine actions to be taken further.

Premonition of later disastrous occurrences in designing and implementation of any product would greatly strive to conserve human effort and cost of production. There are certain domains which absolutely need these predictions to avoid such heavy losses. Science and Engineering, Financial exchanges, Weather forecasting, Business, Medical applications and much more domains urge the implication of information extracting strategies for the betterment of future. This process has its origin back to centuries. Thomas Bayes laid the foundation in the seventeenth century (which was then never
believed to be right) for focusing the similarities among the recorded data and predicting further actions. Named after the founder, Bayes theorem is today’s methodology for efficient analysis over the occurrences based on probability of previous records. The analysis of those events produces the likeliness and chances of occurrences of events with all feasibilities.

Following this theorem, in eighteenth century, Regression analysis proposed another mode of extracting information from the recorded events by a focal point mechanism on the relationship among the events. A range of happenings may occur but there always exists a similarity among them, leading to a prediction of time and percentage of next occurrence. There are events which stimulate the sequence of activities to be followed after a specific process. The flow of events initiates the rest in every cycle of execution, that is, an earthquake hit region would need the initiation of supplying medicines and revamping operations. Corresponding events substantiate the specific happening irrespective of the dependency level. Forecasting and predictions of designs and implementations of any product reserve the importance of regression analysis. There are multiple components individually designed and implemented for the constitution of a whole product. All those components need to be cooperative for its successful functions. The failure in the functions would disturb the whole process ultimately. Failure analysis and implemented regression reports help determine the relationships among the used components and the impact of a failed event. Regression analysis defines the dependability level on each other effectively.

The need of extracting the information from mere data is thus validated to be a primary process in the planning of future actions. The computer society has enlarged to a very large extent concluding its presence in almost every nook and corner. Memory resources, data storage and tools
for manipulating the same have also improved to a greater extent. Algorithms and innovative concepts have been proposed in various modes to extract useful information from mere recordings of time and situation. They have been evolving with advancements in computers and its associated technologies. A simple database from file systems has grown to a distributed databases on different machines installed in various nations. Handling of such distributions needs primitive as well as sophisticated techniques and terminologies. Data mining has a very strong reason to be applied in many fields for various purposes.

1.2 RESEARCH FOCUS

The separation of useful information from a large set of data items needs deep and sensible approaches to identify and isolate the relationships among each individual unit of entries. Implication of these methodologies has helped in its unique way to recognize the dependability, relationship, reliability of data items. Data are entered into the storage spaces as individual units as they take place. Registering all events for a specific time period would group those entities into a data itemset. The itemset comprises of occurrences of activities by every user denoting the time and sequence. Yet there needs a novel method to associate, relate, unite those entries in order to extract information. Information regards how the activities are commenced and the associated activities are stimulated. These analysis and reports on association are of high importance in banking, stock exchange and insurance companies.

This research contemplates on the concepts oriented with communication between independent data item sets of assorted organizations. There exists an obvious link between transactions of a single organization. The data to be requested and provided to the organization for rendering services has a certain connection between other transactions, termed to be
Intra-transactions. They are subjected to the services provided by the organization and never beyond. Intra-transactions are bound within the nature of the organization. Considering a supermarket, all the purchases made by a customer are limited to the billing section irrespective of the categorization of the items purchased. The transactions would be organized into the association of items, combination of items, and preferences of customers to offers provided by the organization. The stock level of the supermarket, profit range, organizing employees, their salary and bonuses are other transactions within a supermarket organization. All these activities are claimed to be Intra-transactions of very specific working environment. The sale of a seasonal product would initiate the promoting offers to attract customers. A product which would never go alone in the sales without a combination product, can be combined to be a combo pack with slash in its price. These are based on the objectives of the organization to enhance the flow of customers, to withstand their competitive level among other similar organizations and the quality measures need to be maintained with affordable cost. From these descriptions, Intra-transactions can thus be categorized and easily maintained with the dedicated framework and terminologies.

The primary constrain is that the same terminologies of Intra-transactions environment need some modifications in case of Inter-transactions association rule mining. A new approach has to be designed for comparison of inter related but typically different domains. There are multiple examples to portray this scenario. Organizations of the same characteristics need to contend with each other to be a stable competitor in the market. Without tracking the others activities in a constructive manner, one will not be able to compete in the market. Considering the computer technology, there exist a number of stable and well renowned organizations manufacturing and marketing their products to achieve a progress every round the clock. The organizations such as Microsoft, SUN and IBM challenge others to predict
and process the similar product with equivalent quality and standards. It
requires an absolute knowledge on the events of the organization every now
and then. Their events are termed to be Inter-transactions, being of any kind
from the introduction of a new product into the network or the consecutive
raise and fall of share prices.

Transactions of a single organization are considerably normal tasks
to obtain patterns of events and to describe the association among various
events. Intra-transactions are simpler since they are subjective to a single
customer or a single day of an obvious single organization. They are denoted
by a Transaction ID, Customer ID, and the sequence of transactions take
place. When subjected to analysis, those transactions give a clear-cut view on
the relationships and dependability level of associated events. Categorized
under the same customer, finding the relativity of most purchased items and
the time of purchase follows an order. Traditional transaction association
methodologies spot those similarities very easily. There are also other
dimensions such as time and space to further define the associations. Analysis
within the components and events subjected to a single organization is thus
evident that identification of patterns by exploiting data items does not need
deep concern. Yet these terminologies lend their support in implementing a
framework for larger and multiple organizations. There still subsist alterations
for implication of Intra-transaction concepts. Apart from these limitations, the
Intra-transaction association forms a compartment of Inter-transaction
association.

Every organization has its very own vision and mission on its
functionalities, differentiating itself from the others of same kind. Features of
different nature lead to limitations when grouped together for exploiting the
data items of different concerns. Transactions are represented in the best way
of understandability and thus stand as an obstacle for comparisons. The
analysis of Inter-transactions associations proves to be more challenging than the previous case. The frequently purchased items cannot be directly compared even if they possess the same criteria in Inter-transaction association. There is always an urge to be more specific rather than general representation of metrics. The Inter-transactions require additional and sincere effort to be put forth in denoting the time and space dimensions of transactions. The general representation would establish an uncertainty which degrades the performance. Apparent description of events should be specific such as “hitting of a storm in Chennai on Tuesday is confirmed” to initiate the recovery mission rather than “a storm may hit some areas in the next week”. Beyond all these limitations, an association rule mining has to be identified in varying domains. This research focuses on novel methodologies to be implied in different organizations and overcome the obstacles faced by Intra-transaction association rule mining. Limitations in the classical association rule mining are diagnosed and the additional alterations are made to propose an efficient framework for achieving successful in Inter-transaction associations which are discussed in the following sections of this work.

1.3 DATA MINING

Data mining, as its name suggests, is a process of mining data from a composition of numerous data items stored in the memory of an organization. The need for data mining is clear understandability as not all data items stored are capable of explaining the context and theme of the process. It combines the terminologies of computers and statistics for determining the patterns that exist. Only the important parts of entire set can be enough to describe the whole concept and reason behind of every event that constitutes the intact structure of an organization. Information has to be extracted by carrying out the prescribed actions of data mining from large
volumes of relevant, irrelevant and noise composition. Data mining is one among the primary processes of Knowledge Discovery in Databases (KDD).

1.3.1 Knowledge Discovery in Databases (KDD)

KDD is an overall process of discovering useful knowledge from data. The tremendous growth in data has generated the need for new techniques that can intelligently transform massive data into useful information and knowledge. Knowledge Discovery in Databases (KDD) is the automated extraction of novel, understandable and potentially useful patterns implicitly stored in large databases, data warehouses and other massive information repositories. The KDD is a multi-disciplinary field, drawing work from areas including Database Technology, Artificial Intelligence, Machine Learning, Neural Networks, Statistics, Pattern Recognition, Information Retrieval, High performance Computing and Data Visualization.

Knowledge Discovery consists of an iterative sequence of the following steps:

- **Data Cleaning**: To remove noise and inconsistent data.
- **Data Integration**: It is the process in which multiple data sources are combined.
- **Data Selection**: It is the process in which data relevant to the analysis task, are retrieved from the database.
- **Data Transformation**: It is the process in which data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
Data Mining : An essential process in which intelligent methods are applied in order to extract data patterns

Pattern evaluation : It is the process to identifying the truly interesting patterns representing knowledge based on some interesting measures.

Knowledge presentation : It is the process in which visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Figure 1.1 Knowledge Discovery Process

Figure 1.1 shows the details of steps in knowledge discovery process. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base. It is noted that according to this view, data mining is only one step in the entire process. It is an essential one since it uncovers hidden patterns for evaluation.
Data mining is a step in the knowledge discovery process, however, in industry, in media, and in the database research milieu, the term data mining is becoming more popular than the longer term of knowledge discovery in databases. To adopt a broad view of data mining functionality, data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.

Data mining is an essential step in the process of knowledge discovery in databases, in which intelligent methods are applied in order to extract patterns. Other steps in the knowledge discovery process include pre-mining tasks such as data cleaning (removing noise and inconsistent data) and data integration (bringing data from multiple sources to a single location and into a common format), as well as post-mining tasks such as pattern evaluation (identifying the truly interesting patterns representing knowledge) and knowledge presentation (presenting the discovered rules using visualization and knowledge representation techniques).

Several algorithms have been proposed to exploit the data volumes and consolidate the necessity enough for representing the whole ideology. Data mining on the whole comprises of the following activities to extract useful and needed information. They are,

1. Anomaly Detection
2. Association Rule Mining
3. Clustering
4. Classification
5. Regression Analysis
6. Summarization
1.3.2 Anomaly Detection

The anomaly in data mining refers to the unexpected and undesired patterns that appear to disturb the efficiency of databases and transactions. The defects may be due to improper design of the structure, improper planning, unexpected human errors and poor manipulation of data stored. There exists a strict order to be followed in the sequence of transactions. Any deviation from the stated order would detect a defect, and their characteristics are analyzed to determine their impact level on the system. Intrusions, irregular descriptions, defects, outliers, abnormal regions, alterations in prescribed path are some of the anomalies commonly detected in a system. Irregular entries faking the authenticity and breaking the security aspects of a confidential system are due to presence of defects in the design of authorization modules.

Anomaly detection plays an important role in conserving the human effort of redoing a work due to a failure. Prior detections of anomaly enhance the efficiency by preventing a big disastrous failure of the system. Detection in individual modules reduces the bulk testing mechanisms at the final stage of the product. Implementation of anomaly detection needs to be accurate enough to eradicate all chances of defects to produce the output as desired and error free. There are higher range of errors in complex databases which retard the functionality of control flow, data flow, storage and manipulation of transactions from and to the memory resources.

1.3.3 Association Rule Mining

A database consists of large volume of different data variables that have occurred at specific time and space domains. There is a strong justification for identifying the relativity and associations between these variables. The identification of those patterns in an organization leads to
promotions in accessibility based on the frequency of visits made. Rakesh Agrawal et al (1993) proposed the ideology of finding the firm regularities between the variables in a data set. The never-go-alone products possess a strong binding between the associated products and this relativity is called as Association Rule Mining.

By considering an example, this term would be explained clearly. The purchase of bread loafs will be followed by the purchase of jams or eggs or both. Individually all these events are transactions which are interrelated. The number of choices for purchasing bread loafs alone is very minimal. Defining this relativity metric would help in the placement of these items nearby, reducing the stress of the customer to search for or introducing a combo offer as a marketing strategy. Everything is based on the market basket analysis which determines the relativity with the probability of combinations. There are standard terminologies used for association rule mining. A database used for storing the variable records the events taking place every instance. These variables are stored as items which are attributes of transactions. An association rule is the relativity or dependability measure between two consecutive transactions or variables of the database.

A rule connects two variables antecedent (one on the left) and consequent (another on the right). The level of association is represented by the rule. There are two primary processes that need to perform the association rule mining. All the frequently occurred events are identified and minimum support threshold is found for each transaction. Secondly the minimum confidences of the regulated frequently occurred events are computed. The second process is comparatively simpler since computations of confidence metrics are bound to certain algorithms, whereas the first process needs the ordering of events based on the similarities and the number of times initiated needs deep and sincere effort of the mechanism. Left out transactions may
imply a huge impact over the confidence and support levels of transactions. Void transactions are eliminated to avoid unnecessary computations.

The other constraints which determine the level of dependability are support, confidence, lift and conviction of the variables on the other. Support is defined as the metric which defines the ratio of an event or transaction in the whole set of transactions. Confidence is the metric which denotes the support for an event to occur. The confidence describes the high chances of the next event to be occurred along with the previous occurrences. Lift is slightly varied from the support and confidence. Lift measures the independent occurrence of a transaction with the extended support of the combination of related events / transactions. Conviction determines the percentage of occurrence of a transaction without the other transaction.

1.3.4 Clustering

Clustering is a process of grouping a set of objects with similarities separating them from other characteristics of another group. Clustering simplifies the process of data mining, as the similar events are already categorized into a boundary.

Figure 1.2 Cluster Formations
Distances are primary factors to be considered when grouping the set objects, as the transactions takes place collectively. This refers to a process and does not subject to specificity of algorithms. It is well depicted as how the clusters are formed by different data objects in Figure 1.2. Beyond the limitation, it still forms a strong model for clustering in complex business models. The fourth kind being the Density based clustering, groups the data items which are most likely to be occurred rather than less frequent items. The number of occurrences displays the density of the option. The condition of high significance is not to miss out any of the important objects while forming a cluster. All the nodes should fall under a cluster and thus provide its service to the system irrespective of how large and complex the organization is.

1.3.5 Classification

Classification is associated with clustering using a more intelligent strategy. When a new event or transaction takes place, it has to be immediately categorized to the best class it fits in. The classifier algorithm already has obtained the information on the available categories with derived patterns to identify each group of transactions. A new transaction occurring would soon be analyzed for the resemblance of patterns and obtain a best location it suits. Attributes of the observed transaction considered for this classification are the type of entry (integer, float or character), category of observation (blood groups) or a real world entity (temperature measured on a thermometer).

Figure 1.3 shows an example of a classification process. Clustering forms the boundary based on mere similarities without a study on the deeper characteristics of each event. The classification is more specific in allocating a suitable point to be mapped, with probabilistic analysis. A new transaction is compared with all destined classes and a probability of matching features is determined. Unlike clustering, a number of classes are checked for best
optimality and highest probability is chosen to be the right category. In clustering, the distance metrics and vector values are analyzed without any matching features. Various strategies are implemented in classifications, binary or multiclass, vectoring of features and frequentist and Bayesian approaches.

![Monsoon Data Classification](image)

**Figure 1.3 Classification of Data and Prediction of results**

Either two classes or multiple classes are noted for analyzing the probabilities of a new transaction. The features which are used for analysis are whether the gender, height and weight metrics, medical applications and recorded events of weather and earthquake occurrences.

### 1.3.6 Regression Analysis

There are a number of variables in a transaction that may be dependent, independent and their associated attributes or parameters. The independent variables are well organized to function individually but they affect the activities of corresponding dependent variables. In an example,
purchase of jam is purely dependent on purchase of bread loafs whereas the purchase of butter goes with a limited dependability on purchase of bread loaf. Butter is a variable with partial independent level and has an equivalent probability of being purchased separately or with other combinations. Purchase of jam purely depends on bread loafs completely. Hence the dependability and relationships among the variables are determined by the frequency of purchased together, separately or with other variables. These relationships are determined by the regression analysis.

The independent variables are fixed and subjected to alterations which provoke the initiation of dependent variables. There are also a number of chances for establishing a false relationship between variables, which in turn affects the efficiency measures of the organization. Gauss and Legendre proposed the idea of regression analysis concept in 1805 and 1809 respectively. Their model is to form square structures around the orbits of planets and to find out the locations of planets. Later Francis Galton introduced the same concept of finding the relationships of heights from the ancestral parents to their descending. This concept has further been enhanced by introducing concepts of statistical techniques for obtaining a constant value of dependability.

This regression analysis produces the results to be values of relativity which has to be ensured as rightly derived. The two cases of regression analyzes are interpolation and extrapolation. If the derived values of relativity fall under the range of values, it is said to be interpolation. If the predicted values fall beyond the observed values, then it is said to be extrapolation. The primary defect in this analysis is that the computations become complex if the volumes of data are considerably large and improperly manipulated. The regression analysis develops a model of checking the relationships between the variables, subjected to tests before implications.
These tests are used to determine the integrity of the model. The tests validate a model in various manners to encompass all types of association. A defined set of values would predict the distribution of the following function and a presence of an error would alter the path of distribution and report the deviation. These tests are responsible for the efficiency of a model and are used to determine how to prevent the occurrences of errors, removal of such errors at earlier stages.

1.3.7 Summarization

The documents of a service comprise of numerous statements providing the explanation of the inlaid concept through various descriptions and definitions. There are always unwanted statements, phrases, supported texts framed to uplift the original concept. Yet they are not needed for understanding of the theme of a document. The theme is ultimately the motive stressed with additional garnishing. Summarization is a process of extracting the key terms from decorations and providing the same meaning as the whole context tries to achieve. If one tries to read a document in a hurry, the time taken to obtain the meaning would be long as the main idea lies in a line or two. In case of a summarized document, the whole idea could be reconstructed and save the time needed for analysis.

There are two types of summarization techniques namely Abstraction and Extraction. Extraction comprises of selecting a detachment, from the whole set of statements, phrases and other participants. This division produces the original meaning of the document which focuses on the elimination of the supportive words. Extraction merely selects the important keywords from the text document and shortlists the summary whereas abstractions includes strategies of developing an internal semantic description of the original document with the help of natural language support and produces the output document nearly similar to a human effort. The
preprocessing of these extracting keywords from the text document includes the eradication of supportive words by filtering process. The next process is the determination of the tags of each selected key sentence. The tags alone would identify the theme of the document. There are certain types of summarization based on the criteria namely generic summaries, multi summaries and automatic summarization.

![Summarization Tasks](image)

**Figure 1.4 Summarization Tasks**

The summaries consolidated from the queries provided as input to the system form query based summary. Figure 1.4 shows the different tasks in summarization process. Summarization of a single document is easier and multiple copies need more standards and stringent methodologies. There are also summaries of images, video and audio archives in many web based applications.

All above cited tasks are common processes in Data mining, which strive to direct the storage and manipulation of transactions, events from the
memory with maximum efficiency and lesser cost on effort and resources. The following section discusses the contributions of the respective thesis.

1.4 APPLICATIONS OF DATA MINING

Data Mining advocates its role in a number of applications with incredible support for easier and faster accessibility. They are prioritized based on their significance. The different applications of data mining have been incorporated into the following scenarios.

1. Governmental Organizations
2. Geographical Environments
3. Technological Advancements
4. Business Developments

Governmental Organizations

Organizations of the Government act accordingly to the records of every day-to-day life. Proofs of registration of lands and vehicles, birth and death certificates, marriage registrations, passports, licenses, certificates of community, orders of government authorities are some of the well known documents issued. These documents withstand the rights and level of questionability of every individual. These records need to be stored and retrieved during verification. The question arises on the possibility of accessing those documents with accuracy and rapidity in a highly populated region or entire nation. Computerized control of authorities necessitates the implementation of data mining in governmental organizations. Data mining has considerably reduced the time of accessing and taking appropriate decisions in issues regarding the preservation of human rights. Proper maintenance of these standards ensures the efficiency of any organization in any region or nation.
Beyond the need for maintaining these multiple set of records, a nation has to be conscious enough over the security measures. Terrorist activities are analyzed by pattern mining ethics to determine the transactions between groups in suspicious regions, pattern of attacks in highly influential localities and migration of a particular set of attackers prior to a disaster.

**Geographical Environments**

Analysis of entities over a large distributed environment is necessary for determining the spread of a deadly disease, saturation level of customers or dealers in a specific region, assigning marketing executives to a region, for agricultural purposes such as climatic change and rainfall measure etc. Spatial data are obtained as readings of these distributional areas. Obviously, these entities have to be mined in order to obtain a clear view of requisites. Spatial data is a combination of different metrics of space, distance or area which may be similar on the outline but possessing serious differences. There are features and attributes which distinguish the spatial data. Any organization dealing with the processing of spatial data needs to follow some standards to achieve success as the outcome. Challenges in mining the data of geographical environments are allocating a huge space for storage and manipulation, recognizing the right data from aggregation of data set and representations of those in a convenient way and finally retrieving the patterns of association.

Remote sensing is one of the methods implied in the areas where no human intervention is necessary for a comprehensive set of data of various purposes. Numerous sensors are distributed in a region for sensing, collecting and forwarding those data to the base station for analysis. They may be grouped into clusters or networks to remove the redundancy of data transmitted to the base station. This methodology overcomes many difficulties in observing the data from the regions which are unfit for human existence.
Implementation of sensors is comparatively cheaper and faster technology than any other methods. This ideology has evolved from Geographic Information System (GIS) which faces many problems before the integration of data mining concepts. GIS has provided the details on sensed events from different regions, but has limited modes to derive patterns. Integrated with data mining concepts, the current method promotes the use of both technologies into one efficient strategy.

**Technological Advancements**

There are many advancements and research areas for implementing the functionalities of data mining into their core. Bioinformatics, Genetic engineering, Nanotechnology, electrical engineering are some of the core technologies which necessitate the need of data mining concepts. Genetic engineering and bio engineering involve the periodic maintenance of genes and their combinations and their growth metrics. The disease and immune anti bodies are developed by using the same technology, after regular analysis on the growth and other events. Medicines are tested before entering into the market as a suitable remedy for a specific disease. The dosage level of a medicine and its corresponding impact on the disease causing vectors are subjected to a number of studies by various experts. The patterns of drug consumption and recommendations to criteria of people, distribution of disease in different areas are retrieved by the conceptual model.

Electrical industries also make use of data mining concepts in Dissolved Gas Analysis (DGA) and vibrations monitoring of power transformers. Electrical devices of many industries mandate frequent monitoring on the voltage fluctuations, for the safety of product and employees around them.
Educational institutions implement the methods of data mining for analysis of students’ details, their marks, and patterns which affect their level of learning capabilities. Every student has a unique approach of learning and the skill sets of teaching aids need additional standards to meet the requirements of the students. The progress of every student needs to be checked and adequate countermeasures are to be taken. Universities and other large institutions demand the techniques for betterment of a student’s academic performance.

Business Developments

As the basic of data mining concepts commence with the comparison of market basket analysis, the same has lead to the developments in business opportunities. The pair or cluster of products which are usually selected together, defines the marketing ethics of introducing a promoting offer to enhance the likability of customers and profit of the concern. Preferences to the particular product, the next product which is initiated by the first purchase and the offers which attract many customers cannot be determined without data mining. Competitors in the market, examine the statistical analysis of the quality, cost and offers of products to stand consistent among others. Data mining facilitates the determination of association rules between products, between customers and thus manufacturers. The consumer relationship is stabilized with the outcomes of these studies. Customer being the integral part of marketing industry seeks such high processing and providences for being satisfied.

1.5 EXISTING SOLUTIONS

Mining of Association rules is a centre of attraction in the field of data mining. Classical association analysis is Intra-transactional because it focuses on association relationships among itemsets within a same
transaction. Inter-transaction association rule mining extends the mining association rules to describe association relationships among itemsets across several transactions. Many algorithms for mining Inter-transaction association rules have been proposed such as E-Apriori, EH-Apriori, FITI, ClosedProwl and ITP-Miner. Anthony K.H.Tung et al(2003) have developed an algorithm, called FITI (First Intra Then Inter), which discovers frequent Intra-transaction itemsets and uses them to generate frequent Inter-transaction itemsets.

There have been many solutions for mining the Inter-transaction item sets in recent past. An Apriori algorithm is a traditional model to examine association rules among the transactions in a domain. This algorithm uses a bottom up approach to operate on the database of recorded transactions, which may be list of purchases, the raise of stock prices or visits to a website.

Transactions are analyzed to find the similarity of transactions that meet the minimum threshold value and are separated as subset of items. The step of candidate generation would extend the transaction by one if a frequent event has occurred. When no further event is found to possess these similarity constraints, the algorithms ceases. The final outcome of this algorithm would be association rules of every transaction that has been repeated for a minimal number of times for consideration. Representation of Apriori algorithm prefers a Hash tree structure and Breadth First Search (BFS) strategy is followed up. Pruning of the Hash tree is done by eliminating the lesser frequent transactions and the resultant tree would comprise of candidates of transactions with a number of significant repetitions.

First Intra Then Inter algorithm (FITI) algorithm extends its function along with the concepts of Apriori algorithm to mine Inter-transaction domains. As previously mentioned, the part of Intra-transactions forms the subset of Inter-transactions, the retrieval of frequently occurred
transactions in a specific inter domain is necessitated for the former. This algorithm comprises of two phases which have a basement of Apriori algorithm in its first phase. The list of frequent transactions in a domain is determined before subjecting to a wider inter domain. After the construction of a data set of transactions with minimal repetitions, they are named for identification and the links to the subset and descendents of other transactions are made. They are represented in the form of tables to best describe their associations. In the next phase, similar to Apriori, the association rules are framed in a level wise approach.

The next methodology simplifies the boundaries and association rules by limiting the frequent transactions within a closed continuity region. Inter-transactions are extended to a large space for inputting different transactional domains into one for determining associations. The continuity of Inter-transactions never follows any constraints over space and time and thus is never bounded. This factor demarcates the derivation of association rules since a user has to select the useful rule out of a huge list.

Breaking the continuity would not help this scenario as the transactions cannot be avoided for computations as in compressed continuity strategy. Instead, the continuities can be closed within a range and association rules can be defined without elimination of transactions as they fall under some other closures. This algorithm is named as Closed PROjected Window List (ClosedPROWL), denoting the closed continuity and projected window list methodology. A projected window list catalogs the new transactions with existing frequent transactions of time domains. The data sets are stored to be in horizontal and vertical formats merging new time slots with resemblance to existing time slots of events.
1.6 PROBLEM DEFINITION

All the algorithms for mining Inter-transaction association rules developed so far are based on Breadth-First Search (BFS) approaches that search for frequent itemsets level by level. At each level, a database must be scanned once to determine the support for each candidate itemset. It has been shown that Apriori-like approaches well perform in finding frequent Intra-transaction itemsets when the itemsets are short. The performance of such algorithms get down dramatically when mining long frequent itemsets, or using very small support thresholds. Since Apriori-like approaches may generate huge number of candidate sets at each level, they are prone to high memory consumption during the mining process. It is observed that Apriori-like methods for finding frequent Inter-transaction itemsets have the same drawbacks as those for finding frequent Intra-transaction itemsets.

1.6.1 Objectives

The primary objectives of this research work are summarized as follow:

(i) To improve the prediction rate and reduce the execution time for finding the association rules by incorporating some modifications in the existing Apriori algorithm.

(ii) To reduce the search space and improve the scalability by the application of closed itemset mining concept.

(iii) To generate the stronger, useful, abstract and meaningful association rules and also to reduce the overhead of establishing the support and confidence values from the initial stages of mining process by the introduction of generalization concept in closed itemset mining.
1.7 ORGANIZATION OF THE THESIS

The need of knowledge extraction, steps in data mining process, application areas of data mining, existing solutions and objectives of the proposed research are given in Chapter 1.

The detailed literature survey on the basics of association rule mining, closed itemset mining, concept of generalization and recent development of Inter-transaction association rule mining are discussed in Chapter 2.

In Chapter 3, the necessity of having multidimensional Inter-transaction association rules is formulated to improve the prediction rate and the execution time.

Chapter 4 deals with the reduction of number of combinations by grouping a majority of transactions into a limited number of itemsets.

The efficient mining of closed itemset patterns with the proposed module is discussed in Chapter 5.

The Merging of closed itemset concept with generalization and taxonomy is explained in Chapter 6.

Conclusions and future works proceeded by selective bibliography is discussed in the last chapter of the thesis.