CHAPTER 2

LITERATURE REVIEW

2.1 Knowledge Discovery in Data

Data mining [117] is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc. Also data mining is known as one of the core processes of Knowledge Discovery in Database (KDD).

![Knowledge Discovery in Data Diagram]

Figure 2.1: Knowledge Discovery in Data
Many people take data mining as a synonym for another popular term, Knowledge Discovery in Database (KDD). Alternatively other people treat Data Mining as the core process of KDD. The KDD processes are shown in Figure 2.1 [118]. Usually there are three processes. One is called preprocessing, which is executed before data mining techniques are applied to the right data. The pre-processing includes data cleaning, integration, selection and transformation. The main process of KDD is the data mining process, in this process; different algorithms are applied to produce hidden knowledge. After that, comes another process called post processing, this evaluates the mining result according to user’s requirements and domain knowledge. Regarding the evaluation results, the knowledge can be presented if the result is satisfactory, otherwise we have to run some or all of those processes again until we get the satisfactory result. The actually process involves in below steps.

Initially database needs to be cleaned and integrated. Since the data source may come from different databases, which may have some inconsistencies and duplications, we must clean the data source by removing those noises or make some compromises. Suppose we have two different databases, different words are used to refer the same thing in their schema. When we try to integrate the two sources we can only choose one of them, if we know that they denote the same thing. And also real world data tend to be incomplete and noisy due to the manual input mistakes. The integrated data sources can be stored in a database, data warehouse or other repositories.

As not all the data in the database are related to our mining task, the second process is to select task related data from the integrated resources
and transform them into a format that is ready to be mined. Suppose we want to find which items are often purchased together in a supermarket, while the database that records the purchase history may contain *customer ID, items bought, transaction time, prices, number of each items* and so on, but for this specific task we only need *items bought*. After selection of relevant data, the database that we are going to apply our data mining techniques to will be much smaller, consequently the whole process will be more efficient.

Various data mining techniques are applied to the data source; different knowledge comes out as the mining result. That knowledge is evaluated by certain rules, such as the domain knowledge or concepts. After the evaluation, as shown in Figure 2.1, if the result does not satisfy the requirements or contradicts with the domain knowledge, we have to redo some processes until getting the right results. Depending on the evaluation result we may have to redo the mining or the user may modify his requirements. After we get the knowledge, the final step is to visualize the results. They can be displayed as raw data, tables, decision trees, rules, charts, data cubes or 3D graphics. This process is tries to make the data mining results easier to be used and more understandable.

### 2.1.1 Types of Data for Mining

Data mining tasks are of different types. These tasks vary with vary in application and type of data. These data can be categorized as follows:

**Relational database:** Round the world the mostly used database model is relational model. This is at-most simple as data is stored using rows and columns. Columns describe the properties of data. Relational databases are the biggest resource of data for mining. Optimized query languages helps
in easy mining of knowledge from such databases. Mining from such databases primarily relies on discovering features and development. Using aggregate operations in relational database, analysis becomes easy. For example, extracting information of sales department at regular intervals (monthly or yearly or daily). This helps in understanding the behavior of system before taking decisions.

**Transactional database:** Transactional database consists of records where each record refers to a transaction and a set of items associated with that transaction. Mining on such databases results in extracting frequent patterns. This helps in the areas of e-commerce etc. for taking apt decisions in-order to increase the market of the system and its products.

**Spatial database:** A spatial database generally contains information of geographic locations besides general information. Rules generated from these databases elucidates about the relationship between different features. The explanations of spatial association rules are quite similar to that of general association rules [7, 123]. These can take of the form \( X \Rightarrow Y \), where \( X, Y \) are sets of predicates. Among these set of predicates one predicate must be of type spatial [123, 124]. Algorithms developed under spatial association rule mining are identical to general association rule mining. But the generation of predicates and rules from spatial databases relies on constructs of algorithm Apriori. Mr. K. Koperski and J. Hanin in 1996 developed few algorithms for mining rules from spatial databases. These were used for the analysis of databases which contains large geographical data [124]. Geominer, an application, was developed in extracting rules from geographical databases [6].
**Temporal and time-series database:** Temporal data includes time as attribute in database. The difference between traditional transactional data and temporal data is: temporal data includes time as attribute whereas traditional transactional data doesn’t include time. Rules mining from temporal data are more instructive and valuable than rules mined from traditional transactional data. The rules generated from such databases are helpful for taking more promotion strategies in decision making systems.

Klemettinen.M et. al and Han. J developed algorithms for extracting periodic items and episode sequential patterns [125, 126]. Agrawal. R initiated most of the researchers and helped in extending the research in association mining called sequential mining associated with time [127].

**World-Wide Web:** At present the usage of web increased profoundly. This resulted in, the web to be abundant which extended the research of many researchers towards mining from web called web mining. Web mining is classified into three major categories. They are:

- **Web usage mining:** This concentrates on navigation patterns of users on web. Many mining applications that analyses logs of web were developed to understand the navigation patterns of users. Mining such patterns from web helps in modification of web applications in order to increase the effectiveness of web like promoting advertisements on web, including marketeering strategies etc.

- **Web structure mining:** This classifies web documents two types of pages. They are: authoritative and hub. Authoritative pages have original source of information,
whereas hub pages are pages that link to those authoritative pages. This mining mainly relies on structures and links of web documents. Links between web pages indicates the relationship between web pages.

- **Web content mining**: This includes mining of contents of web like images, text, graphics, media etc.. Such mining are termed as text mining, graphic mining & multimedia mining.

### 2.1.2 Types of Mining Techniques

Data Mining is classified into two types. They are: *descriptive data mining* and *prescriptive data mining*. Descriptive mining summarizes the wide-ranging properties data in data repositories. Prescriptive data mining uses historical data to predict interpretations for current data. Classifications, clustering and association rules are various data mining techniques. Besides these techniques sequential mining and web mining also extended their research.

*Association rule mining* was first introduced by R. Agrawal, T. Imielinski, and A. Swami in 1993 at International Conference on Management of Data, ACM SIGMOD [7]. It is significant among the data mining techniques. This discovers frequent items, associations and correlations among items in a transactional database.

The discovered rule helps in taking decisions in a decision making system. For example, in market analysis, identifying frequent patterns helps in improving the profitability of the system.
Application areas of association rule mining are telecommunication networks, market and risk management, inventory control etc. Brief history of association mining algorithms is elaborated in this chapter.

**Classification** is a model that is able to classify objects and predict the missing or attribute value of a future object is known as classification [118]. It involves two processes. The first one is building a model to describe the characteristics of a set of data classes. As the data classes are predefined this is termed as supervised learning. While the second process, involves in prediction of future objects or data.

Classification [118] by decision trees has been studied and a number of algorithms had been designed on it. On decision tree induction a comprehensive survey was done by S. K. Murthy in 1998 [119]. Bayesian classification is another method devised by Duda and Hart [120]. Different methods of classification too were devised by Duda and Hart [120] and James [121] like nearest neighbor methods. In addition to these, neural network and machine learning methods are methods in classification.

For example consider which type of customer would buy a particular laptop. A sample decision trees to the mentioned example is shown in Figure 2.2. The internal nodes of decision tree carries out a decision based on the attribute value and the leaf node decides the class (YES or NO).
Clustering is the process of combining a set of physical or abstract objects into classes of similar objects but separate from other clusters [118]. While classification is termed as supervised learning process clustering is an unsupervised learning process. The difference between classification and clustering is that in the former the record and the class it belongs to is predefined, while in the latter there is no predefined classes. In clustering, objects are grouped based on their similarities. These are quantitatively specified as distance or other measures.

Clustering applications are more often seen in market segmentation. Clustering organizes customers into various groups such that a system can adapt feasible services to each individual group for increasing the profitability of the system. A detailed study of existing clustering methods is given by P. Berkhin in 2002 [122].

2.2 Association Rule Mining

Association Rule Mining has become an emerging area of research. During the process of Knowledge Discovery Association Rule Mining has
its own significance. It still is in its nascent stage but has a broad scope in utilization. This area has been developing for the last fifteen years vigorously.

Initially Association Rule Mining (ARM) was proposed by R. Agrawal, T. Imielinski, and A. Swami in 1993 [7]. For discovering frequent patterns or association rules a transactional database is needed. A transactional database contains set of items (I) and set of transactions (T). A transaction contains subset of set of items. Items are otherwise called as patterns. A rule takes of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = 0$. Rules are just like implications. Left hand side of implication is called antecedent and right hand side of implication is called consequent. The primary algorithm designed was AIS algorithm [7].

2.3 Process of Association Rule Mining

Association rules in general must satisfy two user-defined values: support and confidence.

Association rules are generated in two steps:

Step 1: Transactional database is applied with support value in-order to discover frequent patterns.

Step 2: Frequent patterns are validated using confidence value for generating rules.

2.4 History of Association Rule Mining Algorithms

Association rule mining became the most prevailing technique in discovering frequent items. This section extensively elaborates about the
various algorithms developed particularly for discovering frequent items over a period till date.

Agrawal R., Imielinsky T. and Swami A. developed the first algorithm, *AIS algorithm* for discovering frequent patterns in 1993. In this algorithm association rules are generated with only one item in consequent. For instance rules can be like $X \cap Y \Rightarrow Z$ but not like $X \Rightarrow Y \cap Z$. In AIS algorithm, multiple scanning is used to generate the frequent item sets. Each individual item along with its respective support count is stored during the initial pass over the transactional database followed by excluding the items whose count value is less than the user defined support value. The frequent 1-item set is extended to generate candidate 2-item set. During the second pass of the algorithm candidate 2-item sets are validated against the user defined support value. While validating candidates whose frequency is less than support value are excluded from the set and remaining candidates are considered for further process, results in generating frequent 2-item set. This whole process is repeated until no candidates in further iterations have frequency greater than or equal to support value. Figure 2.3 illustrates the demonstration of AIS algorithm with a simple dataset. One major pitfall for this algorithm is more candidates of smaller size are generated results in non-optimal usage of space and time with multiple passes over the dataset.
Houtsma M. and Swami A. in 1993 presented SETM a more effective than AIS. Candidates are generated as and when the database is scanned. Later new candidates are generated in the similar way as in AIS algorithm, but the transaction identifier TID of the generating transaction is saved with the candidate in a sequential structure. This process results in candidate generation process differing from that of counting. This results in aggregating the sequential structure at the end of the pass in-order to get support count of candidates. The SETM algorithm and AIS algorithm face similar disadvantage that for each candidate item set, there are as many entries as its support value. SETM algorithm demonstrated with an example is shown Figure 2.4.
In 1994, Agrawal R and Srikant R, presented novel algorithms, Apriori and AprioriTID [18]. Apriori algorithm is used for discovering frequent patterns. The algorithm uses Boolean association rules to mine frequent patterns from a transactional database. Level-wise search, where k-item sets (An item set which contains k items is known as k-item set) are used to explore (k+1)-item sets. This algorithm uses candidate generation process where in-frequent patterns are excluded one item at a time. In order to count the item sets efficiently groups of candidates are tested against the data. Apriori uses breadth-first search method and a hash tree structure to
fulfill this. At first it discovers frequent individual patterns and then extended to larger sets. But Apriori algorithm has its own drawback that of over utilization of space and memory while generating complex candidates. Another drawback is redundant scans of database. Apriori algorithm is demonstrated with an example is shown in Figure 2.5.

<table>
<thead>
<tr>
<th>Items</th>
<th>Count Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>7</td>
</tr>
<tr>
<td>I2</td>
<td>8</td>
</tr>
<tr>
<td>I3</td>
<td>6</td>
</tr>
<tr>
<td>I4</td>
<td>2</td>
</tr>
<tr>
<td>I5</td>
<td>3</td>
</tr>
<tr>
<td>I6</td>
<td>1</td>
</tr>
</tbody>
</table>

(C1)

<table>
<thead>
<tr>
<th>Items</th>
<th>Count Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1, I2</td>
<td>5</td>
</tr>
<tr>
<td>I1, I3</td>
<td>4</td>
</tr>
<tr>
<td>I1, I5</td>
<td>3</td>
</tr>
<tr>
<td>I2, I3</td>
<td>4</td>
</tr>
<tr>
<td>I2, I5</td>
<td>3</td>
</tr>
<tr>
<td>I3, I5</td>
<td>1</td>
</tr>
</tbody>
</table>

(C2)

<table>
<thead>
<tr>
<th>Items</th>
<th>Count Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1, I2, I5</td>
<td>3</td>
</tr>
<tr>
<td>I1, I2, I3</td>
<td>8</td>
</tr>
</tbody>
</table>

(C3)

Figure 2.5: APRIORI algorithm with an example

In Apriori, after first pass transactional database is not used for counting the support of candidates. It is similar to that of Apriori algorithm. In-order to incorporate TID of each transaction another set C is generated. The generated set i.e. C is used to count the support of each candidate. This
algorithm has performance levels over the later passes in comparison to Apriori. The algorithm is demonstrated in Figure 2.6.

Figure 2.6: Aprioritid with an example

AprioriHybrid, a new algorithm has incorporated the best of Apriori and Aprioritid algorithms. Its advantages are that of excellent scale-up properties and feasibility of mining association rules over very large database. In earlier passes it uses Apriori algorithm and in later passes it uses Aprioritid algorithm. Different aspects of database mining problems were
examined and this relation the algorithms were developed at the IBM Almaden Research Center. Apriori is a traditional algorithm for discovering frequent patterns and were applied on transactional databases.

A new algorithm [57] had been devised by Mannila H, Toivonen H and Verkamo A.I, to eliminate redundant candidate rules. This is based on the collective analysis information of the previous passes that has been obtained.

In 1995, Mueller A., [60] extended his research in sequential algorithms. He devised an incremental portioning algorithm with slender enhancement in the classical portioning techniques. He evaluated the proficiency of his algorithm through which he proved that cost of CPU and IO are reduced. Many researchers have contributed to this area. One of the initial efforts was by Usama Fayyad et al. [17, 34] presented a framework, described associations among data mining, knowledge discovery and other associated fields.

Park J.S. et al. [63] with bendable accuracy developed an algorithm for discovering frequent patterns. This had two methods for discovering frequent patterns. Both the methods obtain same result over a sampled dataset. For obtaining preferred level of accuracy, value of support must be relaxed based on the size of sample.

Cheung D.W. et al. [28] proposed a fast distributed association rule mining algorithm. During execution of this algorithm less number of candidates is generated and lowers the quantity of messages to be passed. The success rate of this algorithm is high when compared with existed sequential algorithms.
Shintani T. et al. [75] presented four hash based parallel algorithms for association rule mining. The limitation of these was they could be processed on parallel only.

Meo R. et al. [59] invented a new operator similar to SQL. This operator is named as \textit{MINE RULE} introduced a new SQL-like operator. This novel concept cleared the problems associated with rule mining.

In 1997, Thomas S. et al. [79] devised an incremental model. This method works upon addition or deletion of transactions from the database. This method with minimum re-computation extracts different large patterns. Also reduces the overhead of I/O operations while updating the databases with transactions.

M.J.Zaki et al. [86] proposed algorithm for fast discovery of frequent patterns. This uses the structural properties of frequent patterns. All patterns in the database are grouped into clusters. Each cluster denotes frequent pattern set of maximum size. Also each cluster prompts a sub lattice. Dominating lattice traversing methods are applied for effective extraction of frequent patterns.

Brin S. et al. [26] generalized association rules by his proposed algorithm. These rules are called as correlation rules. These play an important role resulting in an application extending standard market-basket setting. For analyzing extensive data Chi square statistics are used.

In 1998, Bing Liu et al. [25] devised association rule mining algorithm based on two techniques: integration and classification. Integration relies on discovering subset of frequent items called Class
Association Rules (CARs). This had resulted in an efficient algorithm for building a classifier based on the set of discovered CARs.

Kai C.H. et al. [50] explored a new thought of discovering frequent patterns. He used the concept of weighted patterns. According to his proposed method patterns are assigned with some weights. To increase the market of patterns or items in a market weights are assigned to items. This resulted in discovery of frequent patterns with weights and showed a new era of research.

Roberto J. Bayardo Jr. [71] devised an algorithm for discovering frequent patterns which are long. The algorithm is proportional to the number of patterns existing in the database rather than the length of longest frequent pattern. This algorithm is named as Max-Miner. Pruning in this algorithm is done by using superset frequency. This reduced the space for considered patterns in the database.

Gardarin G. et al. [36] developed an association rule mining algorithm using bitmap. The concept behind the algorithm is, every pair (transaction–item) is denoted by a bit in an index bitmap and a logical operation AND in lieu of the sort-merge algorithm. The two deviations of this type are: hierarchical bitmap algorithm (H-BM) and naïve bitmap algorithm (N-BM).

Jun Lin et al. [53] developed an anti-skew algorithm with an aim in reducing the number of scans over database. This algorithm uses prior information for discovering rules. This prior information is obtained during sampling the database. This reduces the generation of candidate and eliminates unfair candidates at former stage. In this method maximum number of scans is reduced up to 2.
Hipp J. et al. [42] developed a new algorithm for discovering common frequent patterns named Prutax. Despite being a novel approach it did not yield the expected results.

In 1999, Roberto J. Bayardo Jr. and Rakesh Agrawal [23] proposed an optimized association rule mining problem. In place of total order partial order on rules are permitted in this. Their study focused on the best rules mined which lie at the border support and confidence values.

Most of the association algorithms developed discovers rules using customized support values defined by user. Later Bing Liu et al. [24] proposed notion of discovering frequent patterns using Multiple Minimum Supports. This technique involves the user for customizing multiple support values to the frequency and nature of patterns.

Cohen E. et al. [30] tried to discover interesting frequent patterns without support pruning. They used a combination of sampling and hashing techniques to obtain the result.

In 2000, Webb G.I. [81] tried to mine frequent patterns in single stage unlike in two stage process like Apriori algorithm. He stated that Apriori algorithm involves in bulk number of computations when the database is enormous.

Zaki M.J. [87] proposed scalable algorithms for fast discovery of frequent patterns which uses structural properties of frequent patterns.

Zaki M.J. [88] devised SPADE, an algorithm for fast mining of frequent Sequential Patterns. SPADE divides the main problem into sub problems using combinatorial features. The sub-problems have the
capability of getting solved independently. Also these sub-problems doesn’t rely on main memory and customs simple join operations and proficient lattice search techniques. Frequent sequences are identified by using three scans of database.

Han J. et al. [40] proposed a novel method for discovering frequent patterns. This uses special structure, FP-Tree. This method is free of candidate generation. FP-Tree is a prefix tree which holds the vital information of frequent patterns. This notion resulted in developing the algorithm called FP-Growth. FP-Growth uses pattern growth fragment for discovering frequent patterns. The competence of FP-Growth algorithm is obtained by three procedures:

i. Compressed database which reduces redundancy of database scans.
ii. Adoption of pattern-fragment growth by passes the expensive generation of a large number of candidate sets.
iii. Splitting of the mining tasks into a smaller tasks using divide and conquer policy restricts frequent patterns to be mined in conditional databases. Thus results in reduction of search space.

Hipp J. et al. [43] revealed a comparison of different algorithms for association rule mining.

Pudi V. et al. [66] incorporated some change in order to improve the efficiency of the existing association mining algorithms.

Jiangping Chen [47] proposed an association mining algorithm for
spatial databases. This uses autocorrelation with a cell structure theory. The autocorrelation of spatial data is calculated by using special algebraic data structure.

In 2002, Hegland M. [41] devised an improvement on the basic Apriori algorithm with variants for distributed data, constraints inclusion and taxonomies of data. The advantage of this method is its ability to deal with lengthy pattern sets and reduce amount of patterns returning.

Delic D. et al. [32] evaluated the rough set and the association rule method. Despite their different approaches both these methods were based on the same principle as such generate identical rules. But they do differ in the aspect of efficiency. In addition optimized association rule has been developed which unites the benefits of both the approaches.

Jian Pei [46] doctoral thesis title is “Pattern-Growth methods for frequent Pattern Mining”. For discovering frequent patterns from large databases he developed new methods. One among them is FP-Growth which is still noted an efficient algorithm.

Goethals B. [37] submitted his Ph.D. thesis on “Efficient Pattern Mining”. In this he compared different approaches with a given set of frequent patterns on the largest size of possible candidate pattern. The outcome was a hard and tight combinatorial upper bound on the number of candidate patterns. His verdicts are not limited to a single algorithm. Based on the level wise generation of candidates this can be applied on any rule mining algorithms. Pudi V et al. [62, 63, 64] paralleled the competence of some present algorithms against ARMOR which was developed by him.
In 2003, Cheung W [27] proposed a novel data structure called CATS Tree. This was introduced in his master degree thesis titled “Frequent Pattern Mining without Candidate Generation or Support Constraint”. CATS Tree is an enhancement to FP-Tree. This improves the association rules extraction process by avoiding the generation of candidates. Also improves the storage density. The proposed procedure discovers frequent patterns using one pass over the database.

Tao F. et al. [78] revealed the issue related to noteworthy binary associations in transactional database in a weighted setting. A classical approach of discovering frequent patterns for handling weighted associations is used where each pattern was given a weight. The goal is to extract important rules with patterns having substantial weight rather than extracting petty rules.

Shoemaker C.A. et al. [29] deployed a scheme for discovering frequent patterns handles set-valued properties. From set valued datasets two methods of association mining were executed for calculating the efficiency. This devised the system in extracting rules directly from set valued datasets and also easy creation of input data.

In 2004, Song M. et al. [77] proposed an innovative algorithm called Transaction Mapping (TM). This discovers complete frequent patterns. This algorithm maps ids of transactions of each pattern and compresses in a different space at intervals of continuous transactions. The count of patterns is calculated by the intersection of these intervals using lexicographic tree in a depth first order. The coefficient of compression becomes minimum when the average value of intersection of intervals comparisons stretch a
convinced level, results in switching to transaction id intersection.

Palshikar G.K. et al. [62] used heavy patterns and designed association rule mining algorithms.

Flank A. [35] designed multi-relational association rule mining schemes in relational data repositories. The rules and queries are formatted using simple English.

Győrödi C. et al. [39] revealed about the major differences between the algorithms which discovers frequent patterns using candidate generation and without candidate generation.

Guil F. et al. [38] presented a new algorithm named TSET (as acronym of Temporal Set-Enumeration Tree) for frequent temporal pattern (sequences) mining from datasets. This algorithm uses a distinct tree structure, used for keeping all frequent patterns discovering during the process of knowledge mining.

Xia Y. et al. [83] proposed an efficient algorithm RE (Recursive Estimation) to estimate the support of item sets. This introduces different levels of privacy by different attributes i.e. different attribute values are obtained by using randomization factors.

Ashrafi M.Z. et al. [21] designed an association mining algorithm for geographical databases, called Optimized Distributed Association Mining. In contrast to DARM, ODAM calculates the support of candidates faster. Thus reduces the average number of transactions, size of datasets and messages to be exchanged.
In 2005, Jotwani N. [49] proposed an algorithm based on hierarchical classification of patterns. The algorithm executes the first phase efficiently in single pass with tight bounds on the computational effort and uncertain memory requirements. Thus the algorithm is capable of online discovery of frequent patterns from transaction streams.

Salleb-Aouissi A. et al. [72] designed a Genetic Algorithm for Mining Quantitative Association Rules called QuantMiner. By optimizing the values confidence and support this system identifies optimal intervals in association rules. Few modifications to this algorithm are done by Badawy o. et al. [22]. This algorithm was used by Alfatly E.K.J. [20] in his thesis and introduced new algorithms.

Liu P. et al. [54] used FAS algorithm and designed a knowledge discovery system called AR-Miner. This system is simple because of its easiness of its interface for discovery of frequent items.

Margahny M.H. et al. [58] proposed an association rule mining algorithm which uses ArrayList. This algorithm reduces the traverse of the database.

In 2006, Nan Jiang et al. [61] deliberated dissimilar research concerns in data stream mining. These relate in mining frequent patterns from different applications like monitoring network traffic and web click stream analysis. Data from streams (dynamic) arrive with greater speed and bulk in size continuously when compared to traditional static databases. Data streams change with change in time.

Ishibuchi H. et al. [44] debated the application of evolutionary multi-
objective optimization (EMO) to association rule mining. Evolutionary multi-objective rule selection is termed as a post-in processing in procedure in classification rule mining. Optimal rule sets are discovered from a database using an association rule mining technique.

Kotsiantis S. et al. [51] study covers major theoretical issues of different concepts and techniques in association rule mining.

Kumar K. B. et al. [52] developed an association rule mining algorithm with single pass over database. This algorithm is provided with patterns which are clustered hierarchically. The efficiency of this algorithm, relies on reduction of transaction and hierarchy-aware counting.

Dehuri S. et al. [31] developed association rule algorithms using genetic algorithms for discovering rules from large repositories. These algorithms are fast and multi-objective. The basic objective of association rule mining algorithms are: interestingness of frequent patterns, confidence value and plainness. These are treated as elementary inputs to the algorithms. Set of non-dominated resolutions are the outcomes. In knowledge discovery process the size and dimension of data increases with change in time. The algorithm MOGA (multi-objective genetic algorithm) is slow when compared with classical algorithms. To overcome these pitfalls methods which are parallel in nature which are genetic and homogeneous are developed.

In 2007, Mangalampalli A. et al. [56] detailed a procedure with Fuzzy Logic-based Pre-In Processing for discovering frequent patterns.

Pasquier N. et al. [64] proposed A-Close which discovers only needed
rules without loss of information. This algorithm overcomes the pitfall of algorithms which uses lattice framework for finding frequent patterns. The search space for mining frequent patterns is constricted to closed item set. Webb G.I. [79] used statistical measures for identifying frequent items.

Jianwei Li et al. [48] proposed some algorithms which discover both frequent patterns and clusters in parallel. These methods discover groups which are frequent. K-means and hierarchical algorithms are used in parallel for discover groups which are the most frequent.

In 2008, Sheibani R. et al. [74] proposed association rule mining method based on Fuzzy Cluster.

Zaki M.J. [85] devised a method for closed mining of association rules. This is named as CHARM. This proposed method initially doesn’t mine all frequent patterns. This only discovers closed frequent patterns. The size of set of closed frequent patterns is less than size of set of all frequent patterns. Moreover rules which are significant are mined. Repeated rules are excluded.

The FCBAR method scans the database once. After scanning tables are clustered. Then for each $k^{th}$ cluster table records of transactions are clustered. The length of record is $k$. In contrast the partial cluster tables generate fuzzy large patterns. Thus large amount of data is pruned in-turn amount of time spend in scanning the database is reduced.

Ramraj E. et al. [68] proposed some methods with an aim of enhancing the quality of data mined.

Mangalampalli A. et al. [55] developed a new association rule mining
method based on fuzzy. This is faster and more advanced when compared with his earlier developed methods. This proposed method works on large databases.

Wu Jian et al. [82] using item sequence sets (ISS) devised methods for discovering weighted rules.

In 2009, Yang J. et al. [84] surveyed on association rule mining methods in extraction of positive and negative rules.

In 2010, Piao et al. [65] conducted research on mining negative and positive association rules. A correlation measure in conjunction with mining algorithm was used for association rule. This algorithm mines interesting negative patterns besides excluding worthless rules.

Umaraani et al. [80] proposed an implementable technique to prune datasets. This was a result of a study on a huge database of association rules.

Veeramalai et al. devised an algorithm called Fuzzy-T ARM to extract rules from the dataset, breast cancer. This method adapts datasets which contains multidimensional data with different representation.

Wang P. expanded research in discovering rules from privacy preserving systems and increased the thirst of mining rules in this area.

In 2011, Demitrijevic et al. [33] introduced method for discovering frequent patterns on web log usage data called web usage mining. They implemented a system to experiment on a real life web usage log data set. This would help in object oriented applications.

Sandhu et al. [73] proposed a new method which produced qualitative
rules for finding frequent patterns. These methods depend on quantity and profit. Wang H. et al. defined an algorithm to cut down the blockages of Apriori algorithm.

In 2012, Jadav et al. [45] defined some methods using association rule mining on OLAP cube to achieve accurate results.

Radhka et al. [67] explored research on ontological weights for discovering frequent patterns. This approach tries to cut down and screen discovered rules on the origin of ontological weights.

Raorane et al. [69] research helped in advancing the effectiveness of association rule mining over conventional methods using a market-basket analysis.

Raosulian et al. [70] explored the effect of data mining base on association rule mining in strategic management resulting in an effective mode.

Shrivastava et al. proposed two methods in discovering association rules [76]. They are: non-redundant methods and non-redundant sequential methods for discovering frequent patterns.

2.5 Problems in Association Rule Mining

Association rule mining is still in a nascent stage and needs to be explored and developed. The potential areas in which association rule mining needs to be looked into are the following:

a. Vast research is needed to develop procedures of pattern-based discovery.
b. A deeper understanding of semantic annotation for frequent patterns and contextual analysis of the same is needed so as to interpret the results effectively.

c. Generalized approach rather than the assumptive approach would be more practical so as to make the application universal.

d. Scalable and effective methods need to be developed for data mining.

e. Measurements need to be independent of data base.

f. In order to identify patterns from data streams single scan and online mining methods need to be developed.

g. Efficient systems need to be developed in hidden pattern methods.

h. Knowledge patterns could be derived from data streams with the help of single scan and online mining methods.

i. Measurements which are independent of databases should be devised.

j. Systems with efficient and effective hidden patterns need to be developed.

k. Identification of deep-level association should be done.

l. Multi databases should be explored in order to find new techniques for mining association.

m. An efficient algorithm for data mining on XML databases needs to be developed.

n. Techniques for web usage mining needs to be developed.

o. Social network analysis and mining for business needs to be correlated with association rule mining.

p. Bug detection and analyses are an important aspect of software industry. Frequent pattern mining plays a major role in this aspect.
However more effective and efficient methods are needed so as to design an error free software.

2.6 How to improve the Proficiency of Association Rule Mining Algorithms?

There are four different ways in enhancing the ability of association rule mining algorithms. They are:

i. Bypassing the number of passes over database
ii. Parallelization of discovering rules
iii. Database sampling
iv. Patterns imposed with surplus constrictions

2.6.1 Reducing the number of passes over the database

The drawbacks of Apriori had been vanquished With the FP-Tree a new breakthrough in the development of association rule mining [89]. The resultant frequent item set are generated with only two passes over the database.

FP-tree is an extended prefix-tree structure storing crucial, quantitative information about frequent patterns. Only frequent length-1 items will have nodes in the tree, and the tree nodes are arranged in such a way that more frequently occurring nodes will have better chances of sharing nodes than less frequently occurring ones.

FP-Tree scales much better than Apriori because as the support threshold goes down, the number as well as the length of frequent item sets...
increase dramatically. The candidate sets that Apriori must handle become extremely large, and the pattern matching with a lot of candidates by searching through the transactions becomes very expensive.

Apriori series and FP-tree algorithm result in the same mining result. The result of FP-tree algorithm is derived by two sub processes: one is the construction of the FT-tree itself and the second that of generating frequent patterns from the same.

FP-tree algorithm has been termed as an efficient process for the following reasons. These algorithms take only the relevant information hence redundancy is eliminated. The divide and conquer method well as the double cross checking it uses results in reducing the conditional FP-tree algorithm. The drawback of this algorithm is it cannot be used in interactive data mining as change of support results in data redundancy [89]. It is not suitable for incremental mining also. When new datasets are introduced into the databases, this would result in repetition.

Tree Projection is another efficient algorithm recently proposed in [90]. The general idea of Tree Projection is that it constructs a lexicographical tree and projects a large database into a set of reduced, item-based sub-databases based on the frequent patterns mined so far. The number of nodes in its lexicographic tree is exactly that of the frequent item sets. The efficiency of Tree Projection can be explained by two main factors: (1) the transaction projection limits the support counting in a relatively small space; and (2) the lexicographical tree facilitates the management and counting of candidates and provides the flexibility of picking efficient strategy during the tree generation and transaction projection phrases.
Wang and Tjortjis [91] presented PRICES, an efficient algorithm for mining association rules. Their approach reduces large item set generation time, known to be the most time-consuming step, by scanning the database only once and using logical operations in the process. Another algorithm for efficient generating large frequent candidate sets is proposed by [92], which is called Matrix Algorithm. The algorithm generates a matrix which entries 1 or 0 by passing over the cruel database only once, and then the frequent candidate sets are obtained from the resulting matrix. Finally association rules are mined from the frequent candidate sets. Experiments results confirm that the proposed algorithm is more effective than Apriori Algorithm.

2.6.2 Sampling

Toivonen [93] presented an association rule mining algorithm using sampling. The approach can be divided into two phases. During phase 1 a sample of the database is obtained and all associations in the sample are found. These results are then validated against the entire database. To maximize the effectiveness of the overall approach, the author makes use of lowered minimum support on the sample. Since the approach is probabilistic (i.e., dependent on the sample containing all the relevant associations) not all the rules may be found in this first pass. Those associations that were deemed not frequent in the sample but were actually frequent in the entire dataset are used to construct the complete set of associations in phase 2.

Parthasarathy [94] presented an efficient method to progressively sample for association rules. His approach relies on a novel measure of model accuracy (self-similarity of associations across progressive samples),
the identification of a representative class of frequent item sets that mimic (extremely accurately) the self-similarity values across the entire set of associations, and an efficient sampling methodology that hides the overhead of obtaining progressive samples by overlapping it with useful computation.

Chuang et al. [95] explore another progressive sampling algorithm, called Sampling Error Estimation (SEE), which aims to identify an appropriate sample size for mining association rules. SEE has two advantages. First, SEE is highly efficient because an appropriate sample size can be determined without the need of executing association rules. Second, the identified sample size of SEE is very accurate, meaning that association rules can be highly efficiently executed on a sample of this size to obtain a sufficiently accurate result.

Especially, if data comes as a stream flowing at a faster rate than can be processed, sampling seems to be the only choice. How to sample the data and how big the sample size should be for a given error bound and confidence levels are key issues for particular data mining tasks. Li and Gopalan [96] derive the sufficient sample size based on central limit theorem for sampling large datasets with replacement.

**2.6.3 Parallelization**

Association rule discovery techniques have gradually been adapted to parallel systems in order to take advantage of the higher speed and greater storage capacity that they offer [97]. The transition to a distributed memory system requires the partitioning of the database among the processors, a procedure that is generally carried out indiscriminately.
Cheung et al. [98] presented an algorithm called FDM. FDM is a parallelization of Apriori to (shared nothing machines, each with its own partition of the database. At every level and on each machine, the database scan is performed independently on the local partition. Then a distributed pruning technique is employed. Schuster and Wolff [99] described another Apriori based D-ARM algorithm - DDM. As in FDM, candidates in DDM are generated level wise and are then counted by each node in its local database. The nodes then perform a distributed decision protocol in order to find out which of the candidates are frequent and which are not.

Another efficient parallel algorithm FPM (Fast Parallel Mining) for mining association rules on a shared-nothing parallel system has been proposed by [100]. It adopts the count distribution approach and has incorporated two powerful candidate pruning techniques, i.e., distributed pruning and global pruning. It has a simple communication scheme which performs only one round of message exchange in each iteration. A new algorithm, Data Allocation Algorithm (DAA), is presented in [101] that uses Principal Component Analysis to improve the data distribution prior to FPM.

Parthasarathy et al. [102] have written an excellent recent survey on parallel association rule mining with shared memory architecture covering most trends, challenges and approaches adopted for parallel data mining. All approaches spelled out and compared in this extensive survey are apriori-based. More recently, Tang and Turkia [103] proposed a parallelization scheme which can be used to parallelize the efficient and fast frequent item set mining algorithms based on FP-trees.
2.6.4 Constraints based Association Rule Mining

Many data mining techniques consist in discovering patterns frequently occurring in the source dataset. Typically, the goal is to discover all the patterns whose frequency in the dataset exceeds a user-specified threshold. However, very often users want to restrict the set of patterns to be discovered by adding extra constraints on the structure of patterns. Data mining systems should be able to exploit such constraints to speed-up the mining process. Techniques applicable to constraint-driven pattern discovery can be classified into the following groups:

• post-processing (filtering out patterns that do not satisfy user-specified pattern constraints after the actual discovery process);

• pattern filtering (integration of pattern constraints into the actual mining process in order to generate only patterns satisfying the constraints);

• dataset filtering (restricting the source dataset to objects that can possibly contain patterns that satisfy pattern constraints).

Wojciechowski and Zakrzewicz [104] focus on improving the efficiency of constraint-based frequent pattern mining by using dataset filtering techniques. Dataset filtering conceptually transforms a given data mining task into an equivalent one operating on a smaller dataset. Tien Dung Do et al [105] proposed a specific type of constraints called category-based as well as the associated algorithm for constrained rule mining based on Apriori. The Category-based Apriori algorithm reduces the computational complexity of the mining process by bypassing most of the subsets of the final itemsets. An experiment has been conducted to show the efficiency of
the proposed technique.

Rapid Association Rule Mining (RARM) [106] method uses the data structure tree for representing data. This avoids the generation of candidates. The proficiency of existing association rule mining algorithms relies on the degree of extracting rules, which are important for analysis rather than extracting all rules.

2.7 Application Areas for Association Mining

The various application areas in which association rules can be applied for extracting useful information from the huge dataset are:

2.7.1 Market basket analysis

A typical and widely-used example of association rule mining is market basket analysis. For example, data are collected using bar-code scanners in supermarkets. Such ‘market basket’ databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together. They could use this data for adjusting store layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalog design and to identify customer segments based on buying patterns [113].

For example, if a supermarket database has 100,000 point-of-sale transactions, out of which 2,000 include both items A and B and 800 of these include item C, the association rule "If A and B are purchased then C is purchased on the same trip" has a support of 800 transactions (alternatively $0.8\% = \frac{800}{100,000}$) and a confidence of 40% ($\frac{800}{2,000}$).
One way to think of support is that it is the probability that a randomly selected transaction from the database will contain all items in the antecedent and the consequent, whereas the confidence is the conditional probability that a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent.

Now days every product comes with bar code. The software supporting these barcode based purchasing/ordering systems produces vast amounts of sales data, typically captured in “baskets” (records in which the items purchased by a given consumer at a given time are grouped together). This data was quickly recognized by the business world as having immense potential value in marketing. In particular, commercial organizations are interested in discovering “association rules” that identify patterns of purchases, such that the presence of one item in a basket will imply the presence of one or more additional items. This “market basket analysis” result can then be used to suggest combinations of products for special promotions or sales, devise a more effective store layout, and give insight into brand loyalty and co-branding.

2.7.2 Medical diagnosis

Association rules generated from medical data sets helps in aiding the patients who are at remote locations. They provide needed guidance and can assist a patient to have treatment practically. These rules are generated based on induction; hence they can aid and never ensures apt diagnosis [111].

The training data set and its associated rules always may not diagnose the disease exactly. This is because of variant values obtained for same
medical tests by different labs. Hence the association rules obtained from such data sets is not feasible for generating fruitful prediction [108].

Based on Relational association rules and supervised learning methods a new technique was proposed by Serban [111]. This technique could be used to devise a new interface. This would in turn broadly define the correlation between certain illnesses and new symptoms for a given disease.

2.7.3 Protein Sequences

Proteins are important constituents of cellular machinery of any organism. Recombinant DNA technologies have provided tools for the rapid determination of DNA sequences and, by inference, the amino acid sequences of proteins from structural genes [107].

Proteins are sequences made up of 20 types of amino acids. Each protein has a unique 3-dimensional structure, which depends on amino-acid sequence; slight change in sequence may change the functioning of protein. The heavy dependence of protein functioning on its amino acid sequence has been a subject of great anxiety.

Lot of research has gone into understanding the composition and nature of proteins; still many things remain to be understood satisfactorily. It is now generally believed that amino acid sequences of proteins are not random.

Nitin Gupta, Nitin Mangal, Kamal Tiwari, and Pabitra Mitra [115] have deciphered the nature of associations between different amino acids that are present in a protein. Such association rules are desirable for
enhancing our understanding of protein composition and hold the potential to give clues regarding the global interactions amongst some particular sets of amino acids occurring in proteins. Knowledge of these association rules or constraints is highly desirable for synthesis of artificial proteins.

2.7.4 Census Data

Statistical information is made available to researchers as well the general public by means of census [109]. These are used in planning policy measures (education, health, transport etc.) and also for setting public business like factories, shopping malls, or banks.

Effective public policy can be implemented through clubbing data mining techniques with census data for an effective implementation of the same [110]. But these are still in its infancy and would require a more in depth study in order to be effectively implemented.

2.7.5 CRM of Credit Card Business

Data mining has been used to ascertain the preferences of varied customer base and tailor the needs of these to improve the service to Customer Relationship Management (CRM) like the credit card customers [116]. Shaw had devised an effective way of incorporating Data mining to enhance the service given by banks to its customers [114].

Attainment of quality service to the customers as well as knowledge management is strengthened by means of collective application of association rule technique. These have been applied and a new method was proposed by Song [112] to observe changes in customer behavior based on customer profiles as well as sales data.
In order to implement rule matching the changes from two different datasets and rules generated have to be collected.