Chapter 3

Problem Statement and Research Methodology

Research is a process, which is applied to find unexplored aspects regarding a conception. The most concerned subject of research is problem formulation and decision about the research methodology. A topic selection is followed by the comprehensive literature survey which helps in guiding the problem formulation. After setting the research objectives, one can focus on them and carry out the research.

This chapter is organized as follows: Section 3.1 highlights the objectives of this research work, Section 3.2 explores the significance of study, Section 3.3 gives the research methodology to address the identified objectives and Section 3.4 provides the outcome of chapter.

### 3.1 Research Objectives

From the comprehensive review of existing literature, it was concluded that single level association rules mining is useful only when the user require only general information about data items but these rules are not sufficient to provide specific information. To explore more specific and precise information, association rules should be generated at multiple levels of abstraction. There are various multiple level association rules mining algorithms which are offered by several researchers. Various existing algorithms are applicable to discover multiple level association rules but still, there is the scope of improvement in terms of space and time complexity. To fill this gap, there is the requirement of a straightforward approach which is able to improve the time and storage efficiency. So, in view of the facts stated above, the researcher wants to carry out the research in this direction with following objectives:

1. **Comparative analysis of existing algorithms for mining frequent patterns at multiple levels.**
   
   Initially, existing multiple levels association rules mining algorithms were studied and a theoretical comparison is performed among them on the basis of various parameters like minimum support threshold, minimum confidence, execution time and number of frequent patterns. This critical analysis is revealed in Section 4.1 of chapter 4.

2. **To convert transactional tables into encoded tables by applying efficient encoding method.**
To discover frequent patterns using association rules at multiple level, there is the requirement of a hierarchical encoded transaction table rather than the simple transaction table. The reason behind this was that the encoded transaction table provides the information regarding at item’s level. Therefore, an encoding method, which was proposed by an existing researcher (Han and Fu, 1995), is used to convert the transaction tables into encoded tables. This conversion has been elaborated in Section 4.4 of Chapter 4.

III. **To develop/design an algorithm for multiple level association rules to reduce the number of iteration and to achieve time efficiency.**

After concluding the present research as mentioned in objective I, it was identified that a time efficient algorithm required for generation of multiple level association rules. To accomplish the time efficiency, an algorithm for multiple level association rules was generated using object oriented programming language-Java JDK 7 which is explained in chapter 6. The time efficiency achieved by reduction of database, as well as, it takes less space in memory. The algorithm also reduces the number of scans and time in finding the frequent patterns from the database.

IV. **Discovery of frequent patterns at multiple levels of abstraction by newly developed algorithm.**

After accomplishment of proposed algorithm, it was used to generate the frequent patterns at multiple levels of abstraction with the help of reduced minimum support parameter. The results are shown in Chapter 7.

V. **Comparative analysis of result retrieved by new algorithm.**

The proposed algorithm was compared with existing algorithms so that the significance of the new algorithm can be proved. The outcome of proposed algorithm is shown in chapter 7.

Finally, the formulation of the above mentioned research objectives will help in achieving the goal of the designing of a suitable research methodology. By which, the above mentioned objectives and research issues can be explored and investigated.

**3.2 Significance of Study**

The proposed research work has great importance for different sectors like research studies, nation, society, corporate sector & decision analyst. If any kind of new work has been done by any individual, this work contributes in his/her own individual growth, associated organization’s growth, society’s growth & national growth. The proposed work will be
helpful for all the parties involved in knowledge retrieval process for any kind of real life problem. Few of the contributions are explored here with for the current reference:

- This research work has tremendous significance for finding more specific information for users at multiple levels of abstraction.
- This research work is basically exploratory in nature; it will cover the descriptive model development.
- This research work will enhance the ease and comprehensibility of the users.
- The present research work helps us in understanding the significance of encoding of transactional database into encoded table.
- This research work explores the new ideas of generation of frequent patterns at multiple levels.

### 3.3 Research Methodology

The research methodology is stated in a way that initially, it elaborate the problematic issue and then an appropriate method is formulated to deal with the issue. The phases involved in the research methodology are:

1. Idea generating phase,
2. Problem definition phase,
3. Algorithm designing and implementation and
4. Performance evaluation of proposed algorithm; which are explained below:

First of all, it is essential to figure out the kind of problem. To accomplish this task, the comprehensive survey of the data mining techniques and association rules mining algorithms of single and multiple levels of abstraction was done. On the basis of this study, it was found that multiple levels association rules gives exact and precise information than other traditional single level association rules mining algorithms. Several methodologies to improve the discovery of frequent patterns at various levels of abstraction were studied and performance analysis was made. Moreover, it was also concluded that encoding of original database is required for the discovery of association rules at multiple levels. On the basis of these studies, it was concluded that there is a scope of improvement in the area of multiple level association rule mining. Proceeding further, an algorithm named as TransTrie, which is based on an advanced data structure called Trie, was developed which was a single level association rule mining algorithm. By using this algorithm a novel multiple level association rules mining method named as MLTransTrie was devised. It requires two types of files as input. First is the XML file, which is required to provide the concept hierarchy of data. The
concept hierarchy in form of XML file format describes the levels of data. Specimen is given below:

```xml
<?xml version="1.0"?>
<!DOCTYPE ConceptSet SYSTEM "concept.dtd">
<ConceptSet>
  <Concept><Node>food</Node></Concept>
  <Concept><Node>milk</Node><Parent>food</Parent></Concept>
  <Concept><Node>fruits</Node><Parent>food</Parent></Concept>
  <Concept><Node>amul</Node><Parent>milk</Parent></Concept>
  <Concept><Node>dairyland</Node><Parent>milk</Parent></Concept>
  <Concept><Node>weeds</Node><Parent>fruit</Parent></Concept>
  <Concept><Node>woseeds</Node><Parent>fruit</Parent></Concept>
  <Concept><Node>amul-two-percent</Node><Parent>amul</Parent></Concept>
  <Concept><Node>amul-skimmed</Node><Parent>amul</Parent></Concept>
  <Concept><Node>dairyland-two-percent</Node><Parent>dairyland</Parent></Concept>
  <Concept><Node>dairyland-skimmed</Node><Parent>dairyland</Parent></Concept>
  <Concept><Node>apple</Node><Parent>weeds</Parent></Concept>
  <Concept><Node>orange</Node><Parent>weeds</Parent></Concept>
  <Concept><Node>banana</Node><Parent>woseeds</Parent></Concept>
  <Concept><Node>pineapple</Node><Parent>woseeds</Parent></Concept>
</ConceptSet>
```

Second input file is the actual data file in ARFF format. Sample of the ARFF file format is given below:

```arff
@relation food
@attribute amul-two-percent {0,1}
@attribute amul-skimmed {0,1}
@attribute dairyland-two-percent {0,1}
@attribute dairyland-skimmed {0,1}
@attribute apple {0,1}
@attribute orange {0,1}
@attribute banana {0,1}
```
@attribute pineapple                         {0,1} 
@data

???1,1,??,1
1,??,1,1,??
??,1,1,??
1,?,1,1,??
????,1,1,1,?
??,1,1,?,1,?
??,1,??,1,1
????,1,1,1,1
1,1,1,?,??,?
%
%

Above mentioned both files are required as input for proposed algorithm. Subsequently, this algorithm uses an advance data structure trie. Trie is a tree data structure that allows strings with similar character prefixes to use the same prefix data and store only the tails as separate data. Due to this feature, it is appropriate for candidate set generation. This algorithm has been implemented using widespread high level and object oriented programming language-Java JDK version 7 using eclipse. This is followed by analysis of the proposed method that involved an investigation on the existing multiple level algorithms to evaluate their performance in terms of time and space requirement. In order to get a real life result for proposed algorithm, it has been tested on four real world datasets. These real world datasets Breast-cancer, Credit-g, Mushroom and Soybean are available on UCI Repository of Machine Learning databases [Weblink4]. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. This repository has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets. To study the performance of the algorithm, different support threshold were used.

3.4 Outcome

This chapter outlined the definition of the problem based on which the research objectives were articulated to handle the challenges. It has highlighted the objectives of research and
also outlined the significance of study. A research methodology to address the identified objectives is also given in this chapter.