Discovery of frequent patterns and association rule english language essay

Linguistics, English

CHAPTER 2

LITERATURE SURVEY

The purpose of this chapter is to provide the in depth review of the related topics concerning research area. It will provide review on Frequent Patterns, Association Rule Mining on single level, Multi Level Datasets, non-redundant association rules and interesting measures for Association Rules. In addition, the previous works on Association Rules are presented in this chapter. Review of journals, periodicals and other research publications related to the subject area was executed during the initial phase of the research and updated throughout the research. Several research papers, journals, conference paper, books needed to be reviewed to understand the discovery of Association rules.

Discovery of Frequent Patterns and Association Rule

Frequent Pattern Mining is an important area of Data Mining research. The Frequent Patterns are patterns that appear in a data set frequently. Finding such frequent patterns plays an essential role in mining associations, correlations and many other interesting relationships among data. It also helps in data classification, clustering and other data mining tasks. The process of discovering interesting and unexpected rules from large data sets is known as Association Rule Mining. This refers to a very general model that allows relationships to be found between items of a database. An Association Rule is an implication or if-then-rule, which is supported by data. The Association Rules problem was first formulated by Agrawal et al [2, 4 and 3] and was called the market basket problem. Agarwal et al [2] considered a

large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. The algorithm AIS [2] in this work, generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. The results of this algorithm are applied to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm. In this work, combinations of all items, which support minimum support threshold (minsupp), are called large itemsets and all other items are called small itemsets. The rules discovered in this work have one item in the consequent and a union of any number of items in the antecedent i. e.(X ☝ Y ☾ Z), it will not generate the rules of the form X ☾ Y ☝ Z. Another drawback of AIS [2] algorithm is, it generates too many small itemsets. This requires more number pruning steps and wastage of effort in the generation of Association Rule Mining.. Agrawal [3, 4] proposed two algorithms called Apriori and AprioriTid. The Apriori and AprioriTid algorithms generate the candidate itemsets to be counted in a pass by using only the itemsets found large in the previous pass, without considering the transactions in the database. The basic intuition is that any subset of a large itemset must be large. Therefore, the candidate itemsets having k items can be generated by joining large itemsets having k - 1 items, and deleting those that contain any subset that is not large. This is called Apriori property. This procedure results in generation of a much smaller number of candidate itemsets. There are two processes to find out all the large itemsets from the database in Apriori algorithms. First, the candidate itemsets are generated, and then the database is scanned to check the actual support count of the

corresponding itemsets. Where as in AprioriTid there is no database scan for support count, it considers the candidate set for support count. The candidate k-itemsets are generated after the (k-1)th passes over the database by joining the frequent k-1 itemsets. All the candidate k-itemsets are pruned by checking their sub (k-1) itemsets. If any one of its sub (k-1) candidate itemsets is not in the list of frequent (k-1)-itemsets, then this k-itemsets is pruned out, because, it is not a frequent according to the property of Apriori. In the process of finding frequent itemsets, Apriori avoids the counting of candidate itemsets that are infrequent. In the Appriori and AprioriTid approach, the database is scanned multiple number of times. If there is a single database scan, then the possible itemsets is to be tested for support is exponentially large. Suppose we are given a small set, say a few thousand itemsets, then the support for them can be tested in one scan of the database and the rules can be discovered. Partition [41] algorithm accomplishes in two scans of the database, in one scan it generates a set of all potentially large itemsets by scanning the database once. This set is a superset of all large itemsets, i. e., it, may contain false positives. But no false negatives are reported. During the second scan, counters for each of these itemsets are setup and their actual support is measured in one scan of the database. The algorithm executes in two phases. In the first phase, the Partition [41] algorithm logically divides the database into a number of non-overlapping partitions. The partitions are considered one at a time and all large itemsets for that partition are generated. At the end of first phase, these large itemsets are merged to generate a set of all potential large itemsets. In second phase, the actual support for these itemsets is generated

and the large itemsets are identified. The partition sizes are chosen such that each partition can be accommodated in the main memory, so that the partitions are read only once in each phase. The Partition [41] algorithm was motivated by the need to reduce disk I/O. In this aspect it has a clear advantage over the Apriori algorithm. The Partition algorithm reads the database at most twice irrespective of the minimum support and the number of partitions. Apriori reads the database multiple numbers of times. The exact number depends on the minimum support and the data characteristics and cannot be determined in advance. An important, contribution of Partition [41] approach is that, it drastically reduces the I/O overhead associated with previous algorithms (Apriori, AprioriTid). This feature may prove useful for many real-life database mining scenarios where the data is most often a centralized resource shared by many user groups, and may even have to support on-line transactions. Interestingly, this improvement in disk I/O is not achieved at the cost of CPU overhead. This work demonstrated with extensive experiments that the CPU overhead is actually less than the best existing algorithm for low minimum supports (i. e., cases which are computationally more expensive). In addition, the algorithm has excellent scale up property. All the above algorithms are used to discover the association rules. However, it is nontrivial to maintain such discovered rules in large databases because a database may allow frequent updates and such updates may invalidates some existing strong association rules but also turn some weak rules into strong rules. The Fast Update or Incremental Update [6] algorithm is an incremental updating technique for efficient maintenance of discovered association rules, when new transactions data are added to a

transactions database. The Incremental Update [6] algorithm uses the old large itemsets and pools of candidates are pruned. The Incremental Update [6] algorithm is faster than the Apriori [2, 4] algorithm and also the number of candidate sets are also less than the previous algorithms [2, 3, 4 and 41]. This algorithm determines the frequent itemsets and infrequent itemsets in the incremental portion and reduces the size of the candidate set. Borders [57] algorithm is an extension of Incremental Update [6] algorithm. This algorithm provides an efficient method for generating associations incrementally, from dynamically changing databases. Through the Incremental Update [6] algorithm, whenever a new frequent itemset is generated, it accesses the old data. Where as in Borders [57] algorithm, if the changes do not produce any new frequent itemsets, then there is no access to the old data. The Borders algorithm is based on the notion of border sets, introduced in Mannila and Toivonen [31]. A set X is a border set if all its proper subsets are frequent sets (i. e., sets with at least minimum support), but it itself is not a frequent set. Thus, the collection of border sets defines the borderline between the frequent sets and non-frequent sets. The Borders algorithm [57] works by constantly maintaining the count information for all frequent sets and all border sets in the current relation. Borders algorithm [57] outperforms Incremental Update [6], sometimes by more than an order of magnitude. Furthermore, in most cases Borders requires no access at all to the old data. Thus, in most cases, increments are almost instantaneous. All the previous algorithms [2, 3, 4, 6 and 5] solve the Association rule mining in bottom-up breadth-first search direction. In these algorithms, the computations starts from frequent 1-itemsets(i. e. minimal

length frequent itemset) and continue until all k-frequent itemset (maximal length frequent itemset) are found. During the execution, every frequent itemset is explicitly considered. Such algorithms reasonably perform well when all k-itemsets (maximal itemsets) are short. However, performance drastically decreases when some of the maximal frequent itemsets are relatively long. This is due to the fact that a maximal frequent itemset of size 100 implies the presence of 2100 – 2 non trivial frequent itemsets [21] is generated. So, discovering the maximum frequent itemset is a search problem in a hypothesis search space. The search for the maximum frequent itemset can proceed from 1-itemset to n-itemset (bottom-up) or from n-itemsets to 1-itemset (top-down). The Pincer-Search [29] algorithm combines both bottoms-up and top-down directions. This algorithm performs well even when the maximal frequent itemsets are long. In the Pincer-search [29] algorithm, the bottom-up search is similar to the Apriori[2, 4] algorithms. The top-down search is implemented with a maximal-frequent-candidate-set (MFCS). The Pincer-search [29] algorithm reduces both number of times the database scan and the number of candidate set. Most of the previous algorithms for frequent patterns and Association Rule Mining adopt Apriori like, candidate generation and test approach [22]. However, candidate generation and test may still be expensive, especially when long numerous patterns are encountered. Han et. al [22] proposed a new methodology called Frequent Pattern Growth (FP Growth) [22], which mines frequent patterns without candidate generation. This method adopts divide-and-conquer concepts, to project and partition database based on the currently discovered frequent patterns. This algorithm [22] generates the frequent

itemsets without candidate set generation. In this, in the First step, group the original data items that to be mined, and then focuses on counting the frequency of the relevant data instead of candidate set. In FP Growth [22] algorithm, instead of scanning the entire database, it partitions the patterns to be examined by database projection. This is a divide-and-conquer methodology, which reduces the search space and leads to high performance. In comparison with candidate set based algorithms, FP-growth [22] has the advantages, a FP tree [22] is highly compact. It avoids huge candidate sets generation. It uses partition based divide-and-conquer method which dramatically reduces the size of the subsequent sub-databases. FP-Tree [22] has limitations. Firstly, it is difficult to use this approach in an interactive mining process, where it is possible to change the support threshold [59]. The second limitation is that the FP-Tree [22] algorithm is not suitable for incremental rule mining [59]. There are other limitations to the original FP-Tree algorithm the majority of which are based on finding more sophisticated patterns [22]. Also, as the number of different/unique items increases, the size of the tree increases typically at an exponential rate due to the reduction in common prefixes sharing [5]. The CLOSET+ [39] is another algorithm for frequent itemset mining. It is based on FP-tree [22] algorithm. This algorithm will generate the frequent closed itemsets rather than frequent itemsets. An itemset X is called closed in a dataset S if there exists no proper super itemset Y such that Y has the same support count as X in S. An itemset X is a closed frequent itemset in set S if X is closed and frequent in S [29]. Mining frequent closed itemsets provides an interesting and alternative. Because, it inherits the same analytical power of

frequent itemsets, but generates a much smaller set of frequent itemsets. This leads to less and more interesting association rules than the previous algorithms. It has three technologies, First, it generates the frequent closed itemsets using FP-tree [22] structure. Second, it uses prefix path compression techniques to quickly identify frequent closed itemset. Third, it uses partition based projection for large databases. The major advantage of closed itemsets is that after the frequent itemsets are derived, the closed itemsets and their supports can be derived and determined without accessing the dataset. The number of closed itemsets is often significantly reduced when compared to the number of frequent itemsetsMost of the algorithms introduced above are based on the Apriori algorithm [2, 3 and 4] or FP-Tree [22]. These algorithms try to improve the efficiency by making some modifications, such as reducing the number of pruning passes over the database, reducing the size of the database to be scanned in every pass. However there are two bottlenecks of the Apriori algorithm. One is the complex candidate generation process that uses most of the time, space and memory. Another bottleneck is the multiple scan of the database. Limitation of FP Tree is, it is difficult to use this approach in an interactive mining process where it is possible to change the support threshold [59].

Discovery of Association Rules in Multi-Level & Cross-Level Datasets

Traditionally, association rule mining has been performed at a single concept or abstract level, which was usually either a low abstract/primitive level or at a high abstract/concept level [48, 51]. It is widely accepted that single level rule mining has two major problemsIt is difficult to find strong associations at

a low or primitive level due to the sparseness of data. Mining at high levels of data may give a common knowledge, which are already known and are of little interest [48, 51]. It is quite possible that for a given database, can be mined at single level, but contains data in a hierarchical format. It has been argued that few algorithms take advantage of this type of structure [37]. Han et al [48] defined " Association Rules generated from mining data at multiple levels of abstraction are called multiple level or Multi level Association Rules. Multilevel association rules can be mined efficiently using concept hierarchies under support-confidence framework". The following Figure 2. 1 shows an example of the Amazon hierarchy for their book collection, where each level has a different degree of abstraction. Figure 2. 1. Sample Hierarchical dataset. Such a Multi level can also have cross level of hierarchy. In general, Multi-level can be defined as the items come from one taxonomy level, but the set of rules span more than one taxonomy level. Where as, in Cross-level, the items come from more than one taxonomy or hierarchy level. An undiscovered knowledge can be discovered with multi/cross level, which could not be found by the single level approach and this new knowledge may be highly relevant or interesting to a given user [5, 21]. In fact, the multi-level rule mining is useful in discovering new knowledge, which is missed by conventional algorithms [36]. If a database/dataset has a hierarchical structure, then a multi-level or cross-level rule mining will discover more interesting knowledge. Multi-level rules span multiple levels of abstraction, but the items within one rule come from the same concept level. They can be at different levels and contain more general or more specific information than the single level rules. The

intermediate results from high levels of abstraction can be used to mine lower abstract levels and refine the process [18]. The Multi/Cross level Association Rules can be mined under support confidence framework considering uniform-support, reduced-support and group-based support. Cross Level Association Rule Mining: Cross-level rules are those where the items within a single rule come from different multiple levels of abstraction. Here the two items come from different levels in the hierarchy. Han & Fu [18] proposed that the relaxation on rule conditions among concepts at the same abstract level would allow exploration of what they termed ‘ level-crossing’ relationships, which would contain items at different abstract levels [18][16]. Some of the proposed approaches for multi-level rules will also allow a user to also discover cross-level rules. However, these works focused on the multi-level rules rather than the cross-level rules. When it comes to multi-level rule mining, there are several proposed approaches/methods, of which some are reviewed below. The original Apriori approach to mine single level association rules can be extended to mine the association rules from multiple levels. The simplest algorithm to perform generalized rule mining for multi-level was introduced by Sirkant & Agrawal [4]. It was called ‘ Basic’ and was simply the Apriori approach. But, in this, each item ancestor is also added into the transaction, and then performed the mining process across the expanded transactions. It was considered to be too slow even by its authors [4]. Because ‘ Basic’ was inefficient, two other approaches were proposed by Srikant & Agrawal, those are Cumulate and EstMerge [4]. Cumulate was based on the ‘ Basic’ approach and has two main optimizations over the plain ‘ Basic’ approach. First, all of the items

ancestors are not added into the transaction, instead they are filtered. Second, pruning is applied on any itemsets that contains an item and its ancestor. Experimentation performed showed that Cumulate was two to five times faster than ‘ Basic’ on synthetic datasets [4]. EstMerge[4] is based on estimating the support of the candidate. In this the item support is determined whenever it is necessary, instead of calculating it for all itemsets. It is necessary to count candidates items that don’t have the same support level. This usually involves another pass, before moving on to the next candidate set. EstMerge outperformed ‘ Basic’ by two to five times and in some tests it also outperformed Cumulate [4]. Han & Fu [18][19] proposed several Aprior -based variations, ML_T2L1, ML_T1LA, ML_TML1 and ML_T2LA [18][19], which are said to perform differently depending on the dataset and its distribution. These approaches can be adapted for cross level rules. In this, the generations of frequent itemsets are considered as the combinations of concept encodings that come from multiple levels. Its efficiency against other Apriori-based approaches is not clear as the authors did not compare to other multi-level approaches available. But these approaches have the advantage of sharing data structures and intermediate results between the different concept levels [23, 26]. Han also presented three different approaches for multi-level mining. The first is the Progressive Deepening approach [18][17] which starts at a high abstraction level and finds the strong associations. It then selectively moves down to lower levels of abstraction and finds associations. The second is a Progressive Generalization approach [14], which starts at a low level and finds associations. It then uses these results and generalizes them to higher

levels. The third is an Interactive-up-and-down approach [14], which involves the user interacting with the algorithm to direct it in stepping up or stepping down through the concept levels. Apriori uses a uniform support threshold that is the same support requirement for all itemsets regardless of length or whatever concept level they are located at. Wang, He & Han [50] argue that this means Apriori either misses interesting patterns or finds too many patterns and suffers from a bottleneck during its itemset generation. Their algorithm is known as Adaptive Apriori [50] and defines the best support threshold for each group of items and concept level individually through the use of a support based specification. This however means that it is necessary for a user to determine the most appropriate support threshold for use and the organization of the items [50]. The work has been undertaken in discovering both multi-level and cross level frequent itemsets from multi-level datasets. The approach taken is a top-down progressive method built upon existing algorithms used in mining single level and multi-level rules. The main difference between this approach and other similar approaches is the pruning that takes place. Thakur, Jain & Pardasani's [44] approach uses a reduced uniform minimum support, where each level has its own support threshold and the support threshold is reduced as the algorithm works down the levels. For each level, after the frequent itemsets are discovered, the dataset is pruned so that any items that are not frequent at the current level and are not descendents of a frequent item are removed [44]. In presenting their approach the author's do not consider, how the user should determine and select the support thresholds for each level. So, this aspect is still quite subjective and based largely on the user's opinion or belief. Multi-level

Apriori approaches still have the same two bottlenecks from the original Apriori and thus multi-level FP-Growth based approaches are developed. Since they are deemed to be more efficient and do not suffer the same performance bottlenecks. Adaptive FP-growth was proposed by R. Mao [32] to perform multi-level association rule mining based on a FP Growth approach [5, 10]. Using the FP-Growth approach, it first finds the frequent 1-itemsets and the concept level that it belongs to. The Adaptive-FP approach uses a different support threshold for each concept level, than the same support for each concept level.. Adaptive-FP treats all frequent itemsets regardless of concept level. When, it comes to the rule generation then the FP-Tree is revisited for association rules [32]. Another FP-Growth based approach to multi-level rule mining is FP'-Tree, proposed by Ong, Ng & Lim [36]. Their approach differs from other traditional rule mining algorithms. As they take into account the recurrence relationship/s within a transaction. The recurrence is simply the quantity of an item and can be used easily when there is a transactional dataset. Unlike Adaptive-FP, this approach builds a separate FP-Tree for each concept level that is being mined. Although this approach appears to be promising, but, there is no consideration given to the discovery of cross-level association rules that are contained within the dataset. The focus is solely on obtaining multi-level rules through the use of recurrence & quantity. Apriori and FP-Growth based approaches are the most widely used and developed when it comes to multi-level rule mining. However, some work has been done which is not completely based on one of these two approaches. One approach proposed is based on statistics and others are based on the idea of using fuzzy set theory. Páircéir, McClean &

Scotney [37] developed an approach that uses sufficient statistics to represent one concept level of the dataset. In this work, the data at the lowest concept level is not actually used but an aggregate representation is used. Their approach was focused on combining information from different distributed datasets. There is also privacy concern about individual data and probably will not work outside of this type of environment [37]. Another technique proposed for multi-level association rule mining is the use of fuzzy set theory. Hong, Lin & Chien [25] proposed the FDM (Fuzzy Data Mining) algorithm, which combines fuzzy set theory with linguistic terms. This works on finding patterns and rules from quantitative datasets. In this work, the assumption is made that the membership functions for the items is already known and does not need to be discovered during mining. This assumption does limit the use of FDM. This approach is limited to finding association rules for items that are not on the same path in the dataset hierarchy [25]. Hong, Lin & Wang [24, 25] developed a fuzzy Apriori-based approach for mining multi-level association rules using similar techniques to FDM. It too works on quantitative datasets and also relies on knowing membership function/s in advanced. These approaches may not generate the complete set of rules and thus information may be lost. But, it generates the most important rules because they include the most important fuzzy terms for the items [24, 25]. Kaya & Alhajj [24] proposed an approach that is based on the work of Han & Fu [16, 28] and Hong et al [24, 25], in that they used fuzzy set theory, weighted mining, linguistic terms. They used support and confidence to measure the strength of rule interestingness, and also item importance. By using these approaches it is claimed that the rules are more meaningful

and more understandable to users. The performance of this work produces consistent and meaningful results. It has been tested on a synthetic dataset only and there is no test on a ‘ real-world’ or actual dataset [24]. As already mentioned, often the number of rules discovered is quite large. However, not all of these rules are necessarily unique. Often there are redundant rules, especially when a support-confidence based approach has been used. Zaki [58] has stated that it was known that there was redundancy, but " the extent of redundancy is a lot larger than previously suspected" [58]. Redundant rules give very little if any new information or knowledge to the user and often make it more difficult to find new knowledge (and perhaps even help to overwhelm user when it comes to finding high quality interesting / important rules). This problem of redundant rules is claimed to become more crucial when the data is dense or correlated (such as in statistical datasets) [58]. Suppressing the redundant rules or removing them from the final result set, will make it easier for the user to handle, process and understand the remaining rules, which actually contain the new and unique information. Pasquier et al [38] has argued that there is a good reason to suppress or remove the redundant rules, they can be misleading. In Pasquier et al [38], they also argue that the support-confidence information is important when it comes to characterizing redundant rules and propose to generate a condensed rule set. This is done by maximizing the information in each rule and they aim to achieve this by presenting rules which have a minimal antecedent and a maximal consequent [38]. This approach has been shown to reduce the number of rules and thus remove redundancy. Both Pasquier et al and Zaki’s [38, 58] approaches are based on

frequent closed itemsets, Zaki used a FCA (formal concept analysis) framework, while Pasquier et al [38] approach was based on Galois closure [38, 58]. However, it has been argued that their approach does not remove all of the redundant rules. It is also only focused on single level datasets and rules. Hamrouni et al [13] has also done work in the field of removing redundant association rules and they propose the PRINCE algorithm [13]. The aim of their algorithm is to find distinct ‘ closure systems’ (sets of sets which are closed under the intersection operation, which is the set of closed itemsets and the set of minimal generators) and then derive generic bases of association rules. From these generic bases, all other association rules can be derived. By only extracting the generic bases, the number of rules is reduced and thus redundant rules are removed. The PRINCE [13] algorithm conducts a level-wise browsing of the search space and builds a partial order structure, known as the minimal generator lattice. This is maintained between frequent minimal generators. When this algorithm performs sweeping of the lattice, then the itemset closures are derived to build generic bases of association rules. Testing has shown its performance (in terms of time) to be at least comparable to than other algorithms that perform level wise browsing [25]. T. Calders et al [45] proposed Non-Derivable approach to deal with redundant rules is based on extracting non-derivable association rules. This approach is based on the idea that if the upper and lower bounds of a rule’s confidence are equal, then the rule is uniquely determined by sub rules and can be derived from them. Such a rule is considered to be derivable rule and therefore it is considered as redundant. In this proposal, redundancy is tested on deriving the absolute

bounds of the rule’s confidence instead of estimating them as is done in other approaches [45].

Interesting Measures for Single/Multi Level Association Rules

The aim of association rule mining is to uncover associations between data items. At the same time some rules may not be important, relevant, correct or even interesting to the user. It is even possible for rules to be misleading [21]. And they also may be of low or poor quality. Also, association rule mining has been said to produce too many rules and some of them may produce redundant rules. This has been caused by the focus being on improving the efficiency of generating the rules [55, 56 and 58]. There is no accepted definition on what interestingness is or what ‘ goodness’ is. In some cases it appears that interestingness and ‘ goodness’ are considered to be the same thing. In that, if a rule is interesting then it is a good rule and in other cases they are separate but related issues. To demonstrate the lack of agreement over interestingness in just three different survey papers several measures are summarized. In one paper there are 38 different measures [9], in the second there are 39 measures [35] and in the third there are 17 measures [33]. In this section the review will look at several of the different interestingness measures that can be applied to association rules. Determining and measuring the interestingness, quality of association rules is an important area. Much work has been done, but yet there is no formal definition or even a widely accepted agreement in the context of association rule mining [9]. It currently seems that interestingness is a broad concept that is based on and emphasizes the [9] conciseness, Generality/Coverage,

Reliability, Peculiarity, Diversity, Novelty, Surprisingness, Utility and Action ability [26]. There are many different measures available for measuring the interestingness or ‘ goodness’ of association rules. All of these measures can broadly be defined as belonging to three different categories; objective, subjective and semantic. Objective based measures determine the interestingness of association rules based purely on the raw data. Subjective measures use both the raw data and user data to determine the interestingness. Semantic based measures determine the interestingness of a rule based on the semantic meaning(s) and explanation of the patterns. Subjective based measures are more difficult to define and implement, as information about the user must be taken into consideration. Currently objective measures are the most popular and some are reviewed here. The most widely employed method in association rule mining to determine interestingness is a generality and reliability based measure known as the support-confidence approach, which can lead to interesting rules being found by setting low support thresholds [21]. In a support-confidence approach, the support measures the range of the rule and the confidence measures the precision/accuracy of the rule [38]. Support was chosen as it represents statistical significance [1, 33]. The rule is interested when the rules support is above a user specified threshold. The calculation of support assumes statistical independence and the support-confidence approach is targeted at finding qualitative rules [21, 29]. The support-confidence approach has several deficiencies like, rules with high confidence and support are often trivial. These rules do not convey any information. Many variants of a rule might be produced thus giving the same or similar information repeatedly

(this is known as redundancy). Some of these deficiencies can be compensated by using post-processing techniques. However this adds the need for extra resources to produce the knowledge, which is often not desirable. With the support-confidence approach an estimate of the probability of the rule is identified. These will not generate the real strength of correlation and implication between the items. It is possible for this approach to discover and give to the user misleading rules. This is especially the case when the confidence of the rule is less than the support of the consequent of the rule [21]. The support-confidence approach is popular because it is relatively simple, efficient and produces a decent result set. Another reliability based approach is correlation. This approach measures the correlation between items in the rule. One of the advantages of this approach is both positive and negative relations can be discovered. If it is positive correlations, then the presence of one item implies the presence of another item. However, in negative rule the presence of one item discourages the presence of another item [21]. Several correlation based approaches have been proposed with different scoring functions [21, 26]. The Correlation measure divides the rules into significant and non-significant, but cannot rank them. Another measure is needed to rank the significant rules as the correlation measure is not bounded and comparisons would be meaningless [26]. Gray & Orlowska [11] proposed Weighting Dependency measures for interestingness of association rules. It is based on a generality and reliability based approach. This is used to measure the interestingness in association rules between a rule’s antecedent and consequent. This proposed approach takes the view of integrating the rule

discovery and the clustering of items to solve the granularity problem. This is normally done with the help of taxonomy. Items at low level are missed because of their low support. The proposed measure contains both a discrimination and support component. Using this approach, interestingness is given by the following formula: This approach does not require a taxonomy as it instead uses clustering to build a structure. It was limited to single pairs of attributes and had high complexity. But, it will work on any level of data abstraction and thus can be utilized in multi-level datasets [11]. However, experimentation appears to have been limited to synthetic data and thus this approach may not scale up [33]. A measure proposed by Dong & Li [7] was proposed a distance based peculiarity measure to determine and identify interesting rules. This approach is based on breaking the rules into neighborhoods rules. Then this work identifies the rules whose confidence was significantly different than the rest of the neighborhood [7]. In this approach the surrounding rules of a given rule will influence the level of interestingness of that rule [33][7]. Most techniques focus on finding and evaluating strong rules, which is good and useful in prediction tasks. However, there are weak rules, which show unexpectedness and contradiction on accepted knowledge or belief. These weak rules represent a small number of items, and can be more interesting than strong rules [56]. Liu et al [33, 35] proposed a reliability based approach to discover the weak rules, which they call reliable exceptions. A reliable exception is a weak rule having relatively small support and relatively high confidence and they can be induced. In Liu et al [33, 35] approach, the following process is used. First, generate the strong rules or a predetermined number of the strongest rules

using rule induction. Reliable exceptions will be evaluated with respect to these rules. Second, using contingency table analysis, identify significant deviations between the actual and expected frequency of occurrence of attribute-value and class pairs. Third, a deviation threshold is specified, for positive deviations, any deviation greater than the threshold is to be considered outstanding. Fourth, get all 1 instances containing the attribute-value and class pairs of the outstanding negative deviations. Fifth, calculate the difference between the confidence of the rule for the selected instances and the whole dataset. Now the confidence for the selected instances is always 1. A large difference (near 1) implies that the confidence on the whole dataset is low and thus a reliable exception has been located [38][56]. This evaluation method is used to find the weak exceptions and therefore is not used for predictive tasks. Instead it attempts to find information that goes against normal belief. It can however be used for both subjective and objective rule evaluation, finding and can take in prior knowledge. This approach does not measure the interestingness of rules and thus is not an interest measure. It measures the weak rules to find reliable exceptions. However, this approach could be used for improving the quality of the rules. Good quality rules should include all the available information, and also gives the user exceptions. Conversely, if a rule satisfies the fewer exceptions or contradictions then the rule can be considered as strong rule. Another peculiarity based interestingness measure for association rules was proposed by Zhong, et al [60] Their approach was designed to determine unexpected or peculiar association rules [23, 33]. In this proposed approach peculiarity was taken to mean that the rule represents a particular case that

is described by a small number of rules which are very different to the rest of the rule set [33].

Organization of the Thesis

In chapter 1, contains the introduction of the thesis. In Chapter 2, a review of literature and works from relevant topics and areas is undertaken. The review focuses in depth on association rule mining, measures used within the rule mining. The proposed research and work presented here utilizes the literature reviewed in this chapter to avoid the duplicating existing works and to present new work which complements what has already been done. In Chapter 3, the Non-Redundant rules on Single Level (also called as Non-Taxonomy) Datasets were discussed using concept lattice of Formal Concept Analysis are explained. In this, the frequent closed itemsets are generated which generates the non-redundant association rules. This approach generates lossy association rules. For these reasons, this proposal adopted the MinMax Approximation to generate the Non-Redundant and loss less rules. In this work, discover the frequent itemsets which satisfies the minimum support and minimum confidence threshold. To differentiate the boundary between the redundant and non-redundant rules the certainty factor (CF) [42] is used for safe removal of redundancy. The redundant rules are eliminated without loss of information on single level association rules. But the redundancy can exist if the datasets spread into many levels and such datasets are called Multi Level Datasets. In Chapter 4, The proposed method for discovering the non-redundant association rules on multi-level datasets (Taxonomy datasets) using MinMax approximate Rules and MinMax

Exact Rules were discussed with the emphasis on multi level datasets. In this proposed work, first we encoded the taxonomy information of the concerned multi level datasets. For example, the 1-3-1 and 2-1-1 represents the encoding of one item in the itemset. The first digit 1 and 2 represents the it’s level 1 data and so on. Then the closed itemset and its generator are discovered. Later we retrieved the non-redundant rules using MinMax Approximate, MinMax Exact rules, to successfully eliminate the hierarchical redundant rules without loss of information. To know the discovered rules are quality and useful rules, we proposed the interesting measures. In Chapter 5, the proposed interestingness measures for association rules derived from multilevel datasets are applied to find out quality and usefulness of derived rules. There are many number of measures are available to find the quality of association rules on single level datasets. Perhaps, there is a lack of interesting measures for Multi Level association Rules. For this reason we proposed, diversity and peculiarity measures for Multi Level Association Rules to focus on efficiency, effectiveness and usefulness of derived rules. The diversity measure compares the items within a rule and peculiarity compares the items in two rules, to see how they are different from each other. With these measures we can identify interesting rules which can’t be identified using basic measurements support and confidence. In Chapter 6, contains the experimental results were discussed based on the work proposed in Chapters 3, 4 and 5. The results are obtained when applying the proposed non-redundant rule mining methods to a multi-level dataset along with the proposed interestingness measures for multi-level datasets were analyzed and the conclusions are discussed. Finally, in

Chapter 7, conclusions from the research are drawn and the scope of the work was discussed. The future research and development activities are also discussed in this chapter.

Summary

To reduce the number of mined results, many interestingness measures have been proposed for various kinds of patterns. In this article, we reviewed interestingness measures used in data mining to generate the association rules. We reviewed measures for association rules, classification rules, and summaries. We also proposed the related work on objective, subjective, and semantics-based measures. Objective interestingness measures are based on probability theory, statistics, and information theory. However, objective measures take into account neither the context of the domain of application nor the goals and background knowledge of the user. Subjective and semantics-based measures incorporate the user’s background knowledge and goals, respectively, and are suitable both for more experienced users and interactive data mining. It is widely accepted that no single measure is superior to all others or suitable for all applications.