Abstract of the PhD Work

Name of the Research Scholar	**:**	Samar Wazir
Supervisor	**:**	Prof. Tanvir Ahmad
Co-Supervisior	:	Prof. M. M. Sufyan Beg
Department	**:**	Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi.
Title	**:**	Performance Analysis, Improvements and Optimisation of Techniques used in Association Rules Mining.

Association Rules Mining (ARM) is the field of study in which the relationship among real-world entities, objects, business firms, logs and real-time data is analysed and these results are used to generate improved, productive, predictive performance. ARM is one of the most research centric areas of computing; therefore in this research, various popular methods of ARM are analysed, improved and optimised. When the discussion on ARM starts then the term Frequent Itemset Mining (FIM) which is a sub-discipline of ARM, always come along with Apriori algorithm. FIM is a technique to extract the most frequently occurred pattern from a certain or uncertain (probabilistic or fuzzy) database by following a sequential or parallel pattern of execution. FIM algorithms have been designed to work on either certain database, which usually contains past information and the presence of an item in the database transaction is definite here or uncertain database, which includes information to predict future events and each item in the transaction is associated with its existential probability or membership function.

During the analysis of certain and uncertain algorithm for generating a frequent pattern, a question arises that why there is no strategy to compute frequent items for the combination of both databases. This requirement has been covered in this research work, an algorithm is developed as MasterApriori to calculate frequent items for the combination of both certain and probabilistic uncertain databases by following the approach of Apriori on certain and UApriori on uncertain databases. Here, it is assumed that all items in certain databases are also probabilistic but with 100% probability and then combined them with an uncertain database. The results generated by MasterApriori were novel and interesting, but the use of UApriori reflects in loss of association rules. In UApriori, expected support for an itemset is calculated by multiplying the existential probabilities of each item in the itemset for a particular transaction and then totalling the results for all transactions. Due to the multiplication of probabilities, the expected support for long size itemset becomes zero. So UApriori is unable to calculate expected support for long size itemset which yields in loss of association rules. To solve this problem, the work is extended, and UApriori is replaced by Poisson Distribution based UApriori and Normal Distribution based UApriori in MasterApriori. In this second paper, the loss of association rules is decidedly less compare to previous work. This work is further improved and the probabilistic database is replaced by the fuzzy database by using fuzzy uncertainty. In this case, first probabilistic database is converted into the fuzzy database by the following probability to possibility transformation principle. Thereafter, frequent items are calculated for the combination of certain and fuzzy uncertain databases by following the approach of Apriori on certain and Fuzzy Transaction Data Mining algorithm on the fuzzy uncertain database. The performance of all algorithms has been verified on synthetic and standard databases and a software tool named FIMAK (Frequent itemset Mining Algorithms Kit)

has been developed which is available online and used to generate frequent items by old available algorithms and newly proposed algorithms.

On the fuzzy database, fuzzy support is calculated by using fuzzy max or min operator. Min-Max operators are unable to aggregate multicriteria to generate useful results. Therefore, a new operator has been introduced in ARM and to the best of our knowledge first time, OWA is used in FIM to calculate support on the fuzzy uncertain database. This work has introduced a new method of FIM which follows Apriori pattern of execution, uses OWA operator to calculate support of an itemset and executes on the fuzzy uncertain database. The quality of this algorithm can be estimated as, if in the database some values are incredibly high or low then OWA eliminated those value and results are computed by aggregating the values of a specific range, e.g. in our case "most" case of OWA is considered, and values of range 0.3-0.7 are aggregated. In other instances, items present in the databases have some weights or importance except their existential probability or membership function. Mining of such kind of frequent items is known as weighted frequent itemset mining. There are various algorithms available for mining weighted frequent items on certain or probabilistic/fuzzy uncertain databases, but in this research, a new approach introduced by using OWA for mining weighted frequent itemsets.

Calculation of support of an item or itemset in the database is one of the parameters to show how often an item occurred in a transactional database but if the interest is not only to calculate frequently occurred items but also to calculate those items whose presence affect each other, e.g. if item sold frequently then the sale of other increased or decreased in case of market basket data. So, here the attention is to find out interestingness among items which is called correlation analysis in ARM. Correlation analysis can be obtained by a correlation coefficient. In case of a certain database, the

correlation coefficient between two items can be calculated by lift or chi-square test between items. There are some algorithms available to perform correlation analysis on certain or probabilistic/fuzzy uncertain database between two items only, but no method is existing to perform multiple correlation analysis. In this research work, a new approach is introduced to perform multiple correlation analysis for more than two items by calculating Pearson`s Correlation Coefficient using fuzzy means and OWA. All algorithms are coded in C, implemented on standard databases and performance is analysed by execution time and the number of frequent items generated.