

Online Generation of Association Rules

Abstract – This project focuses on the problem of online mining of association rules in a large database of sales transactions. The main aim of the project is the Online Generation of Association Rules. The idea is to implement an algorithm as efficiently as possible with minimum number of I/O operations and to reduce the number of passes over the data.

To understand the idea of association rules, consider a set of literals called items represented by $I = \{ i_1, i_2, \dots, i_m \}$. The database consists of a set of sales transactions \dot{Y} . Each transaction $T \in \dot{Y}$ is a set of items such that $T \subseteq I$. A transaction T is said to contain a set of items X if $X \subseteq T$. An association rule is a condition of the form $X \Rightarrow Y$ where X and Y are two sets of items. An intuitive implication of the association rule is that presence of a set of items X in a transaction indicates possibility of presence of itemset Y .

The strength of the rule $X \Rightarrow Y$ is specified by parameters support and confidence. The support of the rule is the fraction of transactions that contain both X and Y . The confidence of the rule $X \Rightarrow Y$ is the fraction of transactions containing X that also contain Y .

The project undertaken will find association rules as stated above for a user-specified minimum support and minimum confidence, i.e. finding all rules whose support and confidence values are greater than levels specified.

To mine the association rules efficiently, some preprocessing needs to be performed, in order to make it suitable for repeated online queries. The preprocessed data is stored in such a way that online processing may be done by applying a graph theoretic search algorithm whose complexity is proportional to the size of the output. This means the time taken by the algorithm does not depend on the size of the database but on the size of the output. The algorithm is capable of finding rules with specific items in antecedent and consequent. It also reduces redundancy and is hence more efficient in reducing the irrelevant noise in the process.

The algorithm uses the itemset approach, i.e. all combinations of items are generated that have fractional transaction support above a certain threshold called minsupport. Using these itemsets, rules are generated, and only those rules above a user-specified minconfidence need to be retained. The generated itemsets are stored in main memory in the form of an adjacency lattice. This reduces the number of I/O operations required in the analysis.

Index Terms – OLAP (Online Analytical Processing), Associations Rules, Data Mining, Knowledge Discovery, Itemsets, Adjacency Lattice.

References

- [1] C.C. Aggarwal and Philip S. Yu, "A New Approach to Online Generation of Association Rules" in IEEE Transactions on Knowledge and Data Engineering, Vol. 13, No. 4, pp. 527-540, August 2001.
- [2] C.C. Aggarwal, and Philip S. Yu, "Online Generation of Association Rules", ICDE Conf., 1998