

Question 1 (Time spent: 3h 30min):

Fast Algorithms for Mining Associate Rules

In this paper, two new algorithms were introduced: Apriori and AprioriTid. Both of the algorithms were introduced to mine associate rules and to generate all of the associate rules that have support and confidence of user predefined minimum support or minimum confidence. Earlier introduced algorithms such as, AIS and SETM are fundamentally different from the new ones and it is needed to be mentioned that new algorithm, Apriori and AprioriTid, “have always outperformer earlier algorithms.”(2pg) The key features of newly introduces algorithm are that they “generate item 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.”(3pg) Such a procedures provides the outcome if generating the much smaller number of candidate itemsets. This particular article also describes how Apriori and AprioriTid algorithms can be combined into hybrid algorithm called AprioriHybrid. This particular algorithm has excellent scale-up capabilities, allowing the feasibility of mining association rules over very large databases. Finally, authors of this article do not provide in-depth discussion but simple agree that the usefulness of associative rules can vary and are often application-dependent. Human guidance is much needed when determining the validity of association rules.

A Statistical Theory for Quantity Association Rules:

This paper provides definition of quantitative association rules based on statistical inference theory. As it is stated, association rule contains left-hand side and right-hand side. Left-hand side is a description on subset of the population, and the right-hand side provides description of the *interesting behavior* of the selected subset. To examine their introduced algorithm, authors Yonatan Aumann and Yehuda performed testing on real-word database. Database which was selected to be mined was built to study English writing habits of non-native English-speaking people. “The database contains 643 transactions and 42 attributes: 15 categorical and 27 quantitative. With approximately 27,000 entries and 42 different factors (making for many hundreds of possible patterns), the database is large and computerized tools are necessary.” (267pg) This test discovered more than 30,000 rules, which only 534 were accepted due to a confidence level of 95%. It is interesting to mention rules that included more than one categorical attribute in the profile were determined as non-intergering. It is right to assume that complex rules are difficult to understand and may be presented as non-interesting, rather than simple rules being determined as interesting ones. Authors of this research paper also specify that the analysts, such as computer scientists are not able to determine if found quantitative association rules interest user or not. The only way to successfully measure success of the founding is to ask the end-user.

Question 2 (Time spent: 2h 30min):

1.

Constraint: Antimonotonic – anything that costs more than \$150 can be pruned from search space. Antimonotonic constraint can be applied at each iteration of Apriori algorithm. This would improve the performance of the mining process while making sure that each data mining task is completed.

2.

Constraint: Monotonic – can be successfully mined using FP-growth algorithm. Since the sum of items is at least \$200 and there is one free item in the set rules of constraint are satisfied, so do all of its supersets which indicates that this applies monotonic constraint rules.

3.

Constraint: Convertible (Strongly Convertible) – Items that have average value in the pattern less than \$100 can be prune and items that have value greater than \$500 in the pattern, can be prune as well. Convertible constraint can be successfully mined using FP-growth algorithm.