New Algorithms for Discovering Association Rules

A Thesis
Submitted to the Department of Computer Sciences, University of Technology in a Partial Fulfillment for the Requirements for the Degree of Doctor of Philosophy In Computer Science

By
Emad Kadum Jabbar Alfatly

Supervised By
Prof. Dr. Hilal Hadi Saleh    Dr. Ahmed Tariq Sadiq

2005
Abstract
The goal of data mining is the task of discovering interesting knowledge from large amounts of data where the data is stored in database, data warehouse, or other information repositories.

A knowledge discovery process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evolution, and knowledge presentation. Data mining is a good way for extracting or mining knowledge from a vast amount of data for classification, predication, estimation, clustering or association rules or any activities, which need decision. Association rules identify relationships between attributes in a database. All algorithms of association rule mining consist of ‘first finding frequency of itemsets, which satisfy a minimum support threshold, and then compute confidence percentage for each k-itemsets to construct strong association rules’. In this research three algorithms are proposed, two for discovering association rules (ARAND and ARG) and the third algorithm to reduce data size.

The first proposed algorithm ARAND (Association Rule with logical AND operation) aims to produce association rules depending on logical AND operation by convert the database transaction into binary representation and neglecting any sum (column) less than threshold to find identical column in (k-1)-itemset table with column in k-itemset table which represents the association rules.

The second proposed algorithm ARG (Association Rules Due to Groups) aims to produce algorithm to mine association rules depending on group logical AND operator and no dependency between these groups to get fast algorithm by reducing execute time during execute all groups at the same time. ARG creates group table (k-itemset) selecting each transaction record which has k-itemset=1 from data transaction table and neglecting any sum (column) less than threshold, then uses logical AND operation for each column and checks the result if it equals 1 then that implies existence of an association rule between this column and k_itemset.

The Third proposed algorithm TFA (Threshold Frequent A priori) aims to produce algorithm to reduce size of data due to record duplicate threshold, since any duplicate transaction in data transaction table means no new information and that’s not useful in data mining, so it can be regarded as a step in preprocessing data like clean data, select data. This data reduction aids to reduce search space which implies reducing execution time.

Due to many experiments in ARAND and ARG algorithms without using TFA algorithm, these algorithms are very fast (with little number of items and little number of transaction records) compared with Apriori algorithm because a number of itemset that generate increase exponential with a number of items and that needs too much memory size so we repeated the experiments using TFA algorithm and detected ARAND algorithm is faster than ARG and Apriori Algorithms because no candidates are generate and no computation is made on frequency and confidence.

Moreover we have detected Apriori algorithm with TFA makes it faster. Lastly we tried to present new manner to mine association rules through ARAND and ARG algorithms and also presented data reduction algorithm.