ID2222 – Data Mining

Homework 2: Discovery of Frequent Itemsets and Association Rules

Group 10

Xin Ren

Nikhil Dattatreya Nadig

2018-11-08

Introduction

In this homework, we implement the Apriori algorithm for finding frequent itemsets with support at least ‘s’ in a dataset of sales transactions. We also implement the bonus part which requires to implement an algorithm for generating association rules between frequent itemsets discovered by using the Apriori algorithm in a dataset of sales transactions which requires the support of at least ‘s’ and confidence at least ‘c’, where ‘s’ and ‘c’ are given as input parameters.

Apriori.scala

This file contains the following functions:

- findFrequentItemSets()
- generateAssociationRules()

findFrequentSets first find all frequent 1-itemsets. This is done by creating a map of all the items with a count of 1 and reducing by the key value. Then it finds all frequent k-itemsets recursively: filtering out baskets which do not contain frequent items from previous round; generating length k candidates for each basket using frequent items of basket; calculating support of k-itemsets with map and reduce functions and filtering out those with support less than s.

generateAssociationRules we use the frequent item set to generate association rules of frequent itemsets discovered by using findFrequentItemSets() on a dataset of sales transactions. For each frequent itemset of size greater than 1, we find all of its association rules by finding all of its non-empty subsets. For each association rule, we calculate the confidence, by taking the ratio of the support of itemset and the support of the corresponding subset. We then, filtering out those rules which has confidence value less than the given confidence value.

Main.scala

This defines all the parameters for **findFrequentSets** and **generateAssociationRules.** The support threshold is 1/100th of the total size of the dataset. We also set the confidence for the association rules at 0.5. The data is loaded and processed with Spark. All the frequent itemsets and association rules are printed out. And the execution time for both functions is also printed.

Dataset

T10I4D100K.dat which is a transactions dataset is used for this homework.

How to Run

sbt run

Results

The STDOUT by running the code can be found in results.txt in the zip.