ID2222 – Data Mining

Assignment 2 - Discovery of Frequent Itemsets and Association Rules

Course Coordinator:
Vladimar Vlassov & Sarunas Girdzijauskas

Teaching assistants:
Edward Tjörnhammer
Kambiz Ghoorchian
Mohamed Gabr

Group 2

Óttar Guðmundsson

Örn Arnar Karlsson

2018-11-18

Description

Finding itemsets in a large amount of data can be very valuable. Companies try to find which products customers tend to buy together in order to improve sells. The purpose of this assignment was to implement the Apriori algorithm for finding frequent items in baskets given a support S. A support of an itemset is the number of times the itemset appears in different baskets. For this project we choice Apache Spark.

How to run

We wrote our solution in Scala using a Jupyter notebook with Spark Kernel. A guide to installing Jupyter Notebook, install Spark Scala and connect the kernel to Jupyter can be done using Toree (guide here)

Open the notebook and click “Run all cells” on the toolbar. This will load in the the data as an RDD in Spark, compute the singletons of all elements in the list of transactions and also generate a set for each transaction. Both are then loaded into the Apriori algorithm with the parameters K and S, that stands for how many combinations of items to look for and filter only those that occur more often than S many times.

Note: In some cases, we encountered dead kernels in the Notebook. Attached is the file assign2.scala that has all of the code in a pipeline, computing all of our calculations.

Solution

The solution starts with loading the proposed dataset for the assignment in spark text file. From this variable two String RDD’s are constructed, *singles* (counts the number of times an element appears in all of the transactions using map/reduce) and *transSet* (List of Sets that represent each transaction). The value K is set to 4 since we want to find a maximum of 4 items in a combo, and S set to 1000 since we want to look for items that appear in at least 1% of all transactions.

The apriori starts by filtering the singletons, removing all items that appear fewer than S times and then generating all possible candidate pairs. Each pair is then mapped to each transaction, where we count how many times a candidate pair is a subset of a transaction. This returns a tuple of the candidate pair and times it appears in all transactions, which is then filtered by S again. This is done K many times and simply returns an empty list when no combo of length K is found.

For the bonus part, we implemented the association rule and we also calculated the interest because we were interested in doing that.

For all pairs found in K that appear S or more times, a rule combination is checked. Let us say that for K=3 we have the candidate pair A,B and C that appear in 1200 transactions. Three rules will be generated from this candidate pair, namely, AB -> C, AC -> B and BC -> A. For each rule, the confidence of is calculated by looking at the appearance of set K-1 to see how many times AB happened individually from AB and C. If this confidence is above a selected threshold, this will be added to a list of associated rules for K.

Furthermore

We also calculated the interest of the association rule. We did this by subtracting the appearance of the associated item divided by the number of transactions from the confidence level. If this threshold scored more then 0.7, we considered that this rule was interesting and worth noting down!

Results

We read from the given data list with 100.000 transactions. We set the support as s = 1000 and confidence as c = 0.5. We found 375 singletons with more than the given support. In the following tables, we summarize what sets we found for a given K with a larger support than s. We also calculate the confidence and interest for the rules we found.

K = 2	
Pair	Support
(789, 829)	1194
(39, 704)	1107
(704, 825)	1102
(390, 227)	1049
(682, 368)	1193
(217, 346)	1336
(368, 829)	1194
(39, 825)	1187

For these pairs, we found three rules and we calculated the confidence and interest for them.

Rule	Confidence	Interest
704 => 39	0.617	0.574
704 => 825	0.614	0.583
227 => 390	0.577	0.550

K = 3	
Pair	Support
(39, 704, 825)	1035

For this set, we found three rules and we calculated the confidence and interest for them.

Rule	Confidence	Interest
704,825 => 39	0.939	0.897
39,825 => 704	0.872	0.854
39,704 => 825	0.935	0.904

We guess that pair 39, 704 and 825 stands for milk, beers and diapers.