Homework 2

Discovery of Frequent Itemsets and Association Rules

Denys Tykhoglo

Solution

For this homework, I used the sales transactions dataset that was provided at the homework page in Canvas. The dataset contains 100000 baskets of different sizes and the total number of item types is 1000. All item types are represented as integers in the range from 0 to 999 so it was not necessary to convert item names to their ordinal numbers.

The solution implements the A-Priori algorithm for finding similar item sets for the specified support threshold. After all similar item sets are found, the solution displays a list of all association rules that have the specified support and confidence.

The A-Priori algorithm is implemented inside two methods:

- *firstPass()* implements the first pass of the algorithm that reads the dataset, counts each item occurrence and finds all frequent items (that is, all singular item sets);
- *nextPass(setSize)* implements each pass from the second one and further. Parameter *setSize* corresponds to the pass sequential number, as well as to the size of frequent sets that are sought during the pass. As the A-Priori algorithm suggests, each item set of size *setSize* is counted as a candidate set if and only if each of its items appears in at least *setSize – 1* frequent sets from its basket , while all of these sets have to be of size *setSize – 1.*

Method *getAssociationRules()* returns all association rules for the dataset.

Another method that is used by both A-Priori algorithm implementation and the association rules finder is *getSubsetsOfSize(items, size).* For a given set *items,* it returns all subsets of size equal to *size*.

The solution is implemented using Java SE 8.

How to run

In order to run the solution, it is only necessary to compile the attached *TextSimilarity.java* file. The optional console parameters are the following:

1. The path to the dataset file (string). The default path is *src\main\resources\T10I4D100K.dat.*
2. The total number of items that can be encountered in transactions (integer number). The default value is *1000*.
3. Total number of baskets in the dataset (integer number). The default value is *100000*.
4. Support threshold, which is a fraction of baskets that a set needs to be part of to be considered frequent (decimal number). The default value is *0.01*.
5. Confidence threshold for association rules (decimal number). The default value is *0,5*.

Results

Figure 1 demonstrates the screenshot of a part of the output. Several sets of size 1 and 2 are left out of the screenshot for better readability.

The output consists of two parts:

- The first part displays all frequent item sets in curly brackets and the sets' support values in parentheses. The frequent item sets are grouped by their sizes.
- The second part displays all association rules and their confidence values in parentheses.

```
Frequent item sets of all sizes and their support values (support threshold = 0.01):
1: {1} (1535), {4} (1394), {5} (1094), {6} (2149), {8} (3090), {10} (1351), {12} (3415), {17} (1683),
2: {217,346} (1336), {227,390} (1049), {704,39} (1107), {39,825} (1187), {368,682} (1193), {722,390} (
3: {704,39,825} (1035),

Association rules and their confidence values (confidence threshold = 0.5):
704 -> 39  (0.61705685)
39 825 -> 704  (0.8719461)
227 -> 390  (0.5770077)
704 39 -> 825  (0.93495935)
704 825 -> 39  (0.9392015)

Process finished with exit code 0
```

Figure 1: Screenshot of a part of the output