Discovery of Frequent Itemsets and Association Rules

Konstantin Sozinov, sozinov@kth.se
Kim Hammar, kimham@kth.se

November 16, 2017

1 Solution

We implemented the solution in pure scala without any big data processing framework. We used the T10I4D100K.dat dataset uploaded on canvas. The source code is split into four classes, `Apriori.scala` that implementes the Apriori algorithm; `AssocRules.scala` that mines association rules from counted itemsets; `DataUtils.scala` that reads the data into a item-basket data model and `Main.scala` that orchestrates the pipeline and prints the results.

2 How to run

Clone this repository and navigate to *frequent_ itemsets* project. Then use:

```
sbt compile //compile
sbt test //test
sbt run //run
sbt assembly //generate fat jar
```

3 Evaluation and results

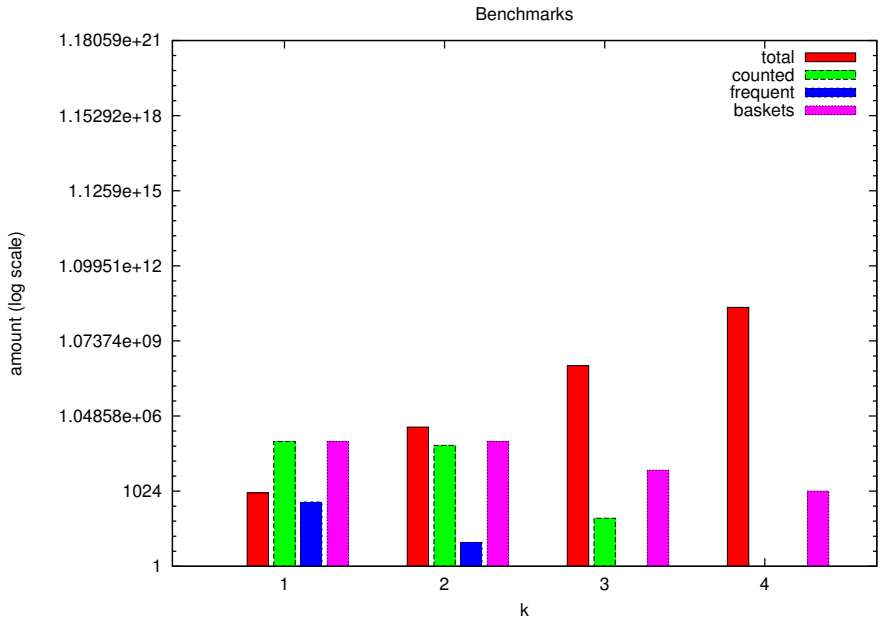

Figure 1: Analysis of number of counts made at each stage (log scale). Frequent itemsets of length 3 was 1, and counted itemsets of length 4 was thus 0. In-between each iteration we also filter the baskets (we hoped to reduce the complexity of the double-loop to count itemsets which has complexity $\mathcal{O}(b \cdot f \cdot k)$ where b is the number of baskets, f is the number of frequent sets and k is the size of each set ($\mathcal{O}(k)$ is the complexity to check if the set is subset of basket).

Example output (s=1000, c=0.5, k=3)

```
Counting all singletons for 100000 baskets

Total unique items to count: 870

Number of frequent singletons 375

Filtering out baskets with no frequent itemsets..
Processing frequent items for 2-sets, approximately 70312.5 sets to check and 99933 baskets
Filtering out baskets with no frequent itemsets..
Finding association rules for 2-sets
Processing frequent items for 3-sets, approximately 40.5 sets to check and 7087 baskets
Filtering out baskets with no frequent itemsets..
Finding association rules for 3-sets
Processing frequent items for 4-sets, approximately 0.5 sets to check and 1035 baskets
Filtering out baskets with no frequent itemsets..
Finding association rules for 4-sets
Done. Evaluating

Number Frequent Items of length 1: 375

Number Frequent Items of length 2: 9
Number of association rules for itemsets length: 2: 3

Association Rule: AssociationRule(Set(Item(227)),Item(390)),
confidence: 0.577007700770077,
interest: 0.550157700770077
Association Rule: AssociationRule(Set(Item(704)),Item(825)),
confidence: 0.6142697881828316,
interest: 0.5834197881828316
Association Rule: AssociationRule(Set(Item(704)),Item(39)),
confidence: 0.617056856187291,
interest: 0.574476856187291

Number Frequent Items of length 3: 1
Number of association rules for itemsets length: 3: 3

Association Rule: AssociationRule(Set(Item(825), Item(704)),Item(39)),
confidence: 0.9392014519056261,
interest: 0.8966214519056261
Association Rule: AssociationRule(Set(Item(39), Item(704)),Item(825)),
confidence: 0.9349593495934959,
interest: 0.9041093495934959
Association Rule: AssociationRule(Set(Item(39), Item(825)),Item(704)),
confidence: 0.8719460825610783,
interest: 0.8540060825610784

Number Frequent Items of length 4: 0
Number of association rules for itemsets length: 4: 0

[success] Total time: 189 s, completed 2017-nov-16 11:09:11
```