Heidelberg University
Winter semester 2017/2018

Group Parallel and Distributed Systems (PVS)
Artur Andrzejak, Diego Costa

Problem Set 9 for lecture Mining Massive Datasets

Due January 8, 2018, 11:59 pm

Exercise 1 **(4 points)**

Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, and so on. Basket 12 consists of items {1, 2, 3, 4, 6, 12}, since these are all the integers that divide 12. Answer the following questions (without programming) and explain how you have obtained the solution:

a) If the support threshold is 5, which items are frequent?

b) If the support threshold is 5, which pairs of items are frequent?

c) What is the sum of the sizes of all the baskets?

d) What is the confidence of the following association rules $R_1 = \{5, 7\} \rightarrow 2$ and $R_2 = \{2, 3, 4\} \rightarrow 5$.

Exercise 2 **(3 points)**

Using the same setup of the Exercise 1, apply the A-Priori Algorithm ("on paper", i.e. without programming) with support threshold 5. Consider $k = 3$, (frequent items, pairs and triples) and submit as your solution the results of each pass of the algorithm.

Exercise 3 **(3 points)**

Consider the collection of twelve baskets depicted below. Each contains three of the six items 1 through 6.

{1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6}

{1, 3, 5} {2, 4, 6} {1, 3, 4} {2, 4, 5}

{3, 5, 6} {1, 2, 4} {2, 3, 5} {3, 4, 6}

Suppose the support threshold is 4. On the first pass of the PCY Algorithm we use a hash table with 11 buckets, and the set {i, j} is hashed to bucket i × j mod 11. Answer the following questions (without programming) and explain how you have obtained your results.

a) By any method, compute the support for each item and each pair of items.

b) Which pairs hash to which buckets?

c) Which buckets are frequent?

d) Which pairs are counted on the second pass of the PCY Algorithm?

Exercise 4 (Bonus, 3 points)

Let there be I items in a market-basket data set of B baskets. Suppose that every basket contains exactly K items. As a function of I, B, and K:

a) How much space does the triangular-matrix method take to store the counts of all pairs of items, assuming four bytes per array element?

b) What is the largest possible number of pairs with a nonzero count?

c) Under what circumstances can we be certain that the triples method will use less space than the triangular array?