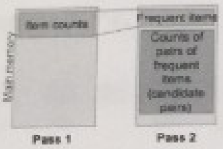
10/10

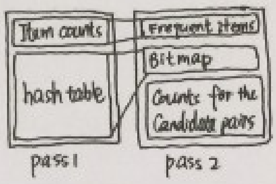
Quiz #4: Frequent Itemsets Week 2

Name: Yijun Lin ID: 3689281438

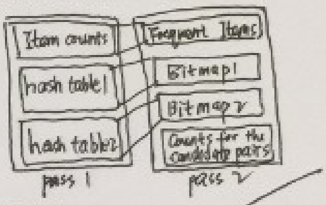
1) (3pts) The figure below shows you the memory footprint of the **A Priori algorithm for counting frequent pairs**. Please draw the memory footprint for counting frequent pairs in (a) the PCY algorithm, (b) the Multistage algorithm, and (c) the Multihash algorithm. You need to clearly label each block in your answer like in the example. (1pt each)



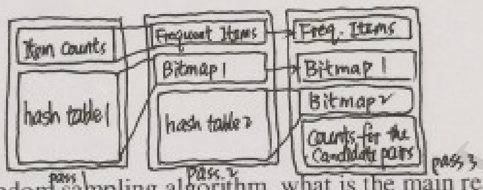
(a)



c)



(b)



3

2) (3pts) When we run the random sampling algorithm, what is the main reason that we can find frequent itemsets of all sizes in one I/O pass not just pairs? (1pt) Briefly explain how you would pick samples from the basket file (1pt). Briefly explain why and how you would need to adjust the support threshold for running the random sampling algorithm (1pt).

(1) Because we take a small sample from the whole input file, there are enough space for us to find the frequent itemsets in the main memory. And we don't need to read that file again, which can ~~deal~~ reduce IO cost

(2) For each basket, there will be a probability P for it to be choosed into the sample. ~~They~~ There will be ~~abt~~ roughly pm baskets in the sample

(3) Assume the fraction of sample of baskets is P and the threshold is S, Then we can make a threshold for the sample as PS, or a little bit lower like 0.9PS. In this way, we can reduce the number of false negatives.

3

3) (4pts) Considering the Toivonen's algorithm, give one example of a singleton and one example of a pair in the negative border. You need to explain why your examples are considered as itemsets in the negative border (1pt each). Explain how and why the Toivonen's algorithm uses the itemsets in the negative border (2pts).

(1) If item {A} is not frequent in the sample. It is in the negative border, because its only subset ϕ is always frequent.

If item {A} and {B} is frequent, but {A,B} is not frequent, then {A,B} is in the negative border.

Because all its subsets {A} and {B} are frequent. and it self is not frequent.

4

(2) In the Toivonen's algorithm. we find all frequent items in the sample with a lower threshold and we also find the infrequent items, but all their subsets are frequent in the negative border.

And then we go through the whole file, to count all the frequent items in the sample and candidates in the negative border. If no candidates in the negative border are frequent in the whole file. then we can say the frequent itemsets are the same as those in the whole file. If there are some candidates are frequent in the whole file. then we had to rechoose the sample and repeat the whole algorithm.

In this way, we can avoid both false positives and false negatives.