Feedback — Week 2C (Advanced)

Help Center

You submitted this quiz on **Sat 14 Feb 2015 1:39 PM PST**. You got a score of **2.00** out of **2.00**.

Question 1

Suppose we perform the PCY algorithm to find frequent pairs, with market-basket data meeting the following specifications:

- *s*, the support threshold, is 10,000.
- There are one million items, which are represented by the integers 0,1,...,999999.
- There are 250,000 frequent items, that is, items that occur 10,000 times or more.
- There are one million pairs that occur 10,000 times or more.
- There are *P* pairs that occur exactly once and consist of 2 frequent items.
- No other pairs occur at all.
- Integers are always represented by 4 bytes.
- When we hash pairs, they distribute among buckets randomly, but as evenly as possible; i.e., you may assume that each bucket gets exactly its fair share of the *P* pairs that occur once.

Suppose there are *S* bytes of main memory. In order to run the PCY algorithm successfully, the number of buckets must be sufficiently large that most buckets are not large. In addition, on the second pass, there must be enough room to count all the candidate pairs. As a function of *S*, what is the largest value of *P* for which we can successfully run the PCY algorithm on this data? Demonstrate that you have the correct formula by indicating which of the following is a value for *S* and a value for *P* that is approximately (i.e., to within 10%) the largest possible value of *P* for that *S*.

Your Answer		Score	Explanation
○ S = 1,000,000,000; P = 35,000,000,000			
○ S = 200,000,000; P = 400,000,000			
○ S = 200,000,000; P = 1,600,000,000			
◉ S = 500,000,000; P = 5,000,000,000	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

On the first pass, we need 4,000,000 bytes to count items, and the remaining space is used for

buckets. Since we need 4 bytes/bucket, we can use S/4 - 1,000,000 buckets. As S is much larger than a million in all choices, we shall approximate the number of buckets as S/4.
The P infrequent pairs that occur exactly once are expected to distribute evenly among buckets, so there are 4P/S of these per bucket. In order that buckets containing only infrequent pairs be infrequent, we need 4P/S < 10,000, or P < 2500S. This relationship holds easily in all choices, so we shall assume that only buckets containing one of the 1,000,000 frequent pairs are frequent buckets. The number of pairs in a frequent bucket is thus 1+4P/S. Since 4P/S is much larger than 1 in all choices, we shall estimate the number of candidate pairs as 4,000,000P/S. Note that all these pairs consist of two frequent items, so none are excluded from counting during the second pass.

On the second pass, we can neglect the space needed to store the frequent items. PCY requires a hash table of candidate pairs, so we use 12 bytes for each of the 4,000,000P/S candidate pairs. In order for there to be enough space for all these counts, we need S >= 48,000,000P/S, or $P < S^2/48{,}000{,}000$. Since we are asked for the largest possible P, we equate the two sides. For instance, if S = 200,000,000, then $P = 4*10^{16}/4.8*10^7$, or about P = 833,000,000.

Question 2

During a run of Toivonen's Algorithm with set of items {A,B,C,D,E,F,G,H} a sample is found to have the following maximal frequent itemsets: {A,B}, {A,C}, {A,D}, {B,C}, {E}, {F}. Compute the negative border. Then, identify in the list below the set that is NOT in the negative border.

Your Answer		Score	Explanation
○ {C,D}			
○ {A,E}			
○ {B,D}			
◉ {D}	✔	1.00	Correct! This set is not in the negative border because it is itself frequent. We know it is frequent because is it a subset of maximal frequent itemset {A,D}.
Total		1.00 / 1.00	