CSC 485B Assignment 4
Leanne Feng V00825004
Zhenyu Zhang V00793557

2. (6 pt) Here is a collection of twelve baskets. 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.

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

Itemset	Support
{1}	4
{2}	6
{3}	8
{4}	8
{5}	6
{6}	4

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

b) Which pairs hash to which buckets? Which buckets are frequent?

Bucket 0:	Count: 0
Bucket 1: {3, 4} {2, 6} {3, 4} {3, 4} {3, 4}	Count: 5
Bucket 2: {1, 2} {4, 6} {4, 6} {1, 2} {4, 6}	Count: 5
Bucket 3: {1, 3} {1, 3} {1, 3}	Count: 3
Bucket 4: {3, 5} {3, 5} {3, 5} {3, 5} {1, 4} {1, 4}	Count: 6
Bucket 5: {1, 5}	Count: 1
Bucket 6: {2, 3} {2, 3} {2, 3}	Count: 3
Bucket 7: {3, 6} {3, 6}	Count: 2
Bucket 8: {2, 4} {5, 6} {5, 6} {2, 4} {2, 4} {2, 4}	Count: 6
Bucket 9: {4, 5} {4, 5} {4, 5}	Count: 3
Bucket 10: {2, 5} {2, 5}	Count: 2

Bucket 1, 2, 4, 8 are frequent.

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

Frequent item: {1}, {2}, {3}, {4}, {5}, {6}

In 1, 2, 4, 8 bucket.

{3, 4} {2, 6} {1,2} {4, 6} {3, 5} {1, 4} {2, 4} {5, 6} are counted.

Suppose now we run the Multistage Algorithm. The first pass is the same previously, and for the second pass, we hash pairs to nine buckets, using the hash function that hashes {i, j} to bucket i + j mod 9.

d) Determine the counts of the buckets on the second pass. Does the second pass reduce the set of candidate pairs?

Bucket 0:	Count: 0
Bucket 1: {4, 6}	Count: 3
Bucket 2: {5, 6}	Count: 2
Bucket 3: {1, 2}	Count: 3
Bucket 4:	Count: 0
Bucket 5: {1, 4}	Count: 2
Bucket 6: {2, 4}	Count: 4
Bucket 7: {3, 4}	Count: 4
Bucket 8: {2, 6}, {3, 5}	Count: 5

Yes, the second pass reduce the set of candidate pairs. Bucket 6, 7 and 8 are frequent.

Note that all items are frequent, so the only reason a pair would not be hashed on the second pass is if it hashed to an infrequent bucket on the first pass.

Suppose now we run the Multihash Algorithm. We shall use two hash tables with five buckets each. For one, the set {i, j}, is hashed to bucket 2i+3j +4 mod 5, and for the other, the set is hashed to i + 4j mod 5. Since these hash functions are not symmetric in i and j, order the items so that i < j when evaluating each hash function.

e) Determine the counts of each of the 10 buckets.

For hash function 2i + 3j + 4 mod 5

Bucket 0: {1, 3} {2, 4} {3, 5} {4, 6}	Count: 14
Bucket 1: {1, 5} {2, 6}	Count: 2
Bucket 2: {1, 2} {2, 3} {3, 4} {4, 5} {5, 6}	Count: 14
Bucket 3: {1, 4} {2, 5} {3, 6}	Count: 6
Bucket 4:	Count: 0

For hash function i + 4j mod 5

Bucket 0:	Count: 0
Bucket 1: {1, 5} {2, 6}	Count: 2
Bucket 2: {1, 4} {2, 5} {3, 6}	Count: 6
Bucket 3: {1, 3} {2, 4} {3, 5} {4, 6}	Count: 14
Bucket 4: {1, 2} {2, 3} {3, 4} {4, 5} {5, 6}	Count: 14

f) How large does the support threshold have to be for the Multistage Algorithm to eliminate more pairs than the PCY Algorithm would in this example?（Does there exist a support threshold for the Multihash Algorithm to eliminate more pairs than the PCY Algorithm would in this example?)

Yes, there exist. When support threshold is 3, the Multihash Algorithm eliminates 5 pairs, and pcy algorithm eliminates 6 pairs. So support threshold have to be 4.

3. (3 pt) Apply Toivonen's Algorithm to the data of the previous exercise with a support threshold of 4. Take as the sample the first row of baskets:

{1, 2, 3}, {2, 3, 4}, {3, 4, 5}, and {4, 5, 6}, i.e., one-third of the file.

Our scaled-down support threshold will be 1.

a) What are the itemsets frequent in the sample?

Itemset	Support	Itemset	Support
{1}	1	{4}	3
{2}	2	{5}	2

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

The itemset frequent in sample: {1} {2} {3} {4} {5} {6} {1, 2} {1, 3} {2, 3} {2, 4} {3, 4} {3, 5} {4, 5} {4, 6} {5, 6} {1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6}

b) What is the negative border?
The negative border: {1, 4}, {1, 5}, {1, 6} {2, 5}, {2, 6}, {3, 6}

c) What is the outcome of the pass through the full dataset? Are any of the itemsets in the negative border frequent in the whole?

Itemsets {1, 4}, {1, 5}, {1, 6}, {2, 5}, {2, 6}, {3, 6} from the negative border are not frequent in the whole. So {1} {2} {3} {4} {5} {6} {1, 2} {1, 3} {2, 3} {2, 4} {3, 4} {3, 5} {4, 5} {4, 6} {5, 6} {1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6} in the sample are frequent.

4. (4 pt) Consider the example of slide 32 of similarity.pdf

a. What is the probability that we miss a pair of similar columns when the (Jaccard) similarity threshold is 60%? The values of b and r are as in the example: 20 and 5, Respectively.

$(0.6)\hat{}5 = 0.07776$

$(1 - 0.07776)\hat{}20 = 0.198 \approx 20\%$

The probability that we miss a pair of similar columns when the (Jaccard) similarity threshold is 60% is 20%.

b. How should we change b and r such that the probability of missing a pair of similar columns (for a 60% threshold) is about 1/3000?

$(1 - 0.6^{r})^{b} \approx 1/3000$

When b is smaller or r is larger, the probability of missing a pair of similar columns (for a 60% threshold) is about 1/3000.