CptS 315: Introduction to Data Mining

Homework 1

(Due date: Feb 11th midnight)

Q1. Consider the following market-basket data, where each row is a basket and shows the list of items that are part of that basket.

1. {A, B, C}

2. {A, C, D, E}

3. {A, B, F, G, H}

4. {A, B, X, Y, Z}

5. {A, C, D, P, Q, R, S}

6. {A, B, L, M, N}

a) What is the absolute support of item set {A, B} ? (3 points)

4

b) What is the relative support of item set {A, B} ? (3 points)

0.667

c) What is the confidence of association rule A ⇒ B ? (3 points)

0.667

Q2. Answer the below questions about storing frequent pairs using triangular matrix and tabular method.

a) Suppose we use a triangular matrix to count pairs and the number of items n = 20. If we store this triangular matrix as a ragged one-dimensional array Count, what is the index where count of pair (7, 8) is stored? (3 points)

The index is: $(I - 1) * (n - (I / 2)) + j - I = 100$

b) Suppose you are provided with the prior knowledge that only ten percent of the total pairs will have a non-zero count. In this case, which method among triangular matrix and tabular method should be preferred and why? (3 points)

If there are low percent of the total pairs that will have a non-zero count, we prefer to use tabular method. Tabular approach beats triangular matrix only when at most 1/3 of the total pairs have a nonzero count.

Q3. This question is about the PCY algorithm for counting frequent pairs of items. Suppose we have six items numbered 1, 2, 3, 4, 5, 6. Consider the following twelve baskets.

1. {1, 2, 3}
2. {2, 3, 4}
3. {3, 4, 5}
4. {4, 5, 6}
5. {1, 3, 5}
6. {2, 4, 6}
7. {1, 3, 4}
8. {2, 4, 5}
9. {3, 5, 6}
10. {1, 2, 4}
11. {2, 3, 5}
12. {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 i × j mod 11.

a) By any method, compute the support for each item and each pair of items. (5 points)

1-item

Items	1	2	3	4	5	6
Support	4	6	8	8	6	4

2-items

Items	1,2	1,3	1,4	1,5	1,6	2,3	2,4	2,5	2,6
Suppurt	2	3	2	1	0	3	4	2	1

3,4	3,5	3,6	4,5	4,6	5,6
4	4	2	3	3	2

b) Which pairs hash to which buckets? (5 points)

{I,j}	1,2	1,3	1,4	1,5	1,6	2,3	2,4
Hash ({I,j})	2	3	4	5	6	6	8
2,5	2,6	3,4	3,5	3,6	4,5	4,6	5,6
10	1	1	4	7	9	2	8

Hash	1	2	3	4	5	6	7	8	9	10	11

(Buckets)											
pairs	{2,6} {3,4}	{1,2} {4,6}	{1,3}	{1,4} {3,5}	{1,5}	{1,6} {2,3}	{3,6}	{2,4} {5,6}	{4,5}	{2,5}	/

c) Which buckets are frequent? (3 points)

bucket	1	2	3	4	5	6	7	8	9	10	11
count	5	5	3	6	1	3	2	6	3	2	0

d) Which pairs are counted on the second pass of the PCY algorithm? (2 points)

Items {1,2,3,4,5,6} are frequent items. We got frequent buckets 1,2,4,8, so {2,6},{3,4},{1,2},{4,6},{1,4},{3,5},{2,4},{5,6} are counted on the second pass of the PCY algorithm.

Q4. Please read the following paper and write a brief summary of the main points in at most ONE page. You can skip the theoretical parts. (10 points)

Saul Schleimer, Daniel Shawcross Wilkerson, Alexander Aiken: Winnowing: Local Algorithms for Document Fingerprinting. SIGMOD Conference 2003: 76-85
https://theory.stanford.edu/~aiken/publications/papers/sigmod03.pdf

Summary

The article introduces the experience of two different implementations of winnowing. The first experiment uses the rolling hash function, shows us that 64-bit hashes is the most suitable one that can avoid accidental collisions, and it reduces the throughput of the fingerprinting algorithm by more than a factor of four when using the k- gram. Early in the experiment, they tested the rationality of the hash function by 8MB of randomly generated. Make sure that the data obtained in subsequent experiments will be valid. At the same time, it proved that the implementation of the hash function is sufficient for the fingerprint printing algorithm. It's no doubt that hash and robust winnowing are important in the experiment.The second experiment which is named as plagiarism detection about winnowing algorithm needs ignore whitespace, and excludes some acceptable copies such as disclaimer and something people do not interest at.

The approximate sequence of execution is: first remove irrelevant content from the original text like some symbols. Then divided into k-grams and start use hash, define a window of size w to split the hash, make sure at least one hash value is selected for each window. Using windows can avoid the problem of too large distance between hashes. The strategy adopted by the Winnowing algorithm is pick the smallest hash value in each window (obviously two windows may have the same minimum value). If there are multiple minimums, the rightmost one is selected. This strategy not only guarantees sufficient fingerprint information, but also does not generate too large fingerprints.

In conclusion, the main purpose of the wind selection algorithm is to select the required hash value.