Kate Radford

11537868

CptS 315 HW 1

| 1 |

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

4 (baskets)

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

absolute support / total = 4 / 6 = 0.6

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

(# A & B) / (# A) = 4 / 6 = 0.6

| 2 |

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)

Find pair {i, j}, where i<j, at the position: (i-1)(n-i/2)+j-i

With n=20 and {I,j}={7,8}: (7 - 1)(20 - (7/2)) + 8 -7 = 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)

The tabular approach beats triangular matrix only when at most 1/3 (30%) of all pairs have a nonzero count. Since we are given that only 10% will have a non-zero count, the tabular method as it uses less memory.

| 3 | 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| item | 1 | 2 | 3 | 4 | 5 | 6 |
| support | 4 | 6 | 8 | 8 | 6 | 4 |

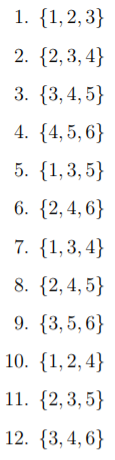
a) By any method, compute the support for each item and each pair of items. (5 points)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Item Pair | {1,2} | {1,3} | {1,4} | {1,5} | {1,6} | {2,3} | {2,4} | {2,5} | {2,6} | {3,4} | {3,5} | {3,6} | {4,5} | {4,6} | {5,6} |
| Support | 2 | 3 | 2 | 1 | 0 | 3 | 4 | 2 | 1 | 4 | 4 | 2 | 3 | 3 | 2 |

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Item Pair | {1,2} | {1,3} | {1,4} | {1,5} | {1,6} | {2,3} | {2,4} | {2,5} | {2,6} | {3,4} | {3,5} | {3,6} | {4,5} | {4,6} | {5,6} |
| hash | 2 | 3 | 4 | 5 | 6 | 6 | 8 | 10 | 1 | 1 | 4 | 7 | 9 | 2 | 8 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| bucket | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 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}{3,5} | {4,5} | {2,5} |
| support | 1+4=5 | 2+3=5 | 3 | 2+4=6 | 1 | 0+3=3 | 2 | 4+4=8 | 3 | 2 |

c) Which buckets are frequent? (3 points)

Therefore, buckets 1, 2, 4, and 8 are frequent.

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

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

| 4 | 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)

A variety of fingerprinting algorithms were developed to detect any amount of copying of any digital document. Detecting partial copies is more complicated than finding exact copies and is a growing problem in our digital world.

The algorithm should have three properties for proper detection: whitespace insensitivity, noise suppression and position independence. Whitespace refers to all types of text changes that are irrelevant, like extra whitespace, capitalization, punctuation, variable names (specific to code matching). Noise suppression means we want to focus on matching unique, specific, and uncommon words or strings (all the rest is just ‘noise’). Lastly, the position, order, or addition of new text to the text should have no influence on copy detection. The last requirement is the most challenging and schema methods vary.

A standard algorithm in fingerprinting is called winnowing and considered to be efficient with a 33% lower bound.