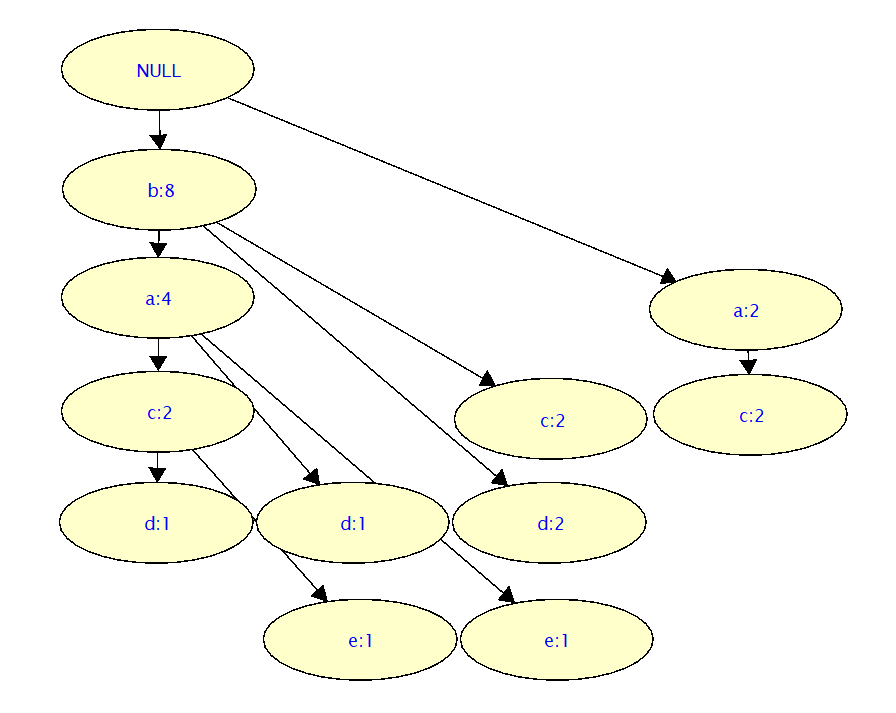
Michael Wang

HW 5

1a)

1. Scan for frequent 1-itemset: **a = 6, b = 8, c = 6, d = 4, e = 2**
2. Generate length-2 candidates: ab, ac, ad, ae, bc, bd, be, cd, ce, de
3. (No need for pruning yet) Test candidates: **ab = 4, ac = 3, ad = 2, ae = 2, bc = 4, bd = 3, be = 2**
4. Generate length-3 candidates: abc, abd, abe, acd, ace, ade, bcd, bce, bde
5. Prune the candidates: abc, abd, abe
6. Test the candidates: **abc = 2, abd = 2, abe = 2**
7. Generate length-3 candidates: abcd, abce, abde
8. Prune the candidates: None

1b)

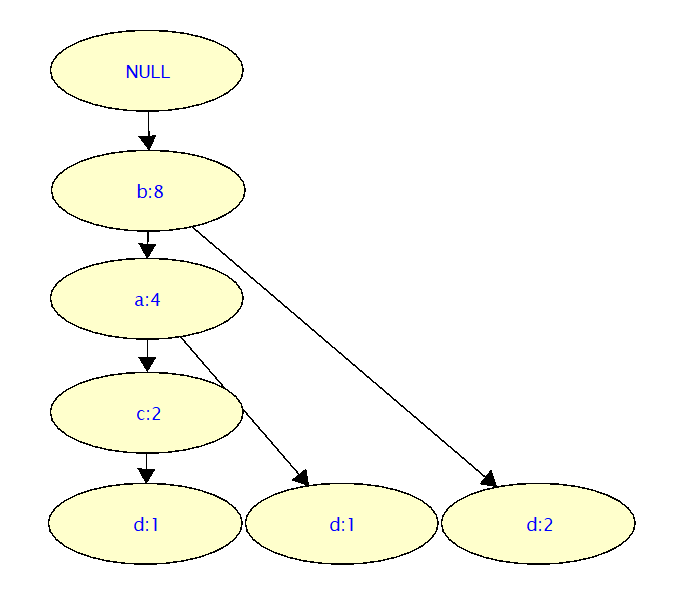


|  |  |
| --- | --- |
| TID | Ordered Frequent Items |
| 1 | b,c |
| 2 | b,a,d |
| 3 | a,c |
| 4 | b,d |
| 5 | b,a,c,e |
| 6 | b,c |
| 7 | a,c |
| 8 | b,a,e |
| 9 | b,d |
| 10 | b,a,c,d |

|  |  |
| --- | --- |
| Item | Frequency |
| b | 8 |
| a | 6 |
| c | 6 |
| d | 4 |
| e | 2 |

1c)

bac: 1, ba: 1, b: 1



1d)

d, bd, bad

2a)

------------------------ FREQUENT PATTERNS

Itemset: "xxx", "zzz" , 60.00%

Itemset: "yyy", "zzz" , 60.00%

Itemset: "xxx", "yyy", "zzz" , 60.00%

Itemset: "xxx", "yyy" , 80.00%

----> 4 printed

------------------------ RULES:

Rule: "xxx" ==> "yyy" , 0.800

Rule: "yyy" ==> "xxx" , 1.000

Rule: "zzz" ==> "xxx" , 1.000

Rule: "zzz" ==> "yyy" , 1.000

Rule: "zzz" ==> "xxx", "yyy" , 1.000

Rule: "xxx", "zzz" ==> "yyy" , 1.000

Rule: "yyy", "zzz" ==> "xxx" , 1.000

----> 7 printed

0.0169999599457 sec

2b)

------------------------ FREQUENT PATTERNS

Itemset: "Forrest Gump (1994)", "Braveheart (1995)" , 27.33%

Itemset: "Forrest Gump (1994)", "Shawshank Redemption, The (1994)" , 28.35%

Itemset: "Pulp Fiction (1994)", "Braveheart (1995)" , 28.57%

Itemset: "Star Wars: Episode IV - A New Hope (1977)", "Star Wars: Episode V - The Empire Strikes Back (1980)" , 28.57%

Itemset: "Pulp Fiction (1994)", "Forrest Gump (1994)" , 30.10%

Itemset: "Pulp Fiction (1994)", "Shawshank Redemption, The (1994)" , 31.63%

----> 6 printed

------------------------ RULES:

Rule: "Braveheart (1995)" ==> "Forrest Gump (1994)" , 0.686

Rule: "Forrest Gump (1994)" ==> "Pulp Fiction (1994)" , 0.686

Rule: "Braveheart (1995)" ==> "Pulp Fiction (1994)" , 0.717

Rule: "Shawshank Redemption, The (1994)" ==> "Pulp Fiction (1994)" , 0.725

Rule: "Star Wars: Episode V - The Empire Strikes Back (1980)" ==> "Star Wars: Episode IV - A New Hope (1977)" , 0.897

----> 5 printed

2.40700006485 sec

2c)

------------------------ FREQUENT PATTERNS

Itemset: "dark comedy", "black comedy" , 0.28%

Itemset: "funny", "quirky" , 0.28%

Itemset: "future", "sci-fi" , 0.28%

Itemset: "fantasy", "adventure" , 0.29%

Itemset: "superhero", "comic book" , 0.30%

Itemset: "atmospheric", "based on a book" , 0.31%

Itemset: "nudity (topless)", "Nudity (Topless)" , 0.33%

Itemset: "sci-fi", "space" , 0.34%

Itemset: "dystopia", "sci-fi" , 0.37%

Itemset: "imdb top 250", "atmospheric" , 0.37%

Itemset: "aliens", "sci-fi" , 0.38%

Itemset: "imdb top 250", "classic" , 0.39%

Itemset: "action", "sci-fi" , 0.39%

Itemset: "stylized", "atmospheric" , 0.47%

Itemset: "funny", "comedy" , 0.57%

Itemset: "based on a book", "adapted from:book" , 0.60%

----> 16 printed

------------------------ RULES:

Rule: "future" ==> "sci-fi" , 0.729

----> 1 printed

2.1)

Assume a global frequent pattern GP has support supp within the dataset. Assume the dataset is partitioned into n sub-datasets. Looking at the first sub-dataset, either GP has the necessary support in which case we are done, or it doesn’t have enough support in which case the rest of the n-1 sub-datasets must on average have a support higher than supp for GP. Repeat this process for sub-dataset 2 and so on. If GP doesn’t have enough support in the first n-1 sub-datasets, we conclude that the final sub-dataset has on average support for GP higher than supp, i.e. if GP doesn’t have enough support in any of the previous sub-datasets then in must have the support in the last one.

2.2)

A trivial demonstration of local frequent patterns not being global frequent patterns is when the dataset is totally partitioned such that each element is in its own sub-dataset. Every pattern would then be a local frequent pattern given that it is represented exactly one out of one times within its sub-dataset. However, it is clear that not every pattern can be a global frequent pattern. Thus, local frequent patterns are not guaranteed to be global frequent patterns.

3a)

=

3b)

These measurements imply a positive correlation between beer and nuts with large lift, large chi-square, and an all confidence significantly larger than either prior probabilities.

4a)

S contains 3 elements.

S has length 6.

S contains 2^6 -1 = 63 subsequences

4b)

1. Generate length-4 candidates by joining- can join if dropping first item in p1 is same as dropping last item in p2: <(ab)(cd)><(ab)ce>
2. Pruning- check if all length-3 sub-sequences are present in L: **<(ab)(cd)>**