**6.6** A database has five transactions. Let *min sup* D 60% and *min conf* D 80%.

*TID items bought*

T100 {M, O, N, K, E, Y}

T200 {D, O, N, K, E, Y}

T300 {M, A, K, E}

T400 {M, U, C, K, Y}

T500 {C, O, O, K, I, E}

(a) Find all frequent itemsets using Apriori and FP-growth, respectively. Compare

the efficiency of the two mining processes.

**Ans:**

Apriori:

M 3 MO 1

O 3 MK 3

N 2 M 3 ME 2

K 5 O 3 MY 2

**C1** = E 4  **L1** = K 5 **C2** = OK 3 **L2** = OKE 3 **C3** = OKE 3

Y 3 E 4 OE 3 KEY 2

D 1 Y 3 OY 2

A 1 KE 4

U 1 KY 3

C 2 EY 3

I 1

F-P Growth:

With min\_sup = 3, we get Ordered Itemset as follows from

**Item S. Count Frequent Pattern**

M 3

O 3

N 2 K 5

K 5 E 4

**C1** = E 4  **L1** = M 3

Y 3 O 3

D 1 Y 3

A 1

U 1

C 2

I 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TID | Itemset | Ordered Itemset | Items | Conditional Pattern Base | Conditional F-P Tree | Frequent Pattern |
| T100 | {M, O, N, E, K, Y} | K, E, M, O, Y | Y | {{K, E, M, O:1}, {K, E, O:1}, {K, M:1}} | {K:3} | {K, Y: 3} |
| T200 | {D, O, N, K, E, Y} | K, E, O, Y | O | {{K, E, M:1}, {K, E:2}} | {KE:3} | {K, O:3}, {E, O:3}, {K, E, O:3} |
| T300 | {M, A, K, E} | K, E, M | M | {{K, C:2}, {K:1}} | {K:3} | {K, M:3} |
| T400 | {M, U, C, K, Y} | K, M, Y | E | {K, C:2}, {K:1}} | {K:4} | {K, E:4} |
| T500 | {C, O, O, K, I, E} | K, E, O | K | {K: 4} |  |  |

Root

K:5

E:4

M:2 M:1

O:1 O: 2

Y:1 Y: 1 Y:1

|  |  |  |
| --- | --- | --- |
| Item ID | S. C | N. L |
| K | 5 |  |
| E | 4 |  |
| M | 3 |  |
| O | 3 |  |
| Y | 3 |  |

Efficiency comparison: If we compare between these two trees, we notice that Apriori has to do multiple scans of the database while FP-growth builds the FP-Tree with a single scan. Candidate generation in Apriori is expensive due to self-join, while FP-growth does not generate any candidates.

(b) List all the *strong* association rules (with support *s* and confidence *c*) matching the following metarule, where *X* is a variable representing customers, and *itemi* denotes variables representing items (e.g., *“A,” “B,”*):

*transaction*, *buys*(*X*, *item*1)^*buys*(*X*, *item*2) => *buys*(*X*, *item*3) [*s*, *c*]

**Ans:**

Given min sup = 60% and min conf = 80%  
So, min sup = 60% = 60/100 \* (5 transactions) = 3 transactions

* Support (Item-set) = No. of transactions where all items in 'Item-set' are purchased
* Frequent Item-sets: A Item-set is said to be frequent if Support (Item-set) >= min-support.

*transaction; buys* (*X; item*1) *=> buys*(*X; item*2) *) buys*(*X; item*3) [*s; c*]

So, we have only one frequent itemset of length 3.  
    And for itemset [O,K, E] there are three possibilities.  
     1)[O, E] → K  
       CONF = support[O,K, E])/support([O,E])  
            =3/3 = 100%  
        Hence [O, E] → K is a strong association rule.   
     2)[O, K] → E  
       CONF = support ([O, K, E])/support ([O, K])  
            = 3/3 = 100%  
        Hence [O, K] → E is a strong association rule.    
     3)[K, E] → O  
       CONF = support ([O, K, E])/ support ([K, E])  
            = 3/4 = 75%  
       As 75% < 80%, [E, K] → O is not a strong association rule.

**6.14** The following contingency table summarizes supermarket transaction data, where *hotdogs* refer to the transactions containing hotdogs, *hotdogs* refers to the transactions that do not contain hotdogs, *hamburgers* refers to the transactions containing hamburgers, and *hamburgers* refers to the transactions that do not contain hamburgers.

|  |  |  |  |
| --- | --- | --- | --- |
| hotdogs | hotdogs | hotdogs |  |
| hamburger | 2000 | 500 | 2500 |
| hamburger | 1000 | 1500 | 2500 |
|  | 3000 | 2000 | 5000 |

(a) Suppose that the association rule “*hotdogs* => *hamburgers*” is mined. Given a

minimum support threshold of 25% and a minimum confidence threshold of

50%, is this association rule strong?

**Ans:**

For the rule, support = 2000/5000 = 40%, and confidence = 2000/3000 = 66.7%. Therefore, the

association rule is strong.

(b) Based on the given data, is the purchase of *hot dogs* independent of the purchase

of *hamburgers*? If not, what kind of *correlation* relationship exists between the

two?

**Ans:**

*Corr {hotdog, hamburger}* = *P (hotdog, hamburger)/ (P (hot dog) P(hamburger))* = 0.4/(0.6 *\** 0.5) = 1.33 *>* 1. So, the purchase of hotdogs is not independent of the purchase of hamburgers. There exists a positive correlation between the two.