DSAA 5002 - HW1

50015976 Ruiming ZHANG

Q1 [15 Marks]

Given the transaction database below,set the minimum support count to 2 and the minimum confidence level to 60% to find the strong association rule. Generate the set C_3 of the candidate 3-itemset, using prunning on Apriori principle.

Solution:

TID	Item
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

1^{st} scan
calculate sup

C_1

Itemset	sup
{A}	3
{B}	3
{C}	4
{D}	1
{E}	4

1^{st} prunning
discard sup<2

L_1

Itemset	sup
{A}	3
{B}	3
{C}	4
{E}	4

generate 2-itemsets

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

C_2

2^{nd} scan
calculate sup

Itemset	sup
{A, B}	1
{A, C}	3
{A, E}	2
{B, C}	2
{B, E}	3
{C, E}	3

C_2

2^{nd} prunning
disgard sup<2

L_2

Itemset	sup
{A, C}	3
{A, E}	2
{B, C}	2
{B, E}	3
{C, E}	3

generate the candidate 3-itemset C_3

Itemset
{A, C, E}
{B, C, E}

C_3

3^{rd} scan

Itemset	sup
{A, C, E}	2
{B, C, E}	2

L_3

Thus we get the candidate 3-itemset C_3.

Q2 [15 Marks]

Reducing the transactions using dynamic hashing and prunning(DHP) algorithm. Set the minimum support count to 2.

Hash function bucket #= h({x y}) = ((order of x)*10+(order of y)) % 7

Solution

TID	Item
T1	A, B, C
T2	B, D, E
T3	A, B, D, E
T4	B, E

1^{st} scan →

C_1

Itemset	sup
{A}	2
{B}	4
{C}	1
{D}	2
{E}	3

1^{st} prunning →

L_1

Itemset	sup
{A}	2
{B}	4
{D}	2
{E}	3

Find all 2-itemset of each transaction ↓

TID	2-itemset
T1	{A B} {A C} {B C}
T2	{B D} {B E} {D E}
T3	{A B} {A D} {A E} {B D} {B E} {D E}
T4	{B E}

Because we have:

- Items = A, B, C, D, E,
- Order = 1, 2, 3, 4, 5
- Hash function : h({x y}) = ((order of x)*10+(order of y)) % 7

Thus we have the hash table below:

bucket	0	1	2	3	4	5	6
count	1	1	1	4	3	2	1
2-itemset	{A D}	{A E}	{B C}	{B D} {D E} {B D} {D E}	{B E} {B E} {B E}	{A B} {A B}	{A C}

Let L_1*L_1 to generate a 2-itemset table, and choose the itemsets where the number of content in its bucket is above the minimum support.

$L_1 * L_1$	# in the bucket
{A B}	2
{A D}	1
{A E}	1
{B D}	4
{B E}	3
{D E}	4

C_2

Resulted C_2
{A B}
{B D}
{B E}
{D E}

L_2

C_2	sup
{A B}	2
{B D}	2
{B E}	3
{D E}	2

Because if an item occurs in a frequent (k+1)-itemset, it must occur in at least k candidate k-itemsets.

TID	Item	2-itemset occurs	
T1	A, B, C	{A B}	Disgard
T2	B, D, E	{B D} {B E} {D E}	Keep {B D E}
T3	A, B, D, E	{A B} {B D} {B E} {D E}	Keep {B D E}
T4	B, E	{E E}	Disacrd

TID	Item
T2	B, D, E
T3	B, D, E

Thus we have reduced the transactions.

Q3 [35 Marks]

An itemset X is said to be a frequent itemset if the frequency count of X is at least a given support threshold.

An itemset Y is a proper super-itemset of X if X $\subset$ Y and X $\neq$ Y.

An itemset X is said to be a closed frequent itemset if (1) X is frequent and (2) there exists no proper super itemset Y of X such that Y is frequent and Y has the same frequency count as X.

An itemset X is said to be a maximal frequent itemset if (1) X is frequent and (2) there exists no proper super-itemset Y of X such that Y is frequent.

Let F_c be the set of (traditional) frequent itemsets each of which is associated with a frequency in the dataset.

For example, if there are three frequent itemsets, $\{I_1\}$ with frequency4, $\{I_2\}$ with frequency 5, and $\{I_1, I_2\}$ with frequency 3, F={$\{I_1\}$, $\{I_2\}$, $\{I_2, I_2\}$} and F_c = {<$\{I_1\}$, 4>, <$\{I_2\}$, 5>, <$\{I_1, I_2\}$, 3>}.

Similarly, let C be the set of closed frequent itemsets without specifying the frequency of itemsets.

Let C_c be the set of closed frequent itemsets each of which is associated with a frequency of itemsets.

Let M be the set of maximal frequent itemsets without specifying the frequency of itemsets.

Ler M_c be the set of maximal frequent itemsets each of which is associated with a frequency in the dataset.

The following shows six transactions with four items. Each row corresponds to a transaction where 1 corresponds to a presence of an item and 0 corresponds to an absence.

A	B	C	D
0	0	1	1
1	1	0	0
0	0	1	1
1	0	1	1
1	0	0	0
0	0	0	1