Supplementary Notes #07

Data Mining and Data Warehousing

Solutions to exercises on Sequential Pattern Discovery

Answer:

Phase I: Sort Phase

- Represent each bidder and auction item with Bidder ID and Auction Item ID respectively.
- Sort the database with Bidder ID as the major key and transaction-time as the minor key.

Bidder ID	Bidder
1	Tony
2	John
3	Ivy
4	David
5	Edith
6	Polly
7	Mandy

Auction Item ID	Auction Item
1	DVD
2	Movie Ticket
3	Watch
4	Toy
5	Clothes
6	Phone

Transaction Date	Bidder ID	Auction Item ID		
16-Mar	1	1	2	3
20-Mar	1	2		
23-Mar	1	3		
24-Mar	1	4		
17-Mar	2	4	5	
27-Mar	2	4	5	
21-Mar	3	1	5	
22-Mar	3	2	6	
23-Mar	3	1	3	
25-Mar	3	4	5	
18-Mar	4	1	4	
28-Mar	4	1	2	4
30-Mar	4	3	4	
22-Mar	5	2	4	5
26-Mar	5	3		
19-Mar	6	1		
20-Mar	6	2	5	
21-Mar	6	1	3	4
23-Mar	6	4	5	
29-Mar	7	2	3	
31-Mar	7	5	6	

Phase 2: Large Itemset Phase

- Find all large itemsets with support = 40% (i.e. count >= 3)

Identifier	Auction Item ID	count
1	1	7
2	2	7
3	3	7
4	4	10
5	5	8
6	1 3	3
7	4 5	4

Phase 3: Transformation Phase

- Delete non-large itemsets
- Map large itemsets into integer ID

Bidder ID	Original Sequence			Transformed Sequence	After Mapping
1	1	2	3	{(1),(2),(3),(1 3)} {(2)} {(3)} {(4)}	**{1,2,3,6} {2} {3} {4}**
	2				
	3				
	4				
2	4	5		{(4),(5),(4 5)} {(4),(5),(4 5)}	**{4,5,7} {4,5,7}**
	4	5			
3	1	5		{(1),(5)} {(2)} {(1),(3),(1 3)} {(4),(5),(4 5)}	**{1,5} {2} {1,3,6} {4,5,7}**
	2				
	1	3			
	4	5			
4	1	4		{(1),(4)} {(1),(2),(4)} {(3),(4)}	**{1,4} {1,2,4} {3,4}**
	1	2	4		
	3	4			
5	2	4	5	{(2),(4),(5),(4 5)} {(3)}	**{2,4,5,7} {3}**
	3				
6	1			{(1)} {(2),(5)} {(1),(3),(4),(1 3)} {(4),(5),(4 5)}	**{1} {2,5} {1,3,4,6} {4,5,7}**
	2	5			
	1	3	4		
	4	5			
7	2	3		{(2),(3)} {(5)}	**{2,3} {5}**
	5				

Phase 4: Sequence Phase

- Use the set of itemsets to fiind the desired sequence with AprioriAll

Large 1-sequences	count	support
<1>	4	57%
<2>	6	86%
<3>	6	86%
<4>	6	86%
<5>	5	71%
<6>	3	43%
<7>	4	57%

Large 2-sequences	count	support
<1 1>	3	43%
<1 2>	4	57%
<1 3>	4	57%
<1 4>	4	57%
<2 3>	5	71%
<2 4>	4	57%
<2 5>	3	43%
<3 4>	3	43%
<3 5>	3	43%
<4 4>	3	43%
<5 3>	3	43%
<5 4>	3	43%
<5 5>	3	43%
<5 7>	3	43%
<6 4>	3	43%

Large 3-sequences	count	support
<1 1 4>	3	43%
<1 2 3>	4	57%
<1 2 4>	4	57%
<1 3 4>	3	43%
<2 3 4>	3	43%

Large 3-sequences	count	support
<1 2 3 4>	3	43%

Phase 5: Maximal Phase

- The maximal large sequences are:

<1 2 3 4>
<1 1 4>
<2 5>
<3 5>
<4 4>
<5 3>
<5 4>
<5 5>
<5 7>
<6 4>

i.e.

DVD → Movie Ticket → Watch → Toy
DVD → DVD → Toy
Movie Ticket → Clothes
Watch → Clothes
Toy → Toy
Clothes → Watch
Clothes → Toy
Clothes → Clothes
Clothes → (Toy, Clothes)
(DVD, Watch) → Toy