COMP4433 Data Mining & Data Warehousing Applications

Assignment 1 (suggested answers only)

1. In a survey, the following data was collected.

   □ Among 5000 teenagers who wear jeans,
   - ○ 3000 play on-line games
   - ○ 3750 eat chips
   - ○ 2000 both play on-line games and eat chips

   □ Among another 5000 teenagers who do not wear jeans,
   - ○ 3000 play on-line games
   - ○ 4000 eat chips
   - ○ 2250 both play on-line games and eat chips

   a) List ALL strong association rules having the form {item1, item2⇒eat chips} with support≥20% and confidence≥50%.

   Ans.

   For wearing jeans, we have

	Game	^ Game	Sum(row)
Chip	2000	1750	3750
^ Chip	1000	250	1250
Sum(column)	3000	2000	5000

For non-wearing jeans, we have

	Game	^ Game	Sum(row)
Chip	2250	1750	4000
^ Chip	750	250	1000
Sum(column)	3000	2000	5000

Frequent itemsets (min_supp=20%)
{Jeans, Games, Chips} (supp=2000/10000)
{^Jeans, Games, Chips} (supp=2250/10000)

Strong Rules (min_conf=50%)
Jeans, Games ⇒Chips (conf=2000/3000) √
^Jeans, Games ⇒Chips (conf=2250/3000) √

(15 marks)

b) Compute the interest (lift ratio) of the strong association rules found in part (a).

Ans.
Strong Rules
Jeans, Games ⇒Chips (Interest=20%/30%/77.5%~=0.86)
^Jeans, Games ⇒Chips (Interest =22.5%/30%/77.5%~=0.968)

(5 marks)

2. Consider the following stock transactions for association analysis.

Table I Stock Transaction Data

Stock	Transactions made by 10 selected investors today									
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
MSFT	Buy	Buy	Buy		Buy	Buy		Buy		Buy
NFLX	Sell	Buy		Buy		Sell	Sell	Sell		Sell
TSLA	Buy		Buy	Buy	Sell		Buy		Buy	Buy
ZM				Buy	Sell		Buy		Sell	Sell

That is "today investor #1 buys MSFT and TSLA but sells NFLX, investor #2 buys MSFT and NFLX, investor #3 …, and investor #10 buys MSFT, TSLA but sells NFLX and ZM".

a) Compute the support and confidence of the association rules:

   i) Buy MSFT ⇒ Sell NFLX

   ii) Buy MSFT, Buy TSLA⇒ *

   Note here that you don't need to apply the Apriori algorithm and * is a wild card (an unknown itemset in this case and you may assume that an empty box in Table I does not form an item). You may need to think about ALL rules satisfying this form and compute the corresponding support and confidence.

Ans.

a-i) Supp=4/10; Conf=4/7

(5 marks)

a-ii)

Possible Answer:

Considering Table 1 as a transactional form, i.e., the empty boxes carry no information (no item involved), we only have

R1: Buy MSFT, Buy TSLA⇒Sell NFLX

Supp=2/10; Conf=2/3

R2: Buy MSFT, Buy TSLA⇒ Sell NFLX, Sell ZM

Supp=1/10; Conf=1/3

Hence, for Buy MSFT, Buy TSLA⇒ *

Supp=2/10; Conf=2/3 (support from transaction #1 and #10; note that both R1 & R2 are referring to these two transactions)

(10 marks)

b) Find all frequent itemsets using the Aprior algorithm for min_support=20% (i.e., 2 transactions).

Ans. For min _sup=20% (i.e., 2 transactions)

1-itemset	Count	2-itemset	Count	3-itemset	Count
B-MSFT	7	**B-MSFT, S-NFLX**	4	**B-MSFT, S-NFLX, B-TSLA**	2
~~S-MSFT~~	~~0~~	**B-MSFT, B-TSLA**	3	~~**S-NFLX, B-TSLA, B-ZM**~~	1
B-NFLX	2	**S-NFLX, B-TSLA**	3		
S-NFLX	5	~~**S-NFLX, B-ZM**~~ **B-MSFT, S-ZM**	2		
B-TSLA	6	**B-TSLA, B-ZM**	2		
~~S-TSLA~~	~~1~~	**B-TSLA, S-ZM**	2		
B-ZM	~~3~~ 2				
S-ZM	~~2~~ 3				

We should have C_2^6 candidate 2-itemsets here but only frequent ones are listed above

The frequent 1-itemsets, 2-itemsets and 3-itemsets are bolded.

(15 marks)

3. A social network is a social structure made up of a set of users and a set of social ties such as friendship between them. In view of the continuously evolving social network data, you are asked by a social networking company to carry out the following data mining task. After interviewing the company's manager and the database administrator, the following information (Table II and Fig.1) about the social network service data are collected. For example, the friends of B are C, D and E but C, D, and E are not necessarily mutual friends (cf. C's friend list does not include E) in 31 March 2015.

Table II. Social Network Data

User ID	Time	Friends of Corresponding User
A	31 March 2015	E, F
	30 June 2015	B, E, F
	30 Sept. 2015	E, F
B	31 March 2015	C, D, E
	30 June 2015	A, C, E
	30 Sept. 2015	C, D
C	31 March 2015	B, D
	30 June 2015	B, D
	30 Sept. 2015	B, D
D	31 March 2015	B, C, E, F
	30 June 2015	C, E, F
	30 Sept. 2015	B, C, E
E	31 March 2015	A, B, D, F
	30 June 2015	A, B, D, F
	30 Sept. 2015	A, D, F
F	31 March 2015	A, D, E
	30 June 2015	A, D, E
	30 Sept. 2015	A, E

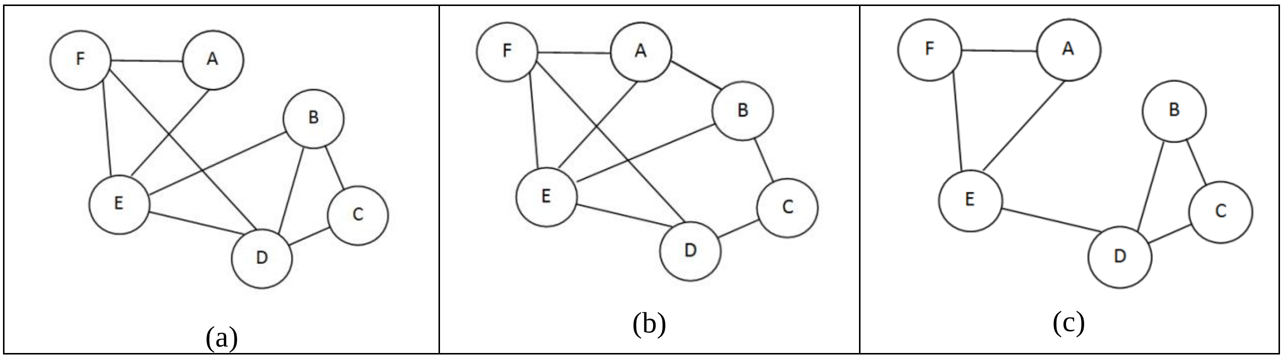

Fig.1 Social network graph of Table II (a) 31 March 2015, *(b)* 30 June 2015, *(c)* 30 Sept. 2015

a) Show the transformation step (step 3 of sequential pattern mining process) for user D and user E using *min_sup*=40%.

Ans.
Step 2 of sequential ARM: (min support=3 users (cf. 40%x6 users=2.4 users))
Frequent itemsets are A, B, D, E, F, (B F)
Step 3 (Transformation) of sequential ARM for user D:
<{(B), (E), (F), (B F)}, {(E), (F)}, {(B), (E)}>

Repeat it for user E:

Step 3 (Transformation) of sequential ARM for user E:
<{(A), (B), (D), (F), (B F)}, {(A), (B), (D), (F), (B F)}, {(A), (F), (D)}>

(20 marks)

b) How many possible sequences, of ANY length, can be extracted from user B?

Ans.

User B has 3 friends, 3 friends and 2 friends respectively in different time. Hence, the number of possible itemsets will be 7, 7, and 3 accordingly. Thus,
Number of length=3 sequences (3-sequences): 7x7x3
Number of length=2 sequences (2-sequences): 7x7+7x3+7x3
Number of length=1 sequences (1-sequences): 7+7+3
Total number of possible sequences=[147 + 91 + 17=255] – the number of repeated sequences

Some would like to think it this way:
1st transaction: 7 itemsets and nil itemset
2nd transaction: 7 itemsets and nil itemset
3rd transaction: 3 itemsets and nil itemset
Total number of possible sequences=[8×8×4-1=255] – the number of repeated sequences (e.g. <{C} {C}>, <{C} {D}>, etc.)

The answer above is good enough to obtain full mark.

One may further compute the repeated sequences and the final answer is 233 possible unique sequences.

Frequent Itemset Phase:

Freq. Itemset	Mapped to
A	1
C	2
D	3
E	4
CD	5
CE	6
DE	7
AC	8
AE	9
CDE	10
ACE	11

Transaction Phase:

User	Sequence	Transformed	Mapping
B	< (C, D, E), (A, C, E), (C, D)>	<{(C), (D), (E), (C D), (C E), (D E), (C D E)}, {(A), (C), (E), (A C), (A E), (C E), (A C E)}, {(C), (D), (C D)}>	< {2, 3, 4, 5, 6, 7, 10}, {1, 2, 4, 8, 9, 6, 11}, {2, 3, 5}>

For the Sequence Phase:

L1: 11 (7+7+3 = 17, but there are 6 recurrences)

L2: 7 * 9 + 4 * 3 = 75

(the 1st set owns 7 unique numbers, the rest two sets own 9 unique numbers; the 2nd set owns 4 unique numbers that not appear in the 1st set, multiple with the 3rd set)

L3: 7 * 7 * 3 = 147

Answer: 11 + 75 + 147 = 233.

(15 marks)

c) For *min_sup*=40%, list ANY TWO frequent sequences with length equal to 1. Repeat it for length equal to 2 and 3. Note that you are NOT required to show the mining steps and there may NOT have such frequent sequences or the required number of frequent sequences.

Ans.
For min_sup=40%, we need at least 3 users' support (cf. 40%x6 users)
Length=3: only 1 such sequence, i.e., <E E E> (supported by users A, D, & F)
Length=2: quite a lot, e.g. <E E>, <D D>, <B, F>, etc.
Length=1: any of A, B, D, E, F (no C here!)

(15 marks)