COMP4433 Data Mining & Data Warehousing Applications

Assignment 1 (Due: 23:59, 8 Oct 2021)

Instructions:
- Answer all questions.
- Interpret the questions logically, show your steps and write down your assumption(s) when necessary.
- Please submit your answer to L@PU before the due date.
- Late Submission Policy
  - 3-hour "grace period" is given.
  - 10% off for every 3-hour late
- Plagiarism Policy
  - Both giver and receiver subject to the same penalty below
  - All the students involved will receive 0 marks for this assessment. In addition, they will receive an additional 50% penalty, e.g., 5 marks for a 10-mark assessment.

1. 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	Buy

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 and ZM but sells NFLX".

a) Compute the support and confidence of the association rules:
   i) Buy MSFT $\Rightarrow$ Sell NFLX
   ii) Buy MSFT, Buy TSLA$\Rightarrow$ *
   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.

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

2. In a survey, the following three questions have been asked:
Q.1 Do you own a smart balance wheel?
Q.2 Do you have a driver license?
Q.3 Do you like selfie?

After computing the corresponding statistics for 10000 participants, the following data are given.
- Among 5000 participants who have a driver license,
  - 3250 own a smart balance wheel
  - 3750 like selfie
  - 2500 own a balance wheel and like selfie

- Among another 5000 participants who DON'T have a driver license,
  - 2750 own a smart balance wheel
  - 4000 like selfie
  - 2250 own a balance wheel and like selfie

a) Compute the frequent 3-itemsets using min-supp=10%. Based on the frequent 3-itemsets, list ALL strong association rules with 3 items using min-support=10% and min-confidence=50%.
b) List ALL 2-itemsets (not necessarily frequent) and compute their support values.

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 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

a) Based on the data in Table II, what is the largest itemset size (i.e. the maximum number of items in an itemset) found by sequential association rule mining algorithm? What is the longest sequence length (i.e. the maximum number of itemsets in a sequence) found by the sequence phase of sequential association rule mining? Note that the minimum support is unknown.

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

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

d) 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.