Försättsblad till skriftlig tentamen vid Linköpings universitet

(fylls i av ansvarig)

Datum för tentamen	2010-06-09
Sal	T1, TER2
Tid	14-18
Kurskod	TDDD41 / 732A02
Provkod	TEN1
Kursnamn/benämning	Data Mining - Clustering and Association Analysis
Institution	*IDA*
Antal uppgifter som ingår i tentamen	7
Antal sidor på tentamen (inkl. försättsbladet)	5
Jour/Kursansvarig	Patrick Lambrix / Jose M Pena
Telefon under skrivtid	2605
Besöker salen ca kl.	15 / 16.30
Kursadministratör (namn + tfnnr + mailadress)	
Tillåtna hjälpmedel	dictionary
Övrigt (exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)	
Vilken typ av papper ska användas, rutigt eller linjerat	
Antal exemplar i påsen	

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A02 and TDDD41
Data Mining –
Clustering and Association Analysis
June 9, 2010, kl 14-18

Teachers: Patrick Lambrix, José M Pena

Instructions:

- Start each question at a new page.
- Write at one side of a page.
- Write clearly.
- If you make assumptions about a question, that are not explicitly stated, you need to write these down. (These assumptions cannot change the exercise or question.)

Help: dictionary

GOOD LUCK!

1. Apriori algoritm (2p+1p+1p+2p=6p)

a. Run the Aprori algorithm on the following transaction database with minimum support equal to 1 transaction. Explain step by step the execution.

Transaction id	Items
1	A, B, C
2	G, H, I
3	I, H, G
4	C, B, A
5	A, G, B

b. In the exercise a, how many candidate itemsets did you produce ? And, for how many of these candidate itemsets did you count support ? Explain your answers.

c. Repeat the exercises a and b with the following constraint: An itemset cannot contain the item G unless that is the only item in the itemset. Explain step by step the execution. Identify what type of constraint we are dealing with. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.

d. Sketch a proof of the correctness of the Apriori algorithm.

2. FP algorithm (2p+2p+1p=5p)

a. Run the FP algorithm on the following transaction database with minimum support equal to 1 transaction. Explain step by step the execution.

Transaction id	Items
1	A, B, C
2	G, H, I
3	I, H, G
4	C, B, A
5	A, G, B
6	A, C

b. Repeat the exercise a with the following constraint: An itemset cannot contain the item G unless that is the only item in the itemset. Explain step by step the execution. Identify what type of constraint we are dealing with. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.

c. Sketch a proof of the correctness of the FP algorithm.

3. Constraints (1p+1p+1p=3p)

a. Given a transactional database, we want to find all the itemsets whose support is exactly 10. Specify the minimun support and constraints (if any) that are needed to run the Apriori or FP grow algorithm and find the desired itemsets.

b. Consider the following contraint: The most expensive item in an itemset cannot be more expensive than the two cheapest items in the itemset together. Is the constraint convertible monotone, convertible antimonotone, both or none ? Explain your answer.

c. Give an example of a constraint that is both convertible monotone and convertible antimonotone. Explain your answer.

4. Clustering by Partitioning (5p)

- Describe the principles and ideas regarding PAM.
  - Give a sketch of the algorithm.
  - Define swapping cost.
  - Draw an example where the swapping cost is 0 and one where the swapping cost is strictly negative.

- Why is PAM more robust than K-means in the presence of outliers?
- Why is CLARANS more efficient than PAM?

5. Hierarchical clustering (4p)

Describe the principles and ideas regarding CHAMELEON.
- Give a sketch of the algorithm.
- Explain k-nearest neighbor graph.
- Explain interconnectivity and closeness.

6. Clustering categorical data (4p)

Describe the principles and ideas regarding the ROCK algorithm.
- Give a sketch of the algorithm.
- Define neighbor, common neighbor, link, G (goodness measure)
- Give an example of neighbor, common neighbor and link involving objects {a, b, c, d}.

7. Distance measure (2+1=3p)

a. What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	(0,0)	(0,0)	Y	Y	Y	Y	Y
Item L	(3,4)	(1,1)	Y	N	Y	N	no-value-available

Attribute A is interval-based and Euclidean distance is used.
Attribute B is interval-based and Manhattan distance is used.
Attributes C and D are binary symmetric variables.
Attributes E and F are binary asymmetric variables.
Attribute G is a symmetric binary variable.

b. What is an ordinal variable? Give an example and define a distance measure.