Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	19/08/2009
Sal	TER 1, 8-12
Tid	
Kurskod	TDDD41/732A02
Provkod	TEN
Kursnamn/benämning	Data Mining - Clustering and Association Analysis
Institution	*IDA*
Antal uppgifter som ingår i tentamen	6 uppgifter 5 sidor
Antal sidor på tentamen (inkl. försättsbladet)	
Jour/Kursansvarig	Patrick Lambrix / Jose Pena
Telefon under skrivtid	2605
Besöker salen ca kl.	
Kursadministratör (namn + tfnnr + mailadress)	
Tillåtna hjälpmedel	lexikon
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
TDDD41/732A02 Data Mining – Clustering and Association Analysis
August 19, 2009, 8-12 am

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+2p+2p+1p=8p)

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

Transaction id	Items
1	A, B, C
2	A, B, D
3	X, Y, Z
4	X, Y, W
5	Z, W

b. What is the Apriori property and where did you use it in exercise above ?

c. Run the Apriori algorithm on the transaction database above to find all the itemsets with support equal or above one transaction and whose items do not all have the same support (i.e. the itemset RST is in the output if and only if its support is equal or above one and R, S and T do not have the same support). Tips: Rephrase the second requirement as a (anti)monotonic constraint. Explain step by step the execution. Make clear when the constraint is checked and what the consequences of the checking are, if any. Do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm.

d. Run the Apriori algorithm on the transaction database above to find all the itemsets with support equal or above one transaction and such that they are included in the itemset ABCXY or in the itemset DZW (e.g. the itemset ABD is not in the output). Tips: Rephrase the second requirement as a (anti)monotonic constraint. Explain step by step the execution. Make clear when the constraint is checked and what the consequences of the checking are, if any. Do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm.

e. Sketch a proof of the correctness of the Apriori algorithm. It suffices to sketch a proof showing that all the frequent itemsets of size k are among the candidates of size k.

2. FP algorithm (3p+1p=4p)

a. Run the FP algorithm on the transaction database in exercise 1 with minimum support equal to one transaction and the constraint that the sum of the prices of the items in an itemset must be strictly smaller than 10. Make clear when the constraint is checked and what the consequences of the checking are, if any. Do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm.

Item	Price
A	10
B	10
C	10
D	10
X	1
Y	1
Z	1
W	1

b. Sketch a proof of the correctness of the FP grow algorithm.

3. Constraints (1p+1p=2p)

a. Give two examples (with explanation) of a convertible monotone constraint that is not monotone.

b. Give two examples (with explanation) of a convertible antimonotone constraint that is not antimonotone.

4. Clustering by Partitioning (4p+1p=5p)

a. Describe the principles and ideas regarding CLARA. Explain the different steps of the algorithm. Explain in particular the notion of *cost* and give an example.
b. Compare CLARANS with PAM in terms of the graph representation of the clustering problem.

5. Hierarchical clustering (3p+3p=6p)

a. Describe the principles and ideas regarding the ROCK algorithm. For what kind of data is this algorithm particularly suited? Explain the major steps.Further, give an example with 4 objects that shows what a *neighbor* and a *common neighbor* are in ROCK and how it is used to define *Link*.

b. Describe the principles and ideas regarding the CHAMELEON algorithm.
What is the main purpose of the algorithm? For what kind of purpose would you use this algorithm? Explain the major steps. What are the strenghts and weaknesses of the algorithm?

6. Density and grid-based clustering (4p+1p=5p)

a. Describe the principles and ideas regarding the DBSCAN algorithm.
What is the main purpose of the algorithm? For what kind of purpose would you use this algorithm? Explain the major steps. What are the strenghts and weaknesses of the algorithm?

b. Consider the following statement: if p is density-connected to q wrt Eps and Minpts then p is density-reachable from q wrt Eps and Minpts. Is this statement true? If yes, then prove. If no, then give a counterexample.