Försättsblad till skriftlig tentamen vid Linköpings universitet

(fylls i av ansvarig)

Datum för tentamen	2009-03-12
Sal	TER2
Tid	14-18
Kurskod	732A02 och TDDD41
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 / 1651
Besöker salen ca kl.	15^{00}, 1645
Kursadministratör (namn + tfnnr + mailadress)	
Tillåtna hjälpmedel	Dictionary
Övrigt (exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)	Pan: 15 / 30

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A02 and TDDD41
Data Mining –
Clustering and Association Analysis
March 12, 2009, 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=4p)

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	A, B, D

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

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

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

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, D
2	A, C, D
3	B, C, D

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

3. Constraints (2p+2p+2p=6p)

a. Run the Apriori algorithm on the transaction database in exercise 1 with minimum support equal to 1 transaction and the constraint that the sum of the prices of the items in an itemset must be greater than 1 (do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm). Explain step by step the execution and when you check the constraint.

Item	Price
A	1
B	2
C	1
D	1

b. Run the Apriori algorithm on the transaction database in exercise 1 with minimum support equal to 1 transaction and the constraint that the sum of the prices of the items in an itemset must be equal or smaller than 2 (do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm). Explain step by step the execution.

c. Give an example of a convertible monotone constraint that is not monotone. Give an example of a convertible antimonotone constraint that is not antimonotone.

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

a. Describe the principles and ideas regarding PAM. Explain the different steps of the algorithm.

b. Given the graph representation of the clustering problem where a node represents k mediods (and thus a potential solution for the clustering), and nodes are neighbors if the sets of objects represented by the nodes differ by one object. Finding a solution for the clustering problem can then be seen as a search in the graph.
What is the difference between PAM and CLARANS regarding searching this graph?

5. Hierarchical clustering (4p)

Describe the principles and ideas regarding Agglomorative Hierarchical Clustering. Show the different steps of the algorithm using the dissimilarity matrix below and single link clustering. Give partial results after each step.

	1	2	3	4	5
1	0				
2	2	0			
3	4	3	0		
4	10	7	9	0	
5	8	5	6	1	0

6. Density-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.

7. Potpourri (1p+1p=2p)

a. Give an example with 4 objects that shows what a *neighbor* and a *common neighbor* is in ROCK and how it is used to define *Link*.

b. What is a CF-tree and how is it used in BIRCH?