Information page for written examinations at Linköping University

Examination date	2012-05-23
Room (1) If the exam is given in different rooms you have to attach an information paper for each room and mark intended place	TER1
Time	14-18
Course code	TDDD41
Exam code	TEN1
Course name **Exam name**	Data Mining - Clustering and Association Analysis Skriftlig tentamen
Department	IDA
Number of questions in the examination	8
Teacher responsible/contact person during the exam time	Patrick Lambrix / Jose Pena
Contact number during the exam time	2605
Visit to the examination room approx.	15.15, 16.30
Name and contact details to the course administrator (name + phone nr + mail)	Carita Lilja, 1463, carita.lilja@liu.se
Equipment permitted	dictionary
Other important information	For pass you need half of the max points.
Which type of paper should be used, cross-ruled or lined	
Number of exams in the bag	

Institutionen för datavetenskap
Linköpings universitet

EXAM
TDDD41 Data Mining – Clustering and Association Analysis
May 23, 2012, 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+1p+1p=6p)

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

Transaction id	Items
1	A, B, C, D
2	F, E, A
3	A, E, G

b. Repeat the exercise 1a with the following additional constraint: Find the frequent itemsets that contain two letters that are consequtive in the alphabet (i.e. A and B are consequtive, B and C too, C and D too, D and E too, E and F too, and F and G too). Explain step by step the execution. Make clear when and how the constraint is used. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.

c. Repeat the exercise 1a with the following additional constraint: Find the frequent itemsets that do not contain any pair of letters that are consequtive in the alphabet. Explain step by step the execution. Make clear when and how the constraint is used. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.

d. What role does the Apriori algorithm play in the search for association rules ?

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

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

a. Repeat the exercise 1a with the FP grow algorithm. Explain step by step the execution.

b. Explain how and where we incorporate a monotonic constraint into the FP grow algorithm.

c. Explain how and where we incorporate an antimonotonic constraint into the FP grow algorithm.

d. What is the main advantage that the FP grow algorithm has over the Apriori algorithm ?

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

a. Give three examples of constraints that are monotone. Explain your answer.
b. Give three examples of constraints that are antimonotone. Explain your answer.
c. Give three examples of constraints that are neither monotone nor antimonotone but that are convertible monotone and convertible antimonotone (i.e. strongly convertible). Explain your answer.

4. Clustering by Partitioning (1p+2p=3p)

a. Describe the principles and ideas regarding CLARANS.
   (i) Give a sketch of the algorithm.
   (ii) Define swapping cost.

b. Compare CLARANS with PAM.

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

6. Clustering categorical data (1p+4p=5p)

a) Why do approaches like PAM not work well for categorical data? Give an example.
b) Describe the principles and ideas regarding the ROCK algorithm.
   Within your description, make sure to give a sketch of the algorithm and to define neighbor, common neighbor, link for objects, link for clusters, and **G** (goodness measure).

7. Density-based clustering (2p)

Describe the principles and ideas regarding the DBSCAN algorithm.
In your description, make sure to give a sketch of the algorithm and to define core point, direct density-reachable, density-reachable, and density-connected.

8. Data mining concepts (1p+1p=2p)

a. The purpose of data mining is to extract interesting patterns from a huge amount of data. What does 'interesting' mean in this case?
b. Data in the real world can be dirty. Give 3 reasons and an example for each.