Databases and Data Mining 2014
Final Exam

LIACS Room 174
Friday December 19th 2014
10.00 – 13.00

- State your name and student number on every page of your answers.
- Every assignment has the same weight. There are 10 assignments.
- Always fully explain your answers.
- Please note that you have a total of 3 hours to answer the questions.
- It is an open book exam: you are allowed to use your book and course notes (slides).
- All electronic equipment should be off the table and switched off.

1. Let a graph *G = (V, E)* be defined by *V* the set of vertices being equal to the set of currently living people, and *E* the set of edges between vertices of *V*, where an edge *(v, w)* is in *E*, if and only if the person represented by vertex *v* '*knows'* the person represented by vertex *w*.
    a. Firstly define 'knows' and then give 3 important characteristics of graph *G*.
    b. Describe an algorithm that produces a synthetic graph *G'* with the same characteristics as *G*.
    c. Give 2 other examples of natural occurring networks with the characteristics you gave in a).
    d. If the degree of separation for G in 1980 is equal to d_{1980} and in 2014 equal to d_{2014}, how do you think they relate to each other? Explain why.

2. Give a succinct constraint. Describe how this succinct constraint can be exploited in the FP-Growth Algorithm.

3. The Pattern-Fusion Algorithm is an efficient solution for colossal pattern mining.
    a. Is it true that standard mining methods like Apriori or FPGrowth can also be used for mining colossal patterns in arbitrary data sets? Explain.
    b. Which characteristic of a colossal pattern is exploited by the Pattern-Fusion Algorithm?
    c. Will the Pattern-Fusion Algorithm find all frequent colossal patterns of a given data set?

4. A database has five transactions. Let min_sup = 50%, and min_conf = 80%.

TID	*Items_bought*
T100	{ A, C ,D, E, G }
T200	{ A, D ,E ,F }
T300	{ B, C, F, G }
T400	{ B, C, F, G, H }
T500	{ A, C, D, F }

a. Find all frequent item-sets using the Apriori-Algorithm (give the intermediate results for the different steps of the algorithm).
b. What is the dimension and size of the data set?
c. Assume that N and D are equal to the size and the dimension of the dataset, respectively. Give an expression for the time- and space-complexity of the Apriori-algorithm.

5. Assume a database *DB* of transactions is given with items/events {A, B, C, …}. After data mining the database, the pattern *A* => *B* is found. (Note: *A* can be something like *"has a dog",* and *B "walks more than 5 km a day"*.)

   a. Give two examples of null-invariant interestingness measures for the correlation/dependence of the two items/events *A* and *B*.
   b. What values of your measures would you expect, if the items/events *A* and *B* are negatively correlated/dependent in *DB*?
   c. What values of your measures would you expect, if the item sets in *DB* are imbalanced and positively correlated with respect to items/events A and B?

6. An optimization in frequent item set mining is mining max patterns.

   a. Describe why mining max patterns can be done more efficiently than mining frequent item sets.
   b. Would the Apriori-Algorithm be useful for mining max patterns in a 100-dimensional data set of more than 10^6 elements?
   c. Given the mined max patterns P of a given data set D, can you determine all frequent item sets in D from the result P?

7. Assume a database DB with T tuples and dimension D is given.
   a. What does the curse of dimensionality mean for data cubing?
   b. Give 2 data cubing methods that could be used to face the curse of dimensionality. Also explain how the curse of dimensionality is countered in each of the methods, respectively.

8. Suppose that a data warehouse *DW* for a Dutch *Company* consists of five dimensions, *location*, *supplier*, *time*, *product_material*, and *product*, and three measures, *count*, *average_costs*, and *total_costs*.
   a) Draw a *snowflake schema* diagram for the data warehouse.
   b) Starting with the base cuboid [*location*, *supplier*, *time*, *product_material*, *product*], what specific OLAP operations should one perform in order to list the total count of a certain product for each *supplier* per *month* (assume *time* has three levels: *day, month, year*)?
   c) Assume we already summarized the measures over all *product_material*, and *time* and we want to compute the remaining 3-dimensional data cube. If there are 1000 *locations*, 200 *suppliers*, and 220 *products*, in what order would you traverse the cube cells when you use multi-way array aggregation for the cube computation?
   d) If each dimension has 4 levels (including *all*), how many cuboids will the cube contain (including the base and apex cuboids)?

9. Give an example where stream-processing and -mining is necessary. Why does a solution implementing stream-processing and -mining using an RDBMS only, in general will not give you the required performance? Sketch a solution that would give an efficient implementation for stream-processing and -mining.

10. Describe the main differences between a Markov Model and a Hidden Markov Model. Assume that a Markov model *M* is given that models a certain class of strings *S* over an alphabet {A, B, C, D}. Assume that also a Hidden Markov Model *HM* is given that models *S*. How would you determine the probabilities *P(S|M)* and *P(S|HM*), respectively?