NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Computer and Information Science

Examination paper for TDT4300 Data warehousing and data mining

Academic contact during examination: Kjetil Nørvåg

Phone: 73596755

Examination date: June 1st 2017

Examination time (from-to): 09.00-13.00

Permitted examination support material: D: No tools allowed except approved simple calculator.

Other information:

Language: English

Number of pages (front page excluded): 3

Number of pages enclosed: 0

Checked by:

_____

Date Signature

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig X **2-sidig** ☐

sort/hvit X **farger** ☐

Problem 1 – Various – 10 % (all having same weight)

a) Explain *pageview* in context of web usage mining.

b) One type of pre-processing in web usage mining is *path completion*. *Why* is this necessary, and *how* can it be done?

Problem 2 – Modeling – 15 %

Miljøbomringen AS will soon be responsible for the tollbooths in all major cities in Norway, and want a data warehouse that can be used to analyze traffic, i.e. the toll passages. As part of this reorganization, all cars must have AutoPass (transponder) for automatic registration of passages. A customer may have several cars and must have one transponder for each car. The price for each passage changes dynamically/continuously for each station independently of others, based on time of day, pollution, traffic jams, etc.

An example of analyzes you should be able to do against the data warehouse:

- Number of passages for each quarter for each station.
- Number of passages for each quarter for each car.
- Average number of passages per month.
- Average price for cars for one particular station.

The description is somewhat imprecisely formulated and it is part of the task to select what should be included. We are primarily looking for you to show modeling principles for data warehousing. Explain any assumptions you find necessary to do.

Create a star schema for the described case.

Problem 3 – OLAP – 15 % (5 % on a and 10 % on b)

a) Explain *roll-up* and *drill-down.*

b) Given a dimension table *Book* in a data warehouse, we want to use *bitmap indexes* on the attributes Language and Binding in order to be able to perform queries more efficiently. Show structure and contents of the bitmap indexes based on the contents in the table below.

Book				
RowID	**BookID**	**Title**	**Language**	**Binding**
1	45	The Hobbit	English	Hardcover
2	63	À la recherche du temps perdu	French	Hardcover
3	88	For Whom the Bell Tolls	English	Paperback
4	143	Madame Bovary	French	Paperback
5	236	La Peste	French	Hardcover
6	463	The Grapes of Wrath	English	Hardcover
7	768	The Great Gatsby	English	Paperback

Problem 4 – Clustering – 10 %

Assume a two-dimensional dataset as shown in the table to the right. Perform clustering using K-means, with k=3 and intial centroids S1=(4,4), S2=(5,8) og S3=(5,11). Use Manhattan distance.

	X	Y
P1	4	8
P2	4	10
P3	4	13
P4	5	3
P5	5	7
P6	7	11

Problem 5 – Classification – 25 % (10 % on a and 15 % on b)

a) You are given a dataset of samples P1 = (4,8), P2 = (8,8), P3 = (8,4), P4 = (6,7), P5 = (1,10), P6 = (3,6), P7 = (2,4), P8 = (1,7), P9 = (6,4), P10 = (6,2), P11 = (6,3), P12 = (4,3), and P13=(4,4). The samples belong to three clusters C1 = {P1,P2,P3,P4}, C2 = {P5,P6,P7,P8} and C3 ={P9,P10,P11,P12,P13}. Consider associated clusters as class labels. Classify the samples A = (6,6), B = (4,6), C = (4,5), and D=(2,6) by employing the k-nearest neighbor (k-NN) method. Use the Manhattan distance metric and k = 3. Describe how the results of the classification are achieved.

b) As part of a larger application we want to be able to predict class (*J* or *N*) based on input data where each record contains a sequence number and the attributes A, B, C, and D:

Nr	**A**	**B**	**C**	**D**	***Class***
1	L	F	R	2	*J*
2	H	T	S	4	*J*
3	H	T	S	4	*J*
4	L	F	S	2	*N*
5	H	F	G	5	*N*
6	H	T	G	2	*N*
7	L	F	S	6	*N*
8	H	K	G	4	*N*
9	H	T	H	2	*J*
10	H	F	S	5	*N*
11	H	K	B	7	*N*
12	L	F	B	9	*N*
13	L	K	R	2	*N*
14	L	F	H	1	*N*
15	L	F	H	7	*N*

Assume that we will use *decision tree* as the classification method. We will use the above dataset as our training data. We use the *Gini index* as measure for impurity, and the following two equations might be of help for solving the problem:

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

$$GAIN_{split} = GINI(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} GINI(i) \right)$$

Task: The goal of the classification is to be able to predict "Class". Compute the $GAIN_{split}$ for splitting by attribute (1) "*A*" and (2) "*B*". Which of these splits would you chose to start building your decision tree? Justify your answer.

Problem 6 – Association rules – 25 % (10 % on a and 15 % on b)

a) Assume the market basket data below. Use the apriori-algorithm to find all frequent itemsets with minimum support of 50 % (i.e., *minimum support count* is 4). Use the $F_{k-1} \times F_{k-1}$ method for candidate generation.

TransactionID	**Item**
T1	ABCDEG
T2	CDFH
T3	AFG
T4	DF
T5	BDEG
T6	BDEG
T7	BCDEGH
T8	ACF

b) Assume the market basket data below. You are now going to use the *FP-growth-algorithm* in order to find all frequent itemsets with minimum support of 40 % (i.e., *minimum support count* is 2).
1) Construct a FP tree based on the dataset.
2) Find frequent itemsets using the FP-growth-algorithm. Use table notation with the following columns in order to show the result:

- Item
- ”Conditional pattern base”
- ”Conditional FP-tree”
- Frequent itemsets

TransactionID	**Item**
T1	ACE
T2	BCE
T3	BCDE
T4	CDE
T5	DE