UNIVERSITY OF TWENTE.

Exam

Module: Business Intelligence and Information Technology
Course: Databases and Business Intelligence
Date: 11 March 2016
Normal Exam time: 13:45 – 15:30

Lecturers: dr. C. Amrit
University of Twente

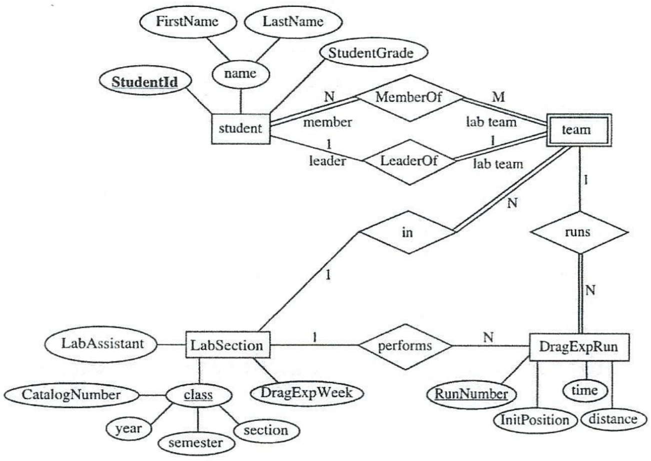
Name:____

Student nr:____

Please **return** the question paper after the exam!

Closed Book Exam
This is a closed book exam: No course materials (slides handouts, books, and papers) can be used during the exam.

Grading
This exam contains 25 multiple-choice questions (from the lectures, book chapters and from the papers that have been included in the course) and 1 optional question (for 2 marks). All the questions have only **1 right answer.**
Each of the following questions is for 1 mark.



1. If the above ER Diagram (where the cardinalities are mentioned in the relationships, and triangle also represents a relationship) were to be converted to a relational schema, the schema would contain

A) 4 tables
B) 5 tables
C) 6 tables
D) 7 tables
E) 8 tables

2. In the same ER Diagram above the number of tables in the schema formed from weak entities and many-many relationships are:

A) 1 table
B) 2 tables
C) 3 tables
D) 4 tables
E) 5 tables

Consider the following schema *Articles*:

ID	Title	Journal	Issue	Year	Pages	TR-ID
42	Cuckoo Hashing Deterministic	JAlg	51	2004	121	T1
43	Social "science" really?	JAlg	41	2001	10	T2
44	Why Databases are cool!	SICOMP	41	2001	131	T1
45	P vs NP resolved	JACM	43	1998	150	T3
46	What G ̈odel missed	SICOMP	44	1978	166	T4
47	Why Databases are cool!	SICOMP	41	2001	131	T5
45	P vs NP resolved	JACM	43	1998	150	T6

It contains information on articles published in scientific journals. Each article has a unique ID, a title, and information on where to find it (name of journal, what issue, and on which pages). Also, if results of an article previously appeared in a "technical report" (TR), the ID of this technical report can be specified.

3. Based on the above, which of the following potential functional dependencies, is NOT a functional dependency.

A. ID -> Title
B. Issue -> Year
C. Title->Journal
D. ID -> Journal, Issue
E. ID -> TR-ID

4. The query
SELECT COUNT(a.ID) FROM Articles a WHERE a.Year > 1997 GROUP BY a.Issue;

will return

A. 2, 1
B. 3, 2, 1
C. 2
D. 3
E. 4

5. Which of the following queries, would return less than 3 tuples if it was run on the instance of Articles

A. SELECT ID FROM Articles WHERE Year<2001;
B. SELECT DISTINCT ID FROM Articles WHERE Year<2001;
C. SELECT AVG(Year) FROM Articles GROUP BY Journal;
D. SELECT ID FROM Articles WHERE Pages>100;
E. SELECT ID FROM Articles WHERE Pages>100 GROUP BY Journal ;

Additionally, consider the relations Authors(auID,name) and Authoring(articleID, authorID), containing information on names of authors, and who is authoring which papers, respectively.

6. The query:

```
 SELECT ID, title, COUNT(*)
FROM Articles, Authoring
WHERE ID=articleID
GROUP BY ID,title;
```

Returns:

A) The set of all titles
B) ID, title and the number of authors per book
C) ID, title and count of all books
D) ID, title and count of journals
E) None of the above

7. The query:

```
 SELECT COUNT(DISTINCT A2.authorID)

FROM Authors, Authoring A1, Authoring A2

WHERE A1.authorID=auID AND name='Robert Tarjan' AND A2.authorID<>auID

AND A1.articleID=A2.articleID;
```

Returns:

A) The set of all titles
B) The number of books authored by Robert Tarjan
C) The number of distinct co-authors of Robert Tarjan

D) The number of distinct authors who have not co-authored with Robert Tarjan
E) The number of distinct authors who have written a book titled 'Robert Tarjan'

Consider the following Employee Department Project DB schema (the primary keys are underlined):

<u>EmpID</u>	Department	<u>ProjID</u>	ProjName	ProjDescription	ProjYear
123	IEBIS	P1	PName1	Social Media	2012
123	IEBIS	P2	PName2	Social Network	2013
324	BA	P1	PName1	Database	2014

<u>EmpID</u>	Name	Address
123	ABC	Abcdstraat, Enschede
324	BCD	Asasdastraat, Enschede

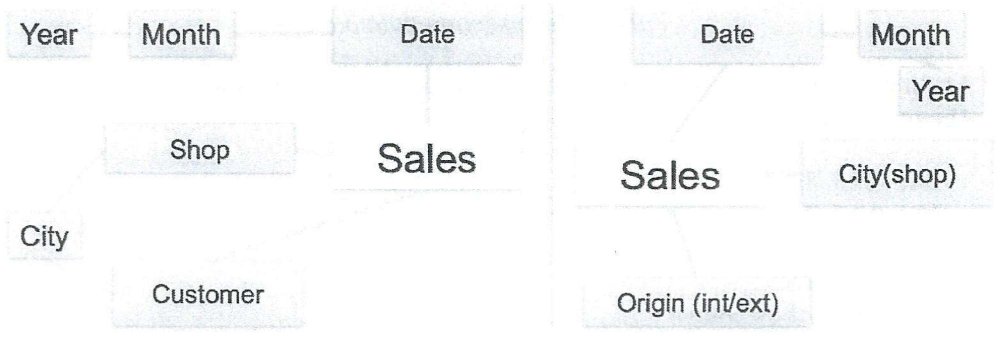
8. The above schema is in/only in the

A. First Normal Form
B. Second Normal Form
C. Third Normal Form
D. Boyce-Codd Normal Form
E. Fourth Normal Form

9. In order to convert the above schema into the next higher normal form, you would:

A. Make each attribute atomic
B. Have no multivalued columns
C. Make each attribute functionally dependent on the primary key
D. Remove transitive dependencies
E. Eliminate trivial multivalued dependencies

10. Director of chain of high-end audio/video shops wants to know per month and city how many sales come from customers in the same city as the shop vs. sales from customers coming from other cities. He needs this to decide if he needs to open shops in all major cities or that customers are willing to travel to go to his shops.



(a) (b)

In the figure above comparing (a) Data Warehouse schema with (b) we can conclude

A. Both (a) and (b) are good Data Warehouse schemas
B. Schema (b) is better than (a) as only the Origin of the sales are important and not the shop's location
C. Customers generally travel and hence the customer city information is very important and hence Schema (a) is better than (b)
D. For the business problem one requires only customer origin information and schema (a) is not an optimum snowflake schema
E. Schema (b) is not a valid star schema so (a) is better than (b)

A supermarket stores all the transactions in a large database. These transactions database can be used for "basket analysis". For the sake of simplicity and time we focus only on the following small part of the database and the items:

A supermarket stores all the transactions in a large database. These transactions data base can be used for "basket analysis". For the sake of simplicity and time we focus o nly on the following small part of the database and the items:

TID	Bread	Milk	Beer	Eggs	Cola
1	1	1	1	1	1
2	0	1	0	1	1
3	1	0	1	1	1
4	0	1	1	0	0
5	1	1	0	1	0
6	1	1	1	0	0
7	1	0	1	0	1
8	1	1	1	1	1
9	1	0	0	0	1
10	1	1	1	1	0
11	1	1	1	1	1
12	1	1	1	0	1
13	1	1	1	0	1
14	1	0	0	0	1
15	0	1	1	1	0

Part of a transaction data base

11. In the above table, the support and confidence of the following association rules: {bread} => {milk}

A. 8/15, 8/11
B. 9/12, 9/15
C. 8/15, 8/12
D. 8/12, 8/15
E. None of the above

12. The support and confidence for {milk} => {bread} is:

A. 8/15, 8/12
B. 9/11, 9/15
C. 8/15, 8/11
D. 8/11, 5/15
E. None of the above

13. The support and confidence of {beer}=>{milk, bread} is:
A. 7/15,7/12
B. 8/11,8/15
C. 7/15,7/11
D. 7/11,7/15
E. None of the above

14. The frequent 1 item sets with support ≥ 50% from Table 1 are:
A. {Bread}, {Beer}, {Eggs},{Milk}, {Cola}
B. {Bread}, {Beer}, {Eggs},{Milk},
C. {Beer}, {Eggs},{Milk}, {Cola}
D. {Bread}, {Eggs},{Milk}, {Cola}
E. {Bread}, {Eggs},{Milk},

15. The frequent 2 item sets with support ≥ 50% from Table 1 are:

A. {bread, milk}, {bread, beer}, {bread, cola},{milk, beer}
B. {bread, milk}, {bread, beer}, {bread, cola}
C. {bread, beer}, {bread, cola},{milk, beer} , {milk, cola}

D. {bread, milk}, {bread, cola},{milk, beer}

E. {bread, milk}, {bread, cola}

16. The frequent 3 item sets with support ≥ 50% from Table 1 are:
A. {bread, milk, cola}, {bread, cola, beer}
B. {bread, eggs, beer}, {bread, cola, milk}
C. {bread, cola, milk}
D. {bread, milk, beer}
E. None of the above

17. Assume that the confidence of the decision rule, *a => b,* is *80%*, then the confidence of the decision rule, *b=>a* is always:

A) 80%.
B) <80%.
C) >=80%
D) <=80%
E) None of the above

18. Which of the following are examples of unsupervised learning algorithms.

A) Support vector machine
B) Naïve Bayes
C) Decision tree
D) Association rule learning
E) Random Forest

19. When comparing traditional databases with OLAP, which of the following are true:

A) In OLAP systems the speed with which queries are executed is much higher than in traditional relational data base
B) In OLAP systems the speed with which queries are executed is much slower than in traditional relational data base
C) OLAP systems contain more data than the individual data bases put together
D) OLAP systems contain more data than the individual data marts put together
E) None of the above

20. Financial profit, number of defects, and employee work time are examples of physical measures whose data are represented in

A) ordinal
B) nominal
C) interval
D) ratio
E) numerical

21. Which of the following is not true about classifiers

A) The simple split methodology splits the data into a training and a test set which may have some overlap
B) The simple split methodology splits the data into a training and a test set which has no overlap
C) Sometimes the data can be split into a training, validation and a testing set
D) The training and the test set is generally split in the ratio 2:1.
E) The training and the test set is generally split in the ratio 1:2

22. In ________, the problem is to group an unlabeled collection of objects, such as documents, customer comments, and Web pages into meaningful groups without any

prior knowledge.

A) search recall
B) classification
C) clustering
D) grouping
E) neural networking

23. Which of the following issues related to BI implementation is false?

A) BI and predictive analytics can lead to serious ethical issues such as privacy and accountability.
B) Developing an effective BI application is no longer complex.
C) Smaller organizations can make the solutions cost effective if they leverage existing databases rather than create new ones.
D) The quality and timeliness of business information for an organization is not the choice between profit and loss--it may be a question of survival.
E) All the above

24. Which of the following is a reason why BI projects fail?

A) Failure to recognize BI projects as enterprise-wide business initiatives and that they differ from typical stand-alone solutions.
B) Lack of business sponsorship and the inability to insure funding.
C) Lack of qualified and available staff.
D) All of the above.
E) None of the above

25. Characteristics of all well-designed dashboards and scorecards include all of the following except:

A) they present a static, real-world view of the data
B) they use visual components to highlight, at a glance, the data and exceptions that require action
C) they are transparent to the user and extremely easy to use
D) they enable drill-down or drill-through to underlying data sources or reports
E) they are based on issues relevant to the business problem at hand

BONUS Question for 2 points,

Please note: There will be no points given for answers without clear Gini index calculation – if you are doing this submit the extra sheets along with the Multiple Choice answer sheets

26. ***Data mining exercise***.

A retailer wants for marketing purposes distinguish between costumers younger than 35 Y and customers older then 35, O. The following table summarizes the data set in the data base of the retailer in an abstract form. The relevant attributes, determined by domain knowledge, are for convenience denoted by A and B. The values for A are A1, A2 and A3. The values for B are B1 and B2.

A	B	Number of Instances	
		Y	O
A1	B1	14	0
A2	B1	0	4
A3	B1	6	2
A1	B2	0	12
A2	B2	6	4
A3	B2	0	6

The retailer wants to use Data Mining techniques to classify the costumers in the class "young", denoted by Y, and "old", denoted by O.

The GINI Index for a given node t in the decision tree is:

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

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Where $p(\frac{j}{t})$ **is the relative frequency of class j at node t.**

n_i **= number of instances at child i,**

n **= number of instances at node t.**

Based on the Gini Index, the best attribute for the root node of the decision tree is:

A. Y
B. O
C. A
D. B
E. A1