COMP 60711

Two hours

EXAM PAPER MUST NOT BE REMOVED FROM
THE EXAM ROOM AND MUST BE RETURNED

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Data Engineering

Date: Wednesday 25th January 2017

Time: 09:45 - 11:45

Please answer ONE Question from Section A and ONE Question from Section B

Use a SEPARATE answer book for each SECTION

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]

Section A

1.

a) Make an analogy between Data Engineering and Software Engineering, illustrating your answer with an example of a problem that Software Engineering addresses, and an example of a problem that Data Engineering addresses.

(6 marks)

b) Describe the Data Life Cycle, giving, for each phase, one example of an issue that has to be addressed during that phase.

(8 marks)

c) Considering the Data Collection phase of the Data Life Cycle and your experience with data quality problems from your coursework, create a data quality methodology that allows organizations to address data quality issues at this phase of the Data Life Cycle. Assume that your methodology will encompass only the following three main steps: (1) State reconstruction, (2) Assessment/Measurement and (3) Improvement.

(6 marks)

d) Explain the steps involved in the process of cleaning data and how integrating big data from multiple and heterogeneous sources can make this process more challenging.

(5 marks)

e) Define data profiling, and enumerate and discuss the challenges involved in profiling very large data sets.

(5 marks)

2.

a) What is the definition of high value data? Provide an example of data of high value that does not degrade over time, and another example of data of high value that degrades over time.

(4 marks)

b) Explain how the ETL (Extract, Transform and Load) process applied in data warehousing relates to each Data Life Cycle phase.

(6 marks)

c) Consider a for-profit organization with Web presence that has decided to record user activity for analysis of user behaviour, but cannot cope with the volume of data to be analysed. Using the best practices in Data Life Cycle Management relating to both identification of objectives and principle of minimalism, develop a detailed Data Life Cycle management process specifically designed to manage the life cycle of this data for this organization.

(7 marks)

d) Consider the following schemas, independently developed for two databases that record information about hotels, room bookings, and guests. In all parts of your answer, state any assumptions you make.

Suppose that a number of hotel chains, belonging to the same owner, store hotel information using the below database schemas. Therefore, the database schemas need to be integrated. During the integration process, schema conflicts have to be reconciled. Answer the questions below, using the following notation: Schema_i.TableName.AttributeName to refer to an attribute in Schema_i; Schema_i.TableName to refer to a table in Schema_i.

Schema_1:

```
Hotel(hotelNo, hotelName, address)
Room(roomNo, hotelNo, type, price, features)
Booking(hotelNo, guestNo, dateFrom, dateTo, roomNo)
HolidayGuest(guestNo, guestName, guestAddress, packageNo)
BusinessGuest(guestNo, guestName, guestAddress)
Package(packageNo, packageName, Company, discount)
```

Schema_2:

```
Hotel(hotelNo, hotelName, classification, streetNo, streetName,
postcode, city, country)
HotelFacilities(hotelNo, facilityNo, numOfUnits)
Facility(facilityNo, facilityName, description)
Bedroom(roomNo, hotelNo, type, additionals, price)
Reservation(roomNo, hotelNo, guestNo, dateFrom, dateTo,
discountGuestType)
Guest(guestNo, guestName, guestAddress, type)
```

i. Identify and describe two different one-to-one table name conflicts.

(1 mark)

ii. Identify and describe two different one-to-one table structure conflicts, one of them being the case of a missing, but implicit attribute.

(2 marks)

iii. Identify and describe one table inclusion conflict.

(1 mark)

iv. Identify and describe two different one-to-one attribute vs. attribute conflicts.

(2 marks)

v. Identify and describe one many-to-many attribute conflict.

(1 mark)

vi. Produce an SQL view (`HotelBill`) that derives a table with the structure shown below from both Schema_1 and Schema_2 databases, where `durationOfStay` can be obtained by subtracting attribute `dateFrom` from attribute `dateTo`, `discount` can be null for some guests, and `totalToPay` can be obtained using `durationOfStay`, `roomPrice` and `discount`.

```
HotelBill(guestNo, roomPrice, durationOfStay,
discount, totalToPay)
```

(6 marks)

Section B

3.

a) In the context of business intelligence:

(i) Characterise the differences between OLTP and OLAP? Why might different database systems be used for each type?

(2 marks)

(ii) Compare and contrast the use of row store and column store.

(3 marks)

(iii) Discuss a business intelligence application that may benefit from data mining solutions instead of OLAP queries. Be specific with the data mining method you recommend and justify your answer.

(3 marks)

b) In the context of classification:

i) Outline a decision tree classification algorithm.

(4 marks)

ii) Explain the potential effects of unbalanced data on the usefulness of a classifier. Suggest the advantages of various methods to handle unbalanced data. Give examples to support your answer.

(4 marks)

c) In the context of association rule (itemset) mining:

(i) The following contingency matrix shows a breakdown of transactions for coffee and tea drinkers (assume 1000 transactions).

	coffee	not coffee	Total
tea	150	50	200
not tea	650	150	800
Total	800	200	1000

Calculate support, confidence and lift for the association rule *{tea} -> {coffee}*.

(2 marks)

(ii) Explain why lift better represents the relationship between tea and coffee drinkers than using support and confidence.

(2 marks)

(iii) Consider ways in which association rule (itemset) mining might be extended. Discuss in your answer adding multi-dimensionality, quantitative intervals and hierarchies to the standard approach.

(4 marks)

d) In the context of "big data":

(i) Describe the major steps in the "Big Data Analysis Pipeline".

(3 marks)

(ii) Discuss the challenges associated with the "Big Data Analysis Pipeline".

(3 marks)

4.

a) In the context of association rule (itemset) mining:

(i) Discuss the terms correlation and causation.

(2 marks)

(i) Outline the working of the Apriori algorithm. Explain the importance of the subset property.

(5 marks)

(ii) Using Apriori, suppose that L3 is the list:

```
{     {a,b,c}, {a,b,d}, {a,c,d}, {b,c,d}, {b,c,w}, {b,c,x},
      {p.q.r}, {p,q,s}, {p,q,t}, {p,r,s}, {q,r,s} }.
```

a. At the join step of the algorithm, which itemsets are placed in C4 (the candidate set)?

(3 marks)

b. Which itemsets are discarded by the prune step of the algorithm?

(3 marks)

Provide your working as appropriate.

b) In the context of classification:

(i) Two classifiers designed to predict patients' susceptibility to allergy are being designed and tested, independently of one another. Each of the classifiers predict that a patient is either positive (allergic) or negative (normal) based on a combination of observable factors. The tests result in the following two confusion matrices, one for each classifier:

Table A:

		predicted		
		allergic	normal	Total
actual	allergic	30	70	100
	normal	20	500	520
	Total	50	570	620

Table B:

		predicted		
		allergic	normal	Total
actual	allergic	70	30	100
	normal	200	320	520
	Total	270	350	620

[PTO]

Calculate the accuracy, recall, and F-measure for each classifier based on these tables. Based on their values, can you recommend one classifier over the other, given this type of application? Justify your answer.
(4 marks)

ii) Compare and contrast information gain, gain ratio and the GINI index.
(3 marks)

(iii) Some classification algorithms run out of memory in trying to fit all data in memory to create the classification model. Discuss ways in which you might address the issue of memory capacity?
(4 marks)

c) Discuss the trade-offs between privacy of the individual on the one hand and the utility of data on the other. Use well-known case studies of data disclosure as appropriate to illustrate these trade-offs. How do we better guarantee that data is private?
(6 marks)

END OF EXAMINATION

The Question Paper must be returned before you leave the examination