COMP60711

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Data Engineering

Date: Thursday 21st January 2016

Time: 14:00 - 16:00

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) Define data profiling, enumerate and discuss the challenges involved in profiling very large data sets. (8 marks)

   b) It is well known that, for companies that maintain a commercial Website, knowing what users look at, what they click on, how much time they spend on a web page, etc. leads to better business decisions and competitive advantage, since this ***user behaviour data*** can be analysed and can provide insights from which valuable opportunities can be derived. In this context, consider a company, (e.g., Amazon, eBay, etc.), and provide a detailed discussion of which data lifecycle management procedures this company should implement to avoid drowning in a torrent of log data, since it is a fact that logging user behaviour generates so much data that many organizations simply cannot cope with the volume, and either turn the functionality off or throw away data after some time. In your discussion, provide details of the company's business and any IT infra-structure solution or programming paradigm that you may find useful to help the company deal with this problem. (6 marks)

   c) What issues can possibly arise during the (big) data lifecycle phase of data acquisition, considering sensor-collected and the fact that sensors are able to produce large volumes of data? What solutions can you suggest to address these issues? (8 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. (8 marks)

2. a) Suggest how general data lifecycle management principles could be applied in the context of the social networking website, Facebook, as a way of managing the torrent of data Facebook needs to deal with on a daily basis. State any assumptions you make. (5 marks)

b) Explain how averages of numerical columns in relational data can be calculated in a distributed manner (i.e., the data is distributed across a number of nodes of a cluster, and the overall average of one of the numerical columns of the entire data set is to be calculated). Then, explain how the median of the values in the same column can be calculated in a distributed manner. (8 marks)

c) Explain the impact that the evolution of information systems into network-based structures has had on the quality of the data that runs through the processes of these systems, and discuss how data quality can be assessed and improved. (4 marks)

d) Consider the following schemas, independently developed for two databases that record information about hotels, room bookings, and guests. In all parts of the question, 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 any of the two schemas; Schema_i.TableName, to refer to a table in any of the two schemas.

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)
```

(Question 2 continues on the following page)

(Question 2 continues from the previous page)

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. Describe two different one-to-one table name conflicts. (1 mark)

ii. Describe two different one-to-one table structure conflicts, one of them being the case of a missing, but implicit attribute. (2 marks)

iii. Describe one table inclusion conflict. (1 mark)

iv. Describe two different one-to-one attribute vs. attribute conflicts. (2 marks)

v. Describe one many-to-many attribute conflict. (1 mark)

vi. Produce a SQL view that derives a table with the following structure from both Schema_1 and Schema_2 databases, where `durationOfStay` can be obtained by simply 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 we use different database systems for each type? (3 marks)

(ii) Compare and contrast the use of row store and column store for storing data. (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) Consider the following data set for a binary classification problem:

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

(i) Calculate information gain when splitting on *A* and on *B*. Which attribute would the decision tree induction algorithm choose? Why? (3 marks)

(ii) Compare and contrast information gain and gain ratio; in which circumstances would you use gain ratio? (2 marks)

(Question 3 continues on the following page)

(Question 3 continues from the previous page)

c) 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

(i) Calculate support, confidence and lift for the association rule *{tea}* -> *{coffee}*. (2 marks)

(ii) Explain whether lift better represents the relationship between tea and coffee drinkers. (2 marks)

d) Consider ways in which standard association rule (itemset) mining might be extended. Discuss in your answer adding multi-dimensionality, quantitative intervals and hierarchies to the standard approach. (6 marks)

e) In the context of "big data": 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? (6 marks)

4. a) Using the Apriori algorithm, 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} }.

(i) At the join step of the algorithm, which itemsets are placed in C4 (the candidate set)? (4 marks)

(ii) Which itemsets are discarded by the prune step of the algorithm? (4 marks)

Provide your working as appropriate.

b) Two classifiers designed to predict patients' susceptibility to allergy are being designed and tested, independently of one another. Each of the classifiers predicts 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

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. (5 marks)

c) Would the performance of a decision tree model be improved by first clustering the training instances and then learning a different tree for each cluster? Justify your answer. (3 marks)

d) 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. (5 marks)

(Question 4 continues on the following page)

(Question 4 continues from the previous page)

e) With particular relevance to itemset mining, discuss the concepts of correlation and causation. (4 marks)

f) Discuss the trade-offs between privacy of the individual and the utility of data. Use case studies as appropriate to illustrate these trade-offs. How do we guarantee that data is private? (5 marks)

END OF EXAMINATION