THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

EXAMINATION

Course : BAC (PT)-61025, BSc SC-63019

Subject : COMP417 Data Warehousing & Data Mining Tech. in Business & Commerce

Group : 205, 291

Session : 2010 / 2011 Semester II

Date : 09 May 2011 Time : 18:30-21:30

Time Allowed: 3 Hours Subject Lecturer: Keith Chan

This question paper has 8 pages (cover included).

Instructions to Candidates:

Answer ALL questions in all parts.
This is an open-book examination.
Students are allowed to use a standard non-programmable calculator.

Do not turn this page until you are told to do so!

Instructions for answering questions in SECTION A

The answer to each question should include, but not limited to, the followings:

1. State any assumptions that you need to make in answering the question.
   - We assume that a "good" customer is defined to be a customer who repays his mortgage loan on time at least 90% of the time.
2. Describe the set of data that you would select for the data mining tasks concerned. In doing so, you should state explicitly the attributes and records that you would use. For example,

   The following attributes are selected:
   - Select attributes 1, 3, and 4 from Table 1.
   - Select attributes 2 and 6 from Table 2.
   - …
   - None of the attributes in Table 4 are selected.
   - …

   etc.

   The following records are selected:
   - Only the male customers in Table 1.
   - All records in Table 2.
   - Only Blue-chip stocks in Table 3.
   - None of the records in Table 4
   - …

   etc.
3. Explain why the attributes above are selected. For example,
   - Attributes 1, 3 and 4 in Table 1 are selected because they are expected to provide important demographic information about a customer.
   - Attributes 2 in Table 2 is selected because it serves as the key for the joined tables.
   - …
   - The first 50 records in Table 1 are chosen because they represent all the female customers.

   etc.
4. Explain how the selected data should be joined to form the final data set for the data mining tasks concerned. For example,
   - Join Tables 1 and 2 using Attribute 1 in Table 1 and Attribute 2 in Table 2 as keys to form a Table A so that Table A has 5 attributes. Give an example of the joined table if you prefer.
   - Join Table A with Table 3 using Attribute 1 in A and 4 in Table 3 as keys to form a Table B so that Table B has 8 attributes. Give an example of the joined table if you prefer.
   - …

   etc.
5. Describe how you would pre-process the data in the final joined table for the data mining tasks concerned. Explain why specific pre-processing steps are required. For example,
   - Attribute 1 in the final joined table is discretized because the ID3 algorithm will be used and the ID3 algorithm work better with discrete data.
   - Attribute 2 to 4 in the final joined table are combined by taking their average because the season-to-season variation is more important for buying behavior to be discovered.
   - As the differences in price provides more information above movement of stock indices, the differences in values between records in the final table are used for data mining.
   - …

   etc.
6. State what data mining algorithms you would use for data mining and explain why you choose to use one algorithm over the other. Explain if there is any alternative data mining algorithm that one can consider that can also be useful for the problems concerned and what your preferences would be if there are possible alternatives. For example,

- ● ID3 is chosen for data mining and the Risk attribute is chosen as the class labels. It is used instead of k-NN because the data involved are discrete and it is hard for k-NN to be used with discrete data. Other than ID3, one can also con
- ● In using the hierarchical clustering algorithm, the Euclidean Distance measure is used and the number of clusters is found in the dendrogram…, etc.
- ● …

etc.

7. Describe the output that you expect to obtain as a result of the application of a data mining algorithm. Explain how the output is related to the answers of the questions that you are trying to look for. Explain if you need to perform further data mining on the output discovered. If so, describe explicitly what you would do by repeating the above descriptions from 1 to 7.

SECTION A: (50 Marks)

In Semester 1 of 2002, some students living in the student hostels of the university decided to get together to form a study group. One thing that they did was to get everyone who lived in the hostels to contribute the study and reference materials of all the subjects that they had previously taken. These materials were then stored electronically in a central document repository that was made available to all students living in the student hostels. Over the years, a total of over 28,000 documents have been collected and indexed in a database. An example of the data that are kept about these documents is given in Table 1.

Table 1 Document Data

Document Number	***Document Type***	***Student ID of Document Contributor***	***Subjects Concerned***	***Possible Other Relevant Subjects***	***Keywords***
DAZ001	Exam paper	10118877d	MM320 Management Information Systems	COMP 321, COMP 331, COMP431	IT, DSS, data mining, DBMS, data warehouse, internet, eCommerce, data security, encryption, artificial intelligence, fuzzy logic, GA.
DCH010	Assignment	09101773d	MM332 Marketing Principles	MM 432	Direct marketing, CRM, market segmentation.
DBY103	Quiz	06738843d	EIE 221 Signal Processing	EIE 223 EIE 331	FFT, Markov Chain, image processing, fuzzy logic.
DYZ104	Lecture notes	04456672x	EE 330 Power Systems	EE 317 EE 318	Reliability, domestic consumption, power meters, GA, fuzzy logic.
DTU505	Lab notes	08765443r	SD 342 Product Design	SD 448	Concept rendering, thermatic design, sketching.
…	…	…	…	…	…

In the above table, a Document Number is used to identify each document. Whenever a student contributes a document to the central repository, a document number will be generated for it and the student will be asked to provide their Student ID and to state what subject the document is obtained from. He or she will also be asked what other subjects the student thinks that it can be relevant to. In addition, the student is to provide a list of keywords that can be found in the document so as to facilitate any searching process.

The documents in the central document repository are made available by students who live in the student hostels to students who live there. Residents who are interested in contributing to or requesting for documents in the central repository are asked to first create an account and to fill out a questionnaire. The questions that are asked include what program the student is in, if the program is an undergraduate (UG) or postgraduate (PG) program or if it is a Higher Diploma (HD) or Associate Degree (AD) program. The student would also be asked which hostels they are living in or have previously lived in and they will be asked if they have participated in any sports (basketball (BB), volleyball (VB), football (FB), table tennis (TT), badminton (BM), swimming (SW), or track-and-field (TF)), or social (chess, bridge, choir, social services, band and cooking competitions) activities during their period of residency. An example of the data that are kept about the document requestors are shown in Table 2.

Table 2. Document Requestors and Contributors

Student ID	***Date Account Created***	***Program***	***PG or UG or HD or AD***	***Hostels Stayed***	***Period of Hostel Residency***	***Sports activities participated***	***Social activities participated***
06118877d	17/12/06	COMP	UG	LS	06-07	BB, VB, FB, TT, BM	Chess
08101773d	01/09/09	SN	UG	WX, MY	08-11	BB, SW, TF	Bridge
03738843d	28/02/04	LGT	HD	KY	03-04	TF, SW	Choir
02456672x	22/09/04	MM	PG	MW, MY	03-05	FB, TT	Chess, Choir
10765443r	15/09/10	RS	AD	MY	10-11	BM, TF	Service
07118877d	12/02/08	COMP	UG	XM	07-09	VB, TT, BM, SW	Service, Cook, Choir, Chess
08101773d	21/01/09	SN	UG	LX	08-09	N/A	Cook, Band
09738843d	11/12/09	ISE	AD	LX, LS	09-11	N/A	Chess
04456672x	07/09/06	MM	PG	KY, MY	05-08	VB	Service, Cook
…	…	…		…	…	…	…

The server that hosts the central document repository keeps track of all checking ins and outs of documents. For each connected session, a unique session ID is generated. In addition, the users' student IDs, the date and time that a session begins and ends, and the documents that are checked in and/or checked out are recorded. An example of such recorded data is given in Table 3.

Table 3. Accesses to the Central Document Repository

Session ID	***Date***	***Beginning Time***	***Ending Time***	***User Student ID***	***Documents checked in (Document No.)***	***Documents checked out (Document No.)***
001	23/3/08	13:13:56	15:04:16	06118877d	DHC121, DHC122	DAB710, DHC277, DKC766, DLL012, DLL104
002	23/3/08	15:06:09	22:26:09	07101773d	DPB211	DPB107, DPB004, DPB218, DPP209
003	23/3/08	22:20:15	22:33:32	05738843d	DBB134, DSU761, DUC441, DFI377, DKJ899, DRT7542	Nil
004	23/3/08	23:01:31	23:03:00	04456672x	Nil	Nil
005	24/3/08	03:15:37	06:24:51	07765443r	Nil	DHK211, DHK212, DHK213, DHL245, DEHH211

006	24/3/08	11:52:01	11:52:31	04118877d	DJJ243	DJ010
007	24/3/08	16:34:12	16:34:31	05101773d	DKL001, DDT900	Nil
008	25/3/08	00:56:45	00:56:31	06118877d	Nil	DPP701, DPP324, DPP415
009	26/3/08	09:27:56	09:27:31	07101773d	Nil	DJJ021
...	...	...	...	...	...	...

Each document requester is sent an email, two weeks after the documents are checked out, to ask them to rate the usefulness of the documents to them and to give comments on how much the document may benefit other students. The ratings, which can be from 1 to 5 with 1 being "bad", 2 being "not so good", 3 being "quite good", 4 being "good" and 5 being "very good", and comments on each document are kept in the central repository and made available to those who may be interested. An example of such a record is given in Table 4 below:

Table 4. Rating and Comments on Documents

Document ID	***Rated by (Student ID)***	***Rated on (Date)***	***Rating From 1 to 5***	***Comments***
DAZ001	06118877d	14/4/08	4	*This should be understood very well before the exam as most of the questions are taken from this set of notes.*
DAZ002	08101773d	15/12/10	3	*This is not very useful as it is a very high level description of the CRM concept. For those of you who do not have the background, you may read this document.*
DAZ003	03738843d	01/02/04	5	*The questions in this test are very similar to those in the exams. I think this is a must-read for anyone who would like to pass the exam without putting in too much effort.*
DAZ004	02456672x	13/10/03	1	*I am not sure who prepared these solutions to the assignment. I think they are all wrong. This is an absolutely useless document. No one should waste time on it.*
DAZ005	10765443r	22/12/10	2	*This supplementary lab preparation notes was prepared two years ago by a TA not familiar with the subject area. The English is also not very good. I do not think that it is very useful.*
DAZ006	07118877d	01/05/10	4	*This test paper help clear some of my misunderstanding about internet security and cryptography. I think it is a useful piece of study material.*
...	...	...	...	...

In order to justify the additional resources used to operate the central document repository and to support the study groups, the hostel management has also obtained data related to the academic results and achievements of the student residents. An example of such data is given in Table 5 below. The data include the student IDs, the gender of the students, the birthday, the code of the subjects taken, the semesters in which the subjects were taken and the grades obtained for the subject. It should be noted that different students take different number of subjects depending on program requirements as well as on the number of years they have been studying at PolyU. Students who have just started their studies will have fewer subjects recorded in the database.

Table 5. GPA of the student residents

Student ID	***Sex***	***Birthday***	***Subj 1 Code***	***Subj 1 Sem***	***Subj 1 Grade***	**...**	**...**	
06118877d	M	03/01/88	COMP202	2007s2	B+	...	...	...
08101773d	M	12/12/90	SN314	2009s1	C	...	...	...
03738843d	M	21/07/86	LGT232	2004s2	B	...	...	...
02456672x	F	16/09/83	MM471	2004s1	C+	...	...	...
10765443r	F	09/04/93	RS277	2011s2	A	...	...	...
...	...	...	...	...	...	...	...	...

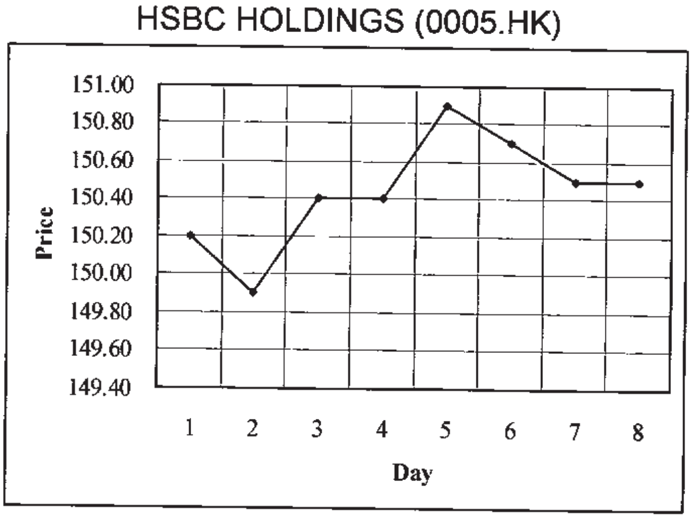
1. As there are over 28,000 documents in the central repository, it has become difficult for students to find documents that are the most useful and relevant for them. To help students with document search, a *recommender system* that is able to recommend the most useful and relevant documents to different students in different disciplines are needed. Describe how you would use data mining in such a recommender system. (18 Marks)

2. In order to evaluate the effectiveness of the services provided through the central document repository, the hostel management would like to find evidence for the improvement of GPA for the active users of the repository. Explain how you would use data mining to help the hostel management find such evidence. (18 Marks)

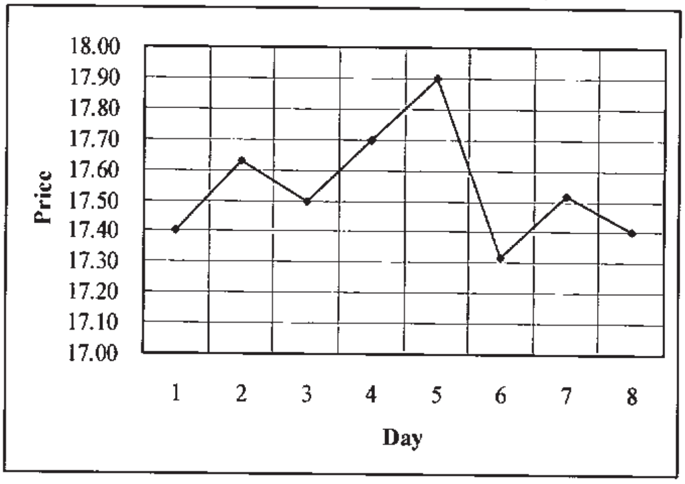
3. It is discovered that, of the 28,000+ documents in the central repository, about 20% of them are either not accessed at all or all accessed only a few times. There have been suggestions that these documents be removed from the repository to make the size more manageable. However, there are also concerns that the removing of documents that are not popular may result in the removing of documents that are very useful but are just not noticed. Explain how data mining can be used to identify such documents. Explain how we can use the data mining results to reject the contribution of a new document that is predicted to be unpopular and not useful. (14 Marks)

SECTION B: (30 Marks)

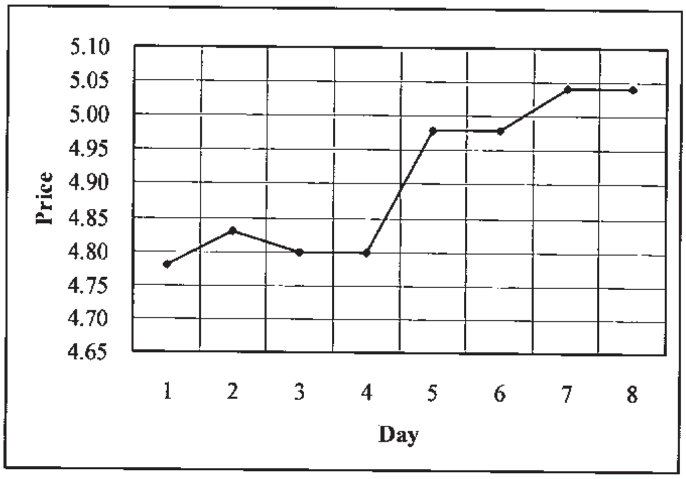
You are given the daily closing price from Day 1 to Day 8 of three different stocks HSBC Holdings, BOC Hong Kong and PCCW as follows.



BOC HONG KONG (2388.HK)



PCCW (0008.HK)



4. Find the lift ratio of the rule: "If the Price of either HSBC or BOC Hong Kong goes down today then the Price of PCCW will neither go up nor down tomorrow." Please show your steps. (3 Marks)
5. Find the support and confidence of the rule "If the price movements of BOC Hong Kong and PCCW were the same yesterday then the Price of PCCW will remain unchanged today." Please show your steps. (3 Marks)
6. Show how you can use ID3 to discover all interesting rules related to between-day stock price movements such as "when compared with the price yesterday, if the stock price of HSBC has gone up and the stock price of BOC Hong Kong has gone down then the stock price of PCCW would have remained unchanged." (7 Marks)
7. Show how you can use the the Apriori Algorithm with a minimum support of 25% and a minimum confidence of 60% to discover rules such as "if the stock price for BOC Hong Kong increases for

two days in a row, then the price of HSBC will decrease for the next two days in a row". Please show your steps. (7 Marks)

8. It is suspected that the stock price movements of the two bank stocks are "similar" to each other and are different from that of the PCCW stock. Show how you can confirm if this is the case by using a clustering algorithm. Please show your steps. (5 Marks)
9. Explain how you can make use of the concept of fuzzy set to improve data mining in this question. Please show some examples of fuzzy sets that you would use for the three stocks above. (5 Marks)

SECTION C: (20 Marks)

The star schema of a data warehouse can be considered to have three dimensions: *time, product* and *location*, and the two measures: *Sales_Amount* and *Unit_Sold.*

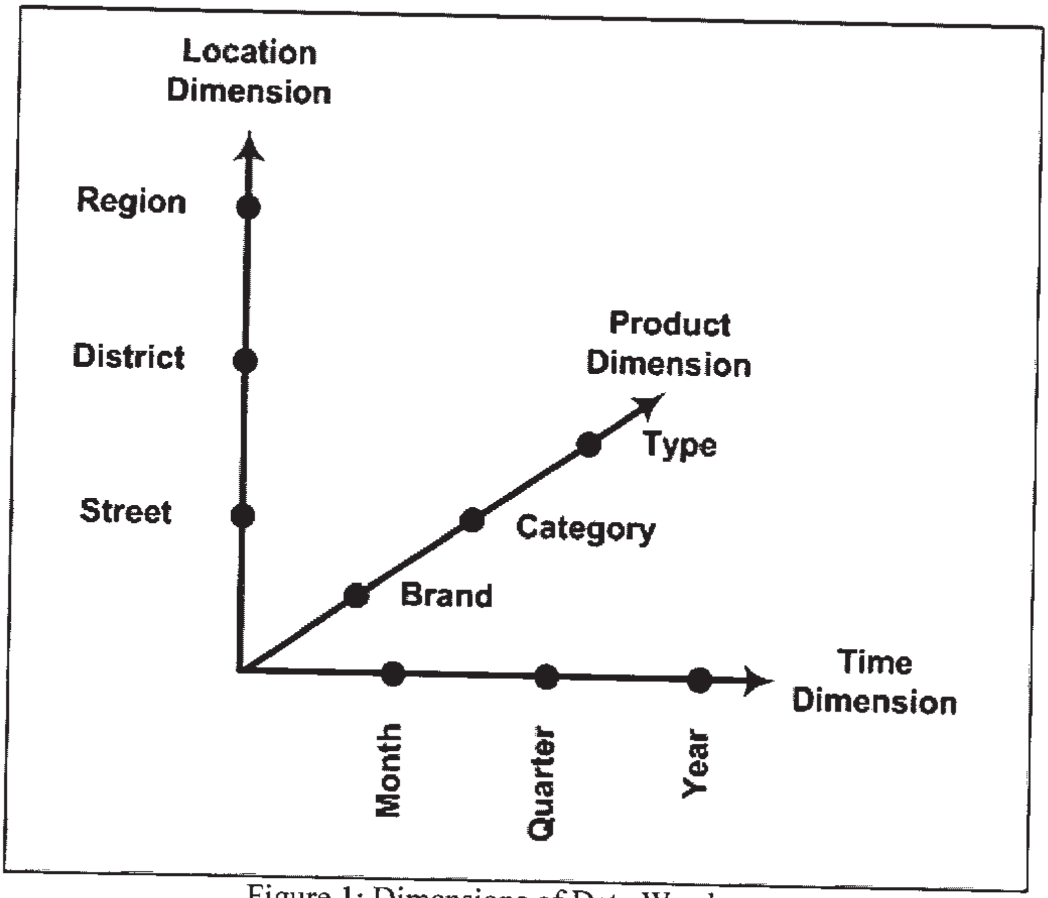


Figure 1: Dimensions of Data Warehouse

10. Draw the *star schema* diagram for the data warehouse. (5 Marks)
11. Starting with the base cuboid {*time, product, location*}, what specific *OLAP* operations (e.g. *roll up time to quarter (level))* should one perform in order to list the total *Sales Amount* of *Notebook PC* in *2009's summer* at *Wanchai district*. (6 Marks)
12. In your opinion, would it be better if the data warehouse use a *snowflake schema* instead of a *star schema*? Why or why not? (3 Marks)
13. Besides the data schema, discuss the criteria that one need to consider in deciding on a suitable hardware configuration for the data warehouse? (6 Marks)

*** END ***

***** Please submit question paper together with your answer booklet *****