CSC400W/CSC403W/CSC416W – Second Course in Databases - 2005

Paper 1 – Open Book Examination

Time: 2 hours
Maximum mark: 50

Answer 5 (five**) of the 9 questions; including at least 1 question from each section**.

SECTION A – Answer at least one question from this section

Question 1

Answer all the following questions with respect to ***any one*** of the papers which were covered in this course.

(a) Summarise the main contribution of the paper in your own words, in one or two sentences – no more than two sentences will be marked.

[1]

(b) Explain briefly the significance of the paper's research goal – i.e. why is it important to solve this problem or tackle this issue?

[1]

(c) If you were teaching students how to write journal/conference papers, would you use this paper as an example of good writing or of bad writing? Justify your answer.

[1]

(d) You are part of a panel of editors choosing papers for a book that will cover a sub-field of database research for which the material in this paper is very relevant. Argue strongly for its inclusion among those selected.

[3]

(e) Now argue strongly for excluding this paper from the book because of its weaknesses. Be as critical as you can, but remain reasonable – your comments will be sent back to the authors with their rejection letter.

[3]

(f) You have just accepted a job as research assistant to the group who wrote this paper. You need to find a research project that follows on from the work in the paper. Briefly describe the research question or goal you would tackle (*not how* you would go about it, just *what* you would investigate).

[1]

Question 2

Answer any 2 (two) parts of this question – each carries 5 marks.

(a) Explain each of the following terms, and briefly describe their relevance to **mobile database** research:

1. PicoDBMS
2. epidemic algorithms

[5]

(b) *Sqldomgen* [McClure and Kruger 2005] has an abstract data model for querying databases dynamically, which it instantiates for any given schema to generate concrete classes capable of manipulating the database dynamically.

1. For the example schema in the paper, give example *sqldomgen* code for function UpdateOrder for adding R400 to the Freight cost of all Orders shipped from the city whose name is stored in parameter ShipCity.
2. Name any two parts of SQL that are not covered explicitly in the examples of this paper, and discuss how *sqldomgen* may or may not cope with these.

[5]

(c) The **Query Optimizer** described in the paper of Babcock and Chaudhuri addresses the problem of uncertainty in query optimisation. In relation to this work, state whether each of the following is true or false, giving a brief reason for each answer:

1. The uncertainty problem tackled by this paper arises because many values in database tuples are incomplete or imprecise – e.g. a person's income may be mistyped, unknown, approximate or inapplicable (e.g. the aged).
2. Using Bayesian inference for their query optimisation is impractical for real-world situations because Bayesian learning takes too long and hence will slow down queries unacceptably.
3. The graph of execution costs in figure 1 of the paper is included to show that the authors' approach is better than existing methods of doing query optimisation.
4. The paper analyses the effect of different confidence thresholds on query execution time in order to show that providing an imprecise or approximate answer to a query will be faster than providing an exact result.
5. The authors advocate an approach where the confidence threshold is set by the database administrator but can be overridden (set to a different value) for a specific query by the end-user.

[5]

(d) Jasper and Uschold produced a framework for **ontology applications**. According to their framework, which type of ontology application is each of the following:

1. A multi-disciplinary team of scientists is starting work on an agricultural development project. They build an ontology together and then use this as a basis for the requirements analysis, system design and data gathering stages of the project.
2. Company A builds an ontology of computer hardware and software using Ontolingua. When they subsequently buy over company B, this ontology is converted into Prolog and used with company B's database to improve database searches.
3. The WordNet ontology is being used in a peer-to-peer database system to map the attributes/columns of one database onto their corresponding attributes/columns in another database – since column names in different databases are often synonyms rather than identical.
4. Two companies wish to share data in their (separate and very different) databases. Each company creates an ontology to describe their model of the application domain. An ontological engineer is then asked to use these two ontologies to derive a mapping from the attributes/columns of the one database onto their corresponding attributes/columns in the other database.
5. You are a keen historian and have an extensive online library of historical documents, as well as a search engine that you built for this yourself. You discover that searching this library for articles of interest is very frustrating. You build an ontology and use this to improve your search engine.

[5]

(e) Answer the following questions with regard to the work of Kantere and Tsois on using ECA rules for **mobile query agents** in P2P database systems:

1. What does the abbreviation ECA stand for?
2. The motivating example given in the paper revolves around a doctor requesting traffic information as he drives his car. Is it significant to their use of ECA rules and mobile agents that the user is mobile? Why (or why not)?
3. What are the advantages and disadvantages of having mobile agents for query processing in P2P systems?

[5]

(f) Lawrence and Barker use a **standardized dictionary** for integrating database schemas. Suppose that the standard dictionary in figure 3 of the paper is being used, and a new database Staff-Reports is to be integrated, with the following schema:

Book (Book_Id, Description, Author_Id)
Employee (Name, Employee_Id)
Company (Name, Company_Id)

where Book gives the id and description of each staff report (book) and the ID of the employee or company who produced it; Employee gives the name and ID of employees; and Company gives the name and ID of each company. Note that books written by a single person have his/her Employee_Id as their Author_Id, while books written by more than one person have their Company_Id as Author_Id instead.

1. Draw up a table to show the Semantic Name that would be associated with each System Name for Staff-Reports (you can omit the "Type" values).
2. Take any one entry in this table and show how it would be specified in X-Spec
3. In the integrated view, what would be the Staff-Reports entry for the global view term [Author] – Id ?

[5]

(g) Consider the paper from Arizona State University that addresses the problem of answering **imprecise queries**, and answer the questions below with respect to this paper:

1. Why do they use a workload database of queries?
2. Why do they not use semantic similarity to measure distance between 2 queries?
3. Why did they use students who were very familiar with the BibFinder test database in their experiment?
4. State briefly any two factors that could have affected the results of this experiment.

[5]

(h) Answer the questions below with reference to the technique of Westmann et al for storing **compressed data** in a database:

1. The length encoding "01" for integers means something different from the length encoding "01" for doubles – explain why.
2. For relation R below:

   R(w : double nulls not allowed; x : double; y : int; z : int nulls not allowed)

   what would be the length encoding byte for tuple 1 of relation R if its w value is 4 bytes long, its x value is 0 bytes long, its y value is 3 bytes long and its z value is 1 byte long?
3. The offset of the z value of this tuple would be looked up in which element of the decoding table?

4. What would the entry in that element of the decoding table be?

[5]

(i) Consider the University of Toronto paper on **mapping of data in P2P systems**, and answer the questions below:

1. Do the mappings in the mapping tables represent mappings between attribute names (e.g. Category in one database matches Subject in another) or do they represent mappings between values (e.g. category “D.2.12” in one database is mapped onto subject “Interoperability” in another database)?
2. Their motivating example is genome databases. Would their approach be useful in mapping between different Client databases (i.e. with people’s names, ID numbers, addresses, phone numbers, purchase history, credit ranking etc.) to enable sister-companies to benefit from each others’ client knowledge? Briefly give a reason for your answer.
3. Why is the open-closed-world semantics not considered in their automated mapping discovery system?
4. Suppose that closed-open-world semantics are being used on the mapping tables below. State whether each of the derived facts below satisfies the constraints represented in these mapping tables or not. Mapping table T1 maps two attributes of DB1 shown in bold onto the Keyword attribute of DB2. In table T2, the Category attribute of DB1 is mapped onto the Author attribute of DB2. In table T3, the Category attribute of DB1 is mapped onto the Keyword attribute of DB3.

Category	**Author**	Keyword
D.12.2	**Li**	portability
H.2.3	**Hu**	usability
H.2.3	**Yu**	SQL

T1

Category	Author
H.2.3	Yu

T2

Category	Keyword
H.2.4	XML

T3

a. Category H.2.3 may map onto usability.
b. Category H.2.3 may map onto SQL.

[5]

SECTION B

Question 3 Object-Oriented Databases

(a) What is the difference between a persistent programming language and an object-oriented database?

[2]

(b) Briefly state any one advantage and any one disadvantage of an object-oriented database compared with a persistent programming language.

[2]

(c) Compare the relative merits of ***any 2 (two)*** of:- the O2 object-oriented DBMS, ODMG C++ and Orthogonally Persistent Java (as described by Atkinson et al in their SIGMOD Paper "Towards an Orthogonally Persistent Java"). Contrast their approaches to identifying which objects are persistent, their handling of pointers to persistent objects, their handling of class extents, their support for a query language as well as noting any other major differences between them.

[6]

Question 4 Object-Relational Databases

(a) You are asked to design a database for the Computer Science Department. It must contain records of students, the programs they submit, the tutorial groups they belong to, the marker who marks the work of each tutorial group and the mark allocated to each program by its marker. Show any one example of how you might use each of the following if you were to use an object-relational database for this task: structured types, large object types, inheritance, references, methods. (Note: do not give a complete database schema, just one example of each of these O-R features).

[5]

(b) Consider the relational database schema S below, which stores information about projects in a company, each of which comprise 3 or 4 stages (e.g. design, testing) and each of which has 2 to 4 special roles (e.g. Supervisor, Tester). The database stores, for each role, the name of the employee filling that role as well as his/her level of expertise and the rate at which the company charges out his/her time. It also records the manager of each stage of every project, as well as the due date when that stage should be completed.

Project (ProjNum, Client, Fee)
Stages (ProjNum, StageNum, Manager, DueDate)
Roles (ProjNum, Role, Employee, EmpLevel, HourlyCharge)

1. Show how you would use nesting in an object-relational SELECT statement to obtain a view V where all of the above information is in a single relation.
2. If someone using your view V wanted to obtain a first normal form relation containing ProjNum, StageNum and Manager, how would they use unnesting to achieve this?

[5]

Question 5 Distributed Databases

(a) Consider a relation Sales (Branch, Salesperson, DateSold, CarModel, CarYear, Price, Comment) which is part of the database of the Ajax company that sells second-hand cars at various branches across the country. The Comment attribute is a text field containing all the notes made by the Salesperson about that car, the buyer and the sale itself. Suppose that Ajax wishes to move to a distributed database system, and have some data stored at the biggest branch in each of the eleven provinces. Suggest how each of the following might sensibly be used in this distributed database:
   1. vertical fragmentation of Sales
   2. horizontal fragmentation of Sales
   3. replication of any of these fragments

[3]

(b) The biggest branch in the Eastern Cape is the PE branch. Suppose somebody at the PE site submits a query that needs to access information at all eleven sites in the distributed database. State briefly what would be the role of each of the following in this transaction, if any:
   1. transaction manager at the PE site
   2. transaction manager at the Cape Town site
   3. transaction coordinator at the PE site
   4. transaction coordinator at the Cape Town site

[4]

(c) Suppose that another relation in the database keeps information about the people employed to sell cars at Ajax:
Emp (Branch, Salesperson, Salary, NumSales)
If this table is neither fragmented nor replicated, but is kept in its entirety at the Head Office branch in Pretoria, what would be the effect of using a semijoin to compute the join of Sales data at the PE branch with this Emp data? Describe which information would be shipped between sites for each step of the semijoin operation. Assume that the query comes from the PE site (and hence the result is needed there).

[3]

Question 6 Data Analysis and OLAP

Staff publications at a local university are kept in a database for subsidy and other purposes. Tuples in this database contain information about the total number of journal publications, book publications and conference publications made per year in each Faculty (the Faculties are: Law, Science, Commerce and Arts). This information has been kept for the period 2001 to 2005.

(a) By means of a simple diagram, show what a cross tabulation of Publications by Faculty and year might look like (do not fill in any values inside the cross tabulation, just show its structure).

[3]

(b) By means of a simple diagram, show what a 3D data cube for Publications might look like (do not fill in any values inside the cross tabulation, just show its structure).

[4]

(c) **EITHER:**

a. Give the SELECT statement that would generate the 3D data cube in (b) above.

OR:

b. Give a SELECT statement to give for each Faculty and publication type (book/journal/conference) the position or rank of that Faculty in the university as regards its 2004 publication count of that type. For example, the tuple (Arts, journal, 4) in the result would mean that the Arts Faculty ranked fourth (i.e. last) over all Faculties in the year 2004 as regards its total production of journal publications.

[3]

Question 7 Data Mining and Warehousing

(a) Show an example of a decision tree for rating first-year students as “unlikely”, “average” or “good” graduates on the basis of information available about them when they first arrive at university: their age, their matric average symbol, and their performance in the university’s AARP Test (which tries to gauge potential as a score between zero and 100). Assume that matric symbols can only be one of the following: A,B,C,D,E,F (where A is 80% and more, B is 70% - 79%, etc. and F is less than 40%). As you do not have the data before you, draw any reasonable decision tree to show that you understand the concept.

[3]

(b) **EITHER** answer the following questions about your decision tree in part (a):

a. In practice, you would have real data of past students at hand and use this to build your decision tree. How would you be able to tell if you have a good decision tree or not?
b. And if you do have a good tree, how would you use this decision tree to guide you in deciding which students to accept into the university?
c. In reality, we have far more information about first-year students than that given above; there are very many attributes (the school they went to, their home address, etc. etc.) which we could take into account. In doing so, how would we prevent the decision tree from becoming too large?

OR answer the following question on association rules:

d. The university keeps records of all course results of all its students at the end of every semester, and has data going back twenty years. They wish to use associations rules and data mining to try and see which situations lead to students being excluded. Give any one example of an association rule for this situation, and explain the difference between the support of this rule and the confidence of this rule.

[3]

(c) ABC is an international company that specializes in line support/consulting – employees answer questions from people anywhere in the world, and ABC is paid for this help. ABC wishes to keep a data warehouse of all its transactions, where each record will store the following: employeeID, clientID, dateQuestionRaised, Topic, FeeCharged, NumberOfMessages (where NumberOfMessages gives the number of times the employee had to reply to the client before the question was fully resolved). Details of employees (their country, salary, home language), clients (country, home language, creditRating) and topics (description, level of difficulty) are also relevant. Show by means of a simple diagram how you would design a star schema for this data warehouse.

[4]

Question 8 Temporal Databases

(a) Consider relation LOAN (clientID, ISBN, dateTaken, dateReturned, FinePaid) that stores information about clients' loans of

books (identified by their unique ISBN number) along with the fine paid for late return, if any. A second relation CLIENTS (clientID, name, dateJoined, dateLeft, finesOwed) stores client details. (Note: in answering the questions that follow, use date values like d1,d2, d3 etc. rather than 07/09/2005 etc. in order to save time.)

1. Give an example of two tuples, t1 and t2, such that t1 BEFORE t2 is true
2. Give an example of two tuples, t1 and t2, such that t1 ENDS t2 is true
3. Give an example of two tuples, t1 and t2, such that t1 MERGES t2 is true
4. Give an example of a temporal query that is difficult to answer in SQL and requires a temporal query language to make it feasible.
5. Give an example of a temporal constraint on the above database that is too difficult to enforce without a temporal query language.
6. Give an example of a temporal query that is considerably simplified if Allen's operators are used, and show how that query could be answered using Allen's operators.
7. Would you say that clientID → finesOwed is a functional dependency or a temporal functional dependency of CLIENTS? Give a reason for your answer.
8. Suppose we join the LOAN and CLIENTS relations for all tuples where clientID is 401, and then we take the temporal join of LOAN and CLIENTS where clientID is 401, and the size of these two results is different. Would you expect there to be more tuples in the temporal join, or fewer tuples? Give a reason for your answer.
9. Suppose we project the name column out of the CLIENTS relation and then we do the temporal projection of the name column out of the CLIENTS relation, and the size of the two results is different. Would you expect there to be more tuples in the temporal projection, or fewer tuples? Give a reason for your answer.

[10]

Question 9 Spatial databases

Answer any 2 (two) of the three parts in this question – each carries 5 marks.

(a) Data kept at the Sea Fisheries Institute (SFI) includes satellite images indicating the sea surface temperature (SST) off our coast at a particular time each day, and current readings taken by 50 current meters that are positioned (anchored) at specific spots in the sea near Cape Town. These current meters automatically record the strength

and direction of the sea current every 6 hours. SFI also keeps “catch” records obtained from fishing boats, which indicate the “grid cell” in the ocean where they fished each day, and the number of fish that they caught there. (SFI has drawn a rectangular grid over our coastal waters so that boats can easily indicate their position simply by giving the grid cell number.)

1. Is the SST data vector data or raster data? If it is raster data, explain why. If it is vector data, indicate what type of vector data (e.g. point, line or polygon).
2. Is the current strength data vector data or raster data? If it is raster data, explain why. If it is vector data, indicate what type of vector data (e.g. point, line or polygon).
3. Is the “catch” data vector data or raster data? If it is raster data, explain why. If it is vector data, indicate what type of vector data (e.g. point, line or polygon).
4. SFI often ask to see which areas of the sea had low SST values and high catch values – what type of query is this: nearest-neighbour, region query, spatial intersection or spatial union? Briefly state why.
5. SFI often ask to see which current meters lie within a particular grid cell – what type of query is this: nearest-neighbour, region query, spatial intersection or spatial union? Briefly state why.

[5]

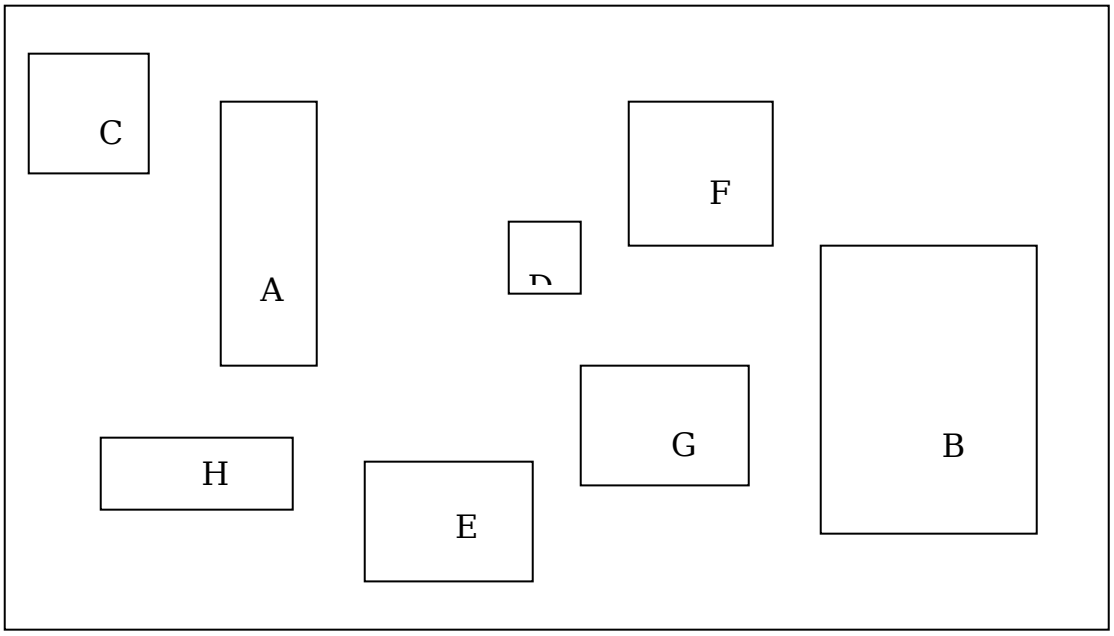
(b) Consider the diagram on the last page of this paper. It represents a part of the sea and the location of current meters in that area. Each current meter is indicated by a small circle. *NB Remember to tear off this late page, put **your name on it**, and place it inside your answer book afterwards!*

1. On the upper diagram in the appendix, show how this area would be divided up using EITHER a k-d tree OR a quadtree, assuming no more than one point can be accommodated in any partition. State which you have chosen to do.
2. Show EITHER the k-d tree index structure corresponding to the k-d tree subdivision you did in the diagram of the appendix, OR show the quadtree index structure corresponding to the quadtree subdivision you did in the diagram of the appendix. State which you have chosen to draw.

[5]

(c) Consider the diagram overleaf. Show how an R-tree would be used to index the objects labeled A to H, assuming that they were inserted in alphabetical order and that the nodes of the R-tree can store at most 3 values. Show all your working. *NB Remember to tear off the appendix**, put your name on it**, and place it inside your answer book afterwards!*

[5]



NAME: _____ Student Number: _____