

DAIS Qualifying Examination  
*Spring 2008*

Department of Computer Science  
University of Illinois at Urbana-Champaign

Feb. 14, 2008

Time Limit: 180 minutes

Please read the following instructions carefully before starting the test:

- The examination has 13 questions, including one basic concept question and 12 topic questions. You are required to answer the basic concept question and any **5** of the 12 topic questions of your choice. If you answer more than five topic questions, the committee will randomly select five to grade.
- The basic concept question has **12** subquestions. You are required to answer **9** of these 12 subquestions. If you answer more than 9 subquestions, the committee will randomly select 9 to grade.
- The 12 topic questions are distributed by areas as follows:
  - **Database Systems:** 3 questions
  - **Data Mining:** 3 questions
  - **Information Retrieval:** 3 questions
  - **Bioinformatics:** 3 questions

Each of these 12 questions generally has three parts. Make sure that you answer all the three parts of any topic question that you have chosen to answer.

- Use a separate booklet for each question that you answer. That is, you will use a total of six booklets (1 for the basic concept question and the rest for the 5 chosen topic questions).
- To ensure the fairness of grading, all the exams will be kept anonymous. Your answer sheets should thus only bear the assigned ID but *no names*.
- Please make your answers clear and succinct; you will lose credit for verbose, convoluted, or confusing answers. *Succinctness does count!*
- The questions have been designed to avoid any ambiguity. However, in case you are not sure how to interpret a question, please assume one interpretation, write down your assumption, and solve the problem accordingly.
- On average, you will have 30 minutes for each of the six questions that you have chosen to answer. So plan your time accordingly, and keep in mind that some questions/subquestions may demand more writing than others.

## Required Question (Basic Concepts): Problem 0

You are required to answer 9 out of the following 12 subquestions. Each subquestion is worth 1 point. Please focus on the major point and keep your answer concise; in general, an answer is expected to be no more than two sentences.

(1) (*SQL*)

Consider the schema ED(employee, department) and DM(department, manager). These relations describe who works in each department, and who manages each department. Each employee works in only one department, and each department has only one manager. Write an SQL query that lists, for each manager, the total salary paid to employees who work in a department managed by that manager. In the answer, list the managers in alphabetical order by name.

(2) (*query optimization*)

Consider the schema Recipe(Dish, Ingredient, Amount) and Produces(Ingredient, Manufacturer), and the query “List all dishes that include more than 1 cup of an ingredient made by manufacturer Nestle.” (For simplicity, assume all ingredients are measured in cups.) Suppose we are trying to decide between the following two query plans. Under which circumstances should we pick the first one, and when should we pick the second one?

1. Find the ingredients made by Nestle (selection in Produces followed by projection onto Ingredient). Then find the Dishes that use that Ingredient (join with Recipe, using a B-tree on Recipe.Ingredient to find the matching tuples for each Ingredient)). Eliminate the tuples that use less than or equal to 1 cup of their Ingredient (selection), and then keep just the names of those Dishes (projection onto Dish).
2. Find the Dishes that use more than 1 cup of some ingredient (selection in Recipe). Then find the Manufacturers of those ingredients (join with Produces, using a B-tree on Produces.Ingredient to find the matching Manufacturers for each Ingredient). Eliminate the tuples where the manufacturer is not Nestle (selection). Finally, keep just the names of the remaining Dishes (projection onto Dish).

(3) (*benchmarks*)

Suppose I invent a new concurrency control and logging technique that will make recovery after a crash much faster. I want to demonstrate that my new technique does not slow down query processing very much. What standard benchmark should I use for this purpose, and why?

**Note: Please start with a new answer sheet.**

- (4) (data cube and OLAP) A 10-dimensional data cube, with the measure *count*, has only 2 nonempty cells in its base cuboid. (1) What is the maximum possible number of aggregate cells in the cube? (2) what is the minimum possible number of aggregate cells in the cube? (3) if the minimum support (i.e., *iceberg condition*) is 2, what is the minimum possible number of aggregate cells in the *iceberg cube*?
- (5) (classification) Use one sentence to distinguish the following pairs of methods: (1) SVM vs. neural network algorithms, (2) decision-tree (such as C4.5) vs. RAINFOREST algorithms, (3) boosting vs. ensemble methods.

- (6) (selection of data mining algorithms) Name or outline one data mining algorithm that best fits each of the following applications: (1) capturing outliers in computer network streams, (2) partitioning high-dimensional microarray data sets into meaningful groups, and (3) promoting the sales of a set of items to a set of known customers based on the history of shopping transaction sequences.

**Note: Please start with a new answer sheet.**

- (7) (*Vector Space Model*) In the vector space retrieval model, if a query has just one term, which term weighting heuristic will be ineffective? Use no more than two sentences to explain why?
- (8) (*Language Model*) Suppose we use the query-likelihood retrieval model and Dirichlet prior smoothing. Would adding a new document into a collection sometimes cause the scores of all the old documents in the collection to change? Why?
- (9) (*PageRank*) Why is PageRank potentially better than just counting the number of in-links of a page?

**Note: Please start with a new answer sheet.**

- (10) (*alignment*) What is the additional computational cost (if any) in going from a global alignment algorithm to a local alignment algorithm, for two-sequence alignment ?
- (11) (*Hidden Markov models*) What is the motivation behind using Markov chains for modeling DNA sequences ?
- (12) (*Phylogeny*) What is the molecular clock hypothesis ?

# Database Systems: Problem 1. Complex Queries over Web Repositories

The paper “Complex Queries over Web Repositories” argues the need for complex queries for accessing information in a Web repository like Stanford WebBase.

## **Part 1**

- a) Give one example query that is *not* from the paper, and *can* be expressed in the proposed query algebra. (*2 points*)
- b) Briefly explain how this query can be expressed in the algebra. Note: We are not asking for an exact query expression. Just sketch the main components. (*2 points*)

## **Part 2**

Compare the proposed query algebra with standard *relational algebra*. Name *two* differences. For each one, identify it, and briefly explain why it is necessary. (*2 points*)

## **Part 3**

Do you think there are queries that *cannot* be expressed in the proposed algebra? Explain your answer, and identify such an example if you think so. (*4 points*)

## Database Systems: Problem 2. Type-annotated Search

In the “Optimizing Scoring Functions and Indexes ...” paper, the authors propose to support *proximity search* over *type-annotated* corpora.

### Part 1

Compare such search with the standard search function of today’s search engines (e.g., Google). Name *two* differences? (*2 points*)

### Part 2

For supporting *proximity search*, the paper proposes a scoring model, which gives a score to a candidate token depending on three quantities: 1) statistical properties of matched selectors in the vicinity, 2) the distance between those selectors and the candidate, and 3) the manner in which contributions from different selectors are aggregated at the candidate token. Do you think there are other quantities that should also be considered? Explain why or why not. (*2 points*)

### Part 3

Since this paper focuses on *proximity search* for finding typed data entities, its scoring scheme considers only the textual patterns (such as distances) between selector keywords and candidate tokens, which are important for measuring the *proximity* factor. We wonder if proximity alone is the *only* factor to consider, for searching typed entities:

- a) Give one example query, for which proximity search alone may not return the right results. Explain why. (*2 points*)
- b) For your example query, explain what other factors must be considered in order to return the desired results. (*4 points*).

## Database Systems: Problem 3. Ranking in RDBMS

### **Part 1**

What is the purpose of *ranking* for querying? Give one application scenario for querying a DBMS where ranking might be crucial. To contrast, give another scenario where ranking is irrelevant. (2 points)

### **Part 2**

Does SQL support ranking or not? Explain your answers either way. If *yes*, what is the construct that supports ranking? If *no*, is there any construct close to it?

(2 points).

### **Part 3**

(a) What does *context-sensitive ranking* in the paper of Agrawal et. al. mean? Contrast it with context-free ranking with examples. (2 point)

(b) Why does context-sensitive ranking complicate query processing? (2 points)

(c) The paper defines “democratic voting” to determine overall the ranking given a set of preferences. We ask you to *disagree* with this particular design, and propose an alternative. Argue why your design addresses the issues. (2 points)

# Data Mining and Data Warehousing: Problem 1

## Part 1

Suppose that a big chain-store has one data center (e.g., a data cube) in each state, and the data is updated daily. The managers of the store would like to know the sum and (standard) deviation of the sales by region (such as mid-west vs. north-east), by month, etc. For each of the following requests, outline an efficient computation method.

- (a) Both sum and deviation measures of each local cube can be updated efficiently in an incremental manner, and (*1 point*)
- (b) both measures of each region can be derived dynamically from the measures stored in the local cubes of the corresponding states. (*1 point*)

## Part 2

Multidimensional data modeling is essential at designing data warehouses and data cubes. Suppose each product carries an RFID (Radio-Frequency Identification) tag, and will traverse from a factory/farm to a distribution center, then to a store, a shelf, and a check-out counter, with information periodically scanned and stored.

- (a) Why do people claim that such RFID data contains a lot of redundancy? (*1 point*)
- (b) Design a data warehouse that may reduce much redundant information and facilitate OLAP on RFID data, and (*1.5 points*)
- (b) outline how such a data warehouse can facilitate path-based data mining, such as find why particular packages of milk got rotten this week. (*1.5 points*)

## Part 3

Most data cube algorithms perform efficient computation for large set of data in low dimensional space.

- (a) Give a few typical applications where high-dimensional OLAP is needed, (*1 point*)
- (b) Outline one method that may perform high-dimensional OLAP efficiently, and (*1.5 points*)
- (c) If one only would like to find top- $k$  measures for a small  $k$ , can you work out an efficient method to compute it? (*1.5 points*)

## Data Mining and Data Warehousing: Problem 2

### Part 1

- (a) Give an example to show why  $\chi^2$  and lift may not be good measures for pattern interestingness in large transaction databases. (*1 point*)
- (b) Present a good measure for mining interesting patterns in large transaction databases, and justify why this is a good measure. (*1 point*)

### Part 2

- (a) What are the major difficulties for mining frequent patterns in data streams? (*1 point*)
- (b) Given a fix amount of memory  $M$ , present an efficient algorithm that makes the best use of the available memory and mines frequent patterns in data streams with small error bound. (*2 points*)

### Part 3

- (a) Explain why frequent-pattern-based classification algorithms may lead to better classification quality than traditional classification methods? (*1 point*)
- (b) Explain why discriminative frequent patterns may lead to better better classification quality than typical associative classifier. (*1 point*)
- (c) Outline an efficient discriminative frequent pattern-based classification method. (*2 points*)

## Data Mining and Data Warehousing: Problem 3

### Part 1

- (a) Use one sentence to distinguish the following pairs of clustering methods: (1)  $k$ -means vs. EM clustering algorithms, (2) DBSCAN vs. OPTICS, (3) AGNES (i.e., Agglomerative nesting) vs. DIANA (i.e., Divisive analysis). (1.5 point)
- (b) Explain why a typical hierarchical clustering algorithm has difficulty to derive high-quality clusters, but BIRCH does not have such weakness. (1.5 point)

### Part 2

- (a) What is the major challenge of clustering high-dimensional data? (1 point)
- (b) Give two algorithms that perform high-dimensional clustering and compare their clustering quality and efficiency. (1 point)
- (c) In micro-array data analysis, each data object may contain tens of thousands of dimensions. Which one would you select from your above two algorithms, and why? (1.5 points)

### Part 3

- (a) A user may often like to give some hints to a clustering task. Illustrate why such a hint may lead to more desirable clustering result. (1 point)
- (b) There are two kinds of hints: one is to specify certain set of objects either must be or cannot be in the same cluster, the other is to use one attribute (such as *Research\_group*, as a hint, to cluster a data set (such as *Students*). Discuss which method is more desirable, and why. (1 point)
- (c) A large bibliographic information network links authors, conferences, and research publications. Outline an effective method that performs user-guided clustering of certain types of entities (such as authors or conferences). (1.5 points)

## **Information Retrieval: Problem 1. IR Evaluation**

The following questions are based on the paper “Alternatives to Bpref” [Sakai SIGIR 07].

### **Part 1**

- a What is the original motivation for Buckley and Voorhees to introduce bpref? (1 point)
- b What does a “condensed list” mean in [Sakai SIGIR 07]? (1 point)
- c The author shows that bpref can actually be regarded two metrics depending on whether a certain condition is satisfied. What is this condition? (1 point)

### **Part 2**

- a nDCG stands for normalized Discounted Cumulative Gain. Why is “Discounted Cumulative Gain” more reasonable than just plain “Cumulative Gain”? Why do we want to further normalize “Discounted Cumulative Gain”? (2 point)
- b What is the main finding of [Sakai SIGIR 07]? Briefly explain what empirical evidence has allowed the author to claim this finding (with no more than 3 sentences). (2 point)

**Part 3** In Web search, a user would not only benefit from the relevant content in a web page, but also benefit from the possibility of navigating into other useful pages from a page. None of the measures studied in [Sakai 07] can reflect the value of a page in helping a user navigate into a useful page. Can you propose some idea for designing a measure that can capture both the relevance of a page’s content and its indirect value of pointing to other relevant pages? Give a rough formula if possible.

(3 points)

## **Information Retrieval: Problem 2. Retrieval Model**

The following questions are based on the paper “Random Field Model for Information Retrieval” [Metzler CIKM 07].

### **Part 1**

- a The Markov Random Field retrieval model can easily go beyond bag-of-words representation and support “dependency features”. Give an example of a dependency feature and explain why adding this feature may potentially help improve retrieval performance over the simple retrieval method based on bag-of-words representation. (*2 points*)
- b In a language modeling approach such as query likelihood, we often have just one retrieval parameter (i.e., a smoothing parameter). How many parameters do we have in the Markov Random Field model? (*1 point*)

### **Part 2**

- a Given  $n$  features to consider and a test collection with known relevant documents for some sample queries, what is the basic procedure proposed for automatically selecting features? When would the procedure stop? (*2 points*)
- b How does the Markov Random Field model leverage the knowledge about the existing retrieval models such as the language modeling approaches and BM25? (*2 points*)

**Part 3** The standard query-likelihood retrieval model is based on matching single words (i.e., unigram language model). Can you propose an extension of such a model to model dependencies between words such as modeling phrases? (That is, can you extend the query-likelihood method to reward a document that matches two query words as a phrase as opposed to matching them as two separate words.) (*3 points*)

## **Information Retrieval: Problem 3. Focused Crawling**

The questions refer to the paper “Focused Crawling for both Topical Relevance and Quality of Medical Information” [Tang et al. CIKM 05].

### **Part 1**

- a What is focused crawling? (1 point)
- b What are the two criteria proposed in [Tang et al. CIKM 05] for prioritizing the pages to be crawled? (1 point)
- c Relevance feedback is used for two different purposes in [Tang et al. CIKM 05]. What are they? (1 point)

### **Part 2**

- a In [Tang et al. CIKM 05], the overall score of a URL in the crawl queue is computed as

$$URLScore = confidence_{rel} * \frac{\sum_{i=1}^m DScore_i}{m}$$

What is  $DScore_i$ ? What is  $m$ ? (1 point)

- b What does “quality locality” mean in [Tang et al. CIKM 05]? (1 point)
- c In [Tang et al. CIKM 05], a decision tree classifier is used to predict the relevance of a link target using features such as words in the anchor text. Where are the training examples from? (1 point)

**Part 3** A main challenge in focused crawling is to prioritize the potential URLs in the queue. We can regard this problem as an iterative retrieval problem: in each iteration, we would “retrieve” the most relevant potential URL from our queue and visit it. Thus many techniques of standard information retrieval can presumably be applied to improve focused crawling. Discuss how pseudo feedback can be applied to focused crawling. Can you think of any other idea for improving focused crawling? (4 points)

## Bioinformatics: Problem 1. CRM finding - Gupta and Liu, 2005.

### Part 1

What is a cis-regulatory module, and why is it important to find them ? (2 points)

### Part 2

Briefly describe the probabilistic framework adopted by Gupta and Liu for finding cis-regulatory modules. (5 points)

### Part 3

How will you extend the basic approach of Gupta and Liu to an integrated framework that does not require any given motifs ? (3 points)

## Bioinformatics: Problem 2. Evolutionary comparisons, Siepel et al., 2005.

### **Part 1**

Why is it useful to find evolutionary conserved sequences genome-wide ? (*2 points*)

### **Part 2**

Briefly describe the phylogenetic Hidden Markov model proposed by Siepel et al. for cross-species comparison. (*5 points*)

### **Part 3**

Identify a possible future extension of the phylo-HMM approach, and explain how you would approach to do such an extension. (*3 points*)

## Bioinformatics: Problem 3. Motif finding, Narlikar et al., 2006.

### **Part 1**

What is the motif finding problem ? What model of motifs is used in the Narlikar et al. paper ?  
(2 points)

### **Part 2**

What is meant by transcription factor structural class ? How can this information be used as an informative prior ? What method does this prior information feed into, as per the Narlikar et al. paper ? (5 points)

### **Part 3**

Can you propose any other type of information that may be used as a prior within the probabilistic framework of Narlikar et al. ? How would you go about implementing such a prior ? (3 points)