

DAIS Qualifying Examination
Spring 2007

Department of Computer Science
University of Illinois at Urbana-Champaign

February 19, 2007
Time Limit: 180 minutes

Please read the following instructions carefully before starting the test:

- The examination has 14 questions, including one basic concept question and 13 topic questions. You are required to answer the basic concept question and any **5** of the 13 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 **13** subquestions. You are required to answer **10** of these 13 subquestions. If you answer more than 10 subquestions, the committee will randomly select 10 to grade.
- The 13 topic questions are distributed by areas as follows:
 - **Database Systems:** 4 questions
 - **Data Mining:** 3 questions
 - **Information Retrieval:** 3 questions
 - **Bioinformatics:** 3 questions

Each of these 13 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 10 out of the following 13 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) (*transaction concurrency*) Serializability in transaction concurrency is a necessary condition for correctness. Do you agree? Explain why or why not.
- (2) (*relation algebra*) What are the basic relational operators? Show one “non-basic” operator and how it can be expressed in terms of the basic ones.
- (3) (*indexing*) All index structures are interchangeable (i.e., one can always be substituted by another). Do you agree? Explain why or why not.
- (4) (*data storage and representation*) What does “pointer swizzling” mean? Give one example situation where it is necessary to perform pointer swizzling.

Note: Please start with a new answer sheet.

- (5) (*data cube and OLAP*) An n -dimensional data cube, with the measure *count*, has p nonempty cells in its base cuboid (where $p > n > 1$). (1) What is the maximal number of (non-empty) 2-D cuboid in the cube? (2) What is the maximum possible number of (non-empty) aggregate cells in the cube?
- (6) (*selection of data mining algorithms*) Name one data mining algorithm that best fits each of the following applications: (1) detecting network intrusions with the availability of some previously recorded intrusion history, (2) finding hidden animal shapes in a multi-colored mosaic plate, and (3) finding when *wine* is selling well in the shop, so is *cheese*.
- (7) (*data mining applications*) Give one good application example for each of the following algorithms: (1) KNN (i.e., K -nearest neighbor), (2) EM, and (3) SVM.

Note: Please start with a new answer sheet.

- (8) (*TF-IDF weighting*) According to TF-IDF weighting, what kind of terms would have high weights?
- (9) (*Query-likelihood retrieval model*) Let $q = q_1 \dots q_m$ be a query and d be a document. Does the following formula correctly describe the query likelihood retrieval model? If not, write down the correct formula.

$$p(q|d) = \sum_{i=1}^m p(q_i|d)$$

where $p(q_i|d)$ is the probability of query word q_i according to a smoothed document language model estimated based on d .

- (10) (*Inverted index*) Would removing a few words with *low* IDF values from an inverted index help reduce the size of the index substantially? Why?

Note: Please start with a new answer sheet.

- (11) (*alignment*) What is the difference between the Needleman-Wunsch and the Smith-Waterman algorithms for pairwise alignment ?
- (12) (*Hidden Markov models*) What are silent (null) states in a Hidden Markov model ? Give an example of how such states can be useful in the application of HMMs to pairwise sequence alignment.
- (13) (*Phylogeny*) What is the maximum parsimony principle for evolutionary tree building ?

Database Systems: Problem 1. Extensible Indexing

Part 1

What's the role of *index* in a database system? How is GiST different from other index structures? (2 points)

Part 2

Suppose we want to index a data type *array*: An array instance A is specified by a sequence of integers: e.g., $A = (1, 2, 10, -5, 0)$.

- (a) Give one example query that is interesting to ask about this data type. (1 points)
- (b) Design a GiST to support such queries. You only need to concisely describe the design of the *key* predicate and the *consistent* method. (3 points)

Part 3

GiST aims at being a *general* indexing structure. It is, however, *not* the first extensible indexing ever proposed.

We ask you to describe *one* other idea of supporting indexing for *new* data types, and how it is different from GiST. (2 points)

Some may argue that: while potentially more general, GiST may not be the choice for practical settings. Do you agree. Explain why or why not. (2 points)

Database Systems: Problem 2. Ranking in RDBMS

Part 1

Why is *ranking* studied for relational DBMS? Give two reasons. (*2 points*).

Part 2

The *RankSQL* work aims at extending the relational algebra to be rank aware. What does it mean to be *rank aware*? (*1 points*)

To be concrete, please identify– not in detail but at a high level– the major extensions to relational algebra made by RankSQL, in order to achieve this rank awareness. (*2 points*).

Part 3

In contrast to its recent development in *databases*, ranking has been a major query paradigm, for several decades, for *text retrieval*, i.e., finding relevant documents from a text collection. We ask you to compare how the issue of ranking is treated in the two areas:

Make three observations that contrast how ranking is treated in *significantly different* ways in the two areas, and explain why. (*5 points*)

Hints: Your observations can be on any aspects of supporting ranking– e.g.: Its purposes? Semantics? Processing techniques? Research issues and focuses? We will consider any interesting points you make.

Database Systems: Problem 3. Physical Database Design

Part 1

In the "AutoAdmin What-if" paper, what is *hypothetical configuration simulation*? Define it in terms of its input, output, and purpose. (2 points)

Part 2

The concept of hypothetical configuration analysis appears to be similar to *cost estimation* in standard cost-based query optimization.

- What are their similarities? Give one point. (1 points)
- What are their differences? Give two points. (2 points)

Part 3

The system as reported in this paper accepts a *workload* as a *set* of (query, frequency) tuples, with the command of the form:

```
CREATE WORKLOAD workloadname AS (Q1, f1), (Q2, f2), ..., (Qn, fn)
```

This simple workload definition may not fully capture real world situations and requirements. We ask you to identify such limitations and propose solutions.

- Identify one limitation: i.e., give a practical situation, in which the above workload specification may not be general enough to capture, and explain what issues may arise. (2 points)
- Propose your strategy to solve the limitation: i.e., specify how the workload specification can be extended, and sketch how the simulation process can be generalized accordingly (or explain why such generalization might be impossible). (3 points)

Database Systems: Problem 4. Information Integration

Part 1

Web data extraction studies the problem of extracting data records from Web pages. Why is this problem important for information integration? (2 points)

Part 2

The problem of Web data extraction can be abstracted as: Given a set of input Web pages, let $x = \{x_0, x_1, \dots, x_n\}$ be the data “tokens” (or blocks) on these pages, which are to be labeled; let $y = \{y_0, y_1, \dots, y_n\}$ be one possible label assignment of the corresponding tokens. The goal of data extraction is to find a best assignment y^* , such that some objective function $G(x, y)$ is maximized.

- What is typically such an objective function G ? (1 points)
- In a particular data extraction setting, how exactly to calculate G ? That is, given input x , for some possible assignment y , how to find the value of $G(x, y)$? (2 points)

Part 3

The paper on simultaneous record extraction and attribute labeling attempts to remove the “First-Order Markov” assumption.

- Explain what this assumption is. (1 points)
- Give an example feature function that violates this assumption. (2 point)
- Removing this assumption, this paper claims to support more general feature functions by its HCRF model. Is it now able to support *arbitrary* feature functions? Explain why or why not. (2 points)

Data Mining and Data Warehousing: Problem 1

Part 1

Data cube computation algorithms have been developed for efficient computation of multidimensional aggregations for low dimensional data (such as less than 10 dimensions).

- (a) Explain why a typical cubing algorithm, such as BUC or StarCubing, encounters difficulties at computing high-dimensional data cubes, such as 50-100 dimensions. (*1 point*)
- (b) Outline an efficient algorithm that can perform fast OLAP operations in such a high-dimensional space with not so huge data sets (such as 10^5 tuples). (*2 points*)

Part 2

Data warehouse has been used in many applications. Suppose one wants to trace the shipping of goods for a shipping company such as UPS so that not only customers but also company managers can check the situations or summaries, by region, by goods category, by month, etc., and drill down to any particular region, day, etc.

- (a) Design such a data warehouse by drawing star schema with concise explanation. (*1 point*)
- (b) Show how to efficient compute data cubes and perform OLAP operations in such a data warehouse. (*2 points*)

Part 3

Although data cube models can be used for computing multidimensional aggregations for numerical values, the multidimensional model can be used for modeling and prediction as well.

- (a) Explain how to extend the multidimensional data cube model for prediction analysis. (*2 points*)
- (b) Design an “outlier cube” that facilitates outlier analysis in multidimensional space. (*2 points*)

Data Mining and Data Warehousing: Problem 2

Part 1

- (a) Give an example to show that two strongly associated itemsets, X and Y , may not be strongly correlated. (1 point)
- (b) Give another example to show that the commonly used *lift* measure, $lift(X, Y) = \frac{P(X, Y)}{P(X)P(Y)}$, may not be a good measure for correlation analysis in transactional databases. (1 point)
- (c) Propose a better measure for it. (1 point)

Part 2

It is often useful to mine frequent itemsets in data streams.

- (a) Explain why lossy counting algorithm proposed by Manku and Motwani (2002) guarantees an error bound at mining approximate frequent itemsets in data streams. (1 point)
- (c) Outline a method that may find approximate correlation patterns in data streams. (2 points)

Part 3

Frequent-patterns have been used for effective classification.

- (a) Explain why discriminative frequent patterns may achieve better classification accuracy than (1) the complete set of frequent patterns, and (2) many other typical classification methods. (2 points)
- (b) The method proposed by Hong et al. (2007) for pattern-based classification is to first mine the full set of closed frequent patterns and then extract a set of discriminative frequent patterns for effective classification. Propose a more efficient method that mines a similar set of patterns directly. (2 points)

Data Mining and Data Warehousing: Problem 3

Part 1

- (a) Explain why a simple hierarchical clustering algorithm, such as AGNES, often generates low quality clusters. (*1 point*)
- (b) Explain why two recently proposed hierarchical algorithms, BIRCH and CHAMELEON, generate high quality clusters. (*1 point*)

Part 2

- (a) What are the major challenges in high-dimensional clustering? (*1 point*)
- (b) Outline a method that efficiently clusters micro-array data sets that contain high-dimensional numerical data. (*2 points*)
- (c) Propose a method that efficiently clusters text documents based on the sets of terms contained in those documents. (*2 point*)

Part 3

Most objects are linked to each other via various kinds of semantic links (or relationships).

- (a) Explain how such links can be used for effective cluster analysis. (*1 point*)
- (b) Outline a scalable method that performs effective clustering using multiple kinds of links. (*2 points*)

Information Retrieval: Problem 1. Expert Finding

The following questions about expert finding are all based on the paper [Balog et al. SIGIR 06].

Part 1

- a How is expert finding (as defined in [Balog et al. SIGIR 06]) different from regular document retrieval? (*1 point*)
- b Given a topic query q , the authors first proposed to rank candidate expert ca based on $p(ca|q)$, but the proposed language models are actually used to estimate $p(q|ca)$. How are these two conditional probabilities related? Under what conditions, ranking based on $p(q|ca)$ would be equivalent to ranking based on $p(ca|q)$? (*2 points*)

Part 2

- a The authors use $a(d, ca)$ to denote “document-candidate association.” How is $a(d, ca)$ computed? Among all the variations of methods for computing $a(d, ca)$, which method performs the best according to their experiment results? (*2 points*)
- b In the model 2 proposed in the paper, $p(q|ca)$ is computed based on $p(d|ca)$ and $p(q|d)$ where d is a document. Let D be all the documents to be considered. Write down the formula for computing $p(q|ca)$ in terms of $p(d|ca)$ and $p(q|d)$. Briefly explain why the formula intuitively makes sense. (*2 points*)

Part 3

Given the same task setup for expert finding as defined in the paper (i.e., given a list of candidate experts with their names and emails, a set of documents where the candidates may be mentioned, rank candidates for a given topic query). Can you propose a method for expert finding based on the vector-space model? (*3 points*)

Information Retrieval: Problem 2. Query refinement

Part 1

- a Four different types of query substitutions are defined in [Jones et al. WWW 06]. Name two of the four types and give an example for each of the two. (*2 points*)
- b Name three specific types of term relationships that are exploited in [Collins-Thompson Callan CIKM 05] for query expansion. (*1 point*)

Part 2

- a In [Jones et al. WWW 06], a query is segmented into phrases. For example, “new york maps” becomes “(new york) (maps)”. What method has been proposed to do this and how are these phrases discovered? (*2 points*)
- b In [Jones et al. WWW 06], when generating related queries, what method has been used to generate related queries for *infrequent* queries? (*1 points*)
- c Have the authors of the paper [Jones et al. WWW 06] shown that the generated query substitutions using their methods are effective for improving retrieval performance? If yes, explain how; if not, can you propose a way to do that? (*1 points*)

Part 3

In [Jones et al. WWW 06], the authors identified several types of poor suggestions made by their methods. Name one of them and propose one idea to avoid this type of poor suggestions. (*3 points*)

Information Retrieval: Problem 3. Similarity measure

Part 1

- a Compared with computing the similarity between two regular documents (e.g., web pages), why is measuring the similarity between two short text snippets more challenging? (*1 point*)
- b Give two different applications where it is necessary to measure the similarity between short text snippets. (*2 points*)

Part 2

- a In [Sahami & Heilman WWW 06], a new similarity measure for short text snippets is proposed. Sketch the basic idea of this approach. (*2 points*)
- b In [Sahami & Heilman WWW 06], the authors identified an *undesirable* property of the baseline Set Overlap similarity measure, which the proposed new method does not have. What is this property? (*1 point*)

Part 3

- a The new similarity measure proposed in [Sahami & Heilman WWW 06] involves a parameter n (the number of documents to retrieve). How should we set this parameter? Why? (*2 points*)
- b Identify at least one deficiency of the new similarity measure proposed in [Sahami & Heilman WWW 06] and propose a way to improve it. (*2 points*)

Bioinformatics: Problem 1. Motif finding: Lawrence *et al.* 1993.

Part 1

How is motif finding, as addressed in the paper of Lawrence *et al.*, related to multiple sequence alignment ? Name the motif model used in this motif-finding algorithm ? (2 points)

Part 2

(i) In the Lawrence *et al.* approach to motif-finding, what is the objective function ? That is, what is the score of a candidate solution (motif) which the algorithm aims to optimize ? (2 points)

(ii) Give a brief description of the algorithm used by this method for finding the highest scoring motifs. (3 points)

Part 3

How would you approach extending the algorithm of the Lawrence *et al.* paper to the discovery of multiple motifs (in the same set of sequences) simultaneously ? Assume that the number of motifs to be discovered is fixed in advance. (3 points)

Bioinformatics: Problem 2. Gene expression: Segal *et al.* 2003.

Part 1

Why should integration of gene expression and non-coding sequence data provide a more powerful analysis than either of these data types separately ? (1 points)

Part 2

In the Segal *et al.* paper, the probability distribution of a gene g 's expression vector $g.\mathbf{E}$, given that the gene belongs to a transcriptional module M , is modeled explicitly. Provide an expression for this probabilistic model. That is,

$$\Pr(g.E_1, g.E_2, \dots, g.E_J | M) = ?$$

(2 points)

For each gene g , the paper defines a set of binary-valued *Regulates* variables $g.\mathbf{R} = \{g.R_1, \dots, g.R_L\}$, where $g.R_i$ takes the value *true* if motif i appears in the promoter region of gene g , and false otherwise. Assuming that motif i is a position-specific scoring matrix (PSSM), write down an expression for the probability of the variable $g.R_i$ being true, given the sequence of the gene's promoter $g.S$. That is,

$$\Pr(g.R_i = \text{true} | g.S) = ?$$

(2 points)

Under the Segal *et al.* model, the *motif profile* of a transcriptional module m is defined to be a set of weights u_{mi} , one for each motif i , such that u_{mi} specifies the extent to which motif i plays a regulatory role in module m . The strength of association of a gene g with a module m then depends on the $g.R_i$ variables and the u_{mi} weights (for all motifs i). Write down the explicit model used by the authors to prescribe this association. That is,

$$\Pr(g.M = m | g.R_1 = r_1, \dots, g.R_L = r_L) = ?$$

(2 points)

Part 3

Can you think of any other types of data, apart from sequence or microarray gene expression data, that could be integrated with one or both of these data types in a similar probabilistic framework ? Briefly mention how you might model the probability of observing such data, for a particular gene g , given that it belongs to the transcriptional module M ? (3 points)

Bioinformatics: Problem 3. CRM prediction: Blanchette *et al.* 2006.

Part 1

What is a CRM ? What is the basic premise behind CRM prediction in the genome-wide approach of Blanchette et al. ? (2 points)

Part 2

For any motif m , and any aligned position p of the genome, Blanchette et al. define $\text{hitscore}_{\text{aln}}(m,p)$ as a combination of the $\text{hitscore}_S(m,p)$ values over each of three species S . The total score of a putative CRM spanning the region from position p_1 to position p_2 , denoted as $\text{TotalScore}(m, p_1, p_2)$, is then defined in terms of the $\text{hitscore}_{\text{aln}}(m, p)$ values for all positions $p \in \{p_1, p_2\}$. State how this TotalScore is defined by the authors. (2 points)

The TotalScore values for each motif m are then used to compute the score of a candidate CRM (p_1, p_2), based on one to five PWMs called “tags”. Describe how these tags are decided and how the score of the CRM is evaluated based on these tags. (3 points)

Part 3

How is the TotalScore computed by the authors ? How would you modify their approach to obtain an accurate computation of the TotalScore ? Why do you think they did not adopt the approach you propose ? (3 points)