

DAIS Qualifying Examination  
*Spring 2009*

Department of Computer Science  
University of Illinois at Urbana-Champaign

Feb 23, 2009

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) (*Concurrency Control*) Give an example of a nonserializable schedule.
- (2) (*Recovery*) What kind of recovery do commercial relational DBMSs use: redo, undo, or undo/redo?
- (3) (*Checkpointing*) Commercial relational DBMSs provide checkpoint facilities. What is the **main** advantage of periodically taking a checkpoint? Explain in one sentence.

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

- (4) (Measures in similarity computation) It is good to select appropriate measures in computing similarity or distances among objects. Give the most appropriate measure for the following computation: (1) correlation between items in a transactional dataset, (2) finding clusters in a large connected graph, and (3) finding similar conferences in the DBLP network.
- (5) (Data mining methods) Use one sentence to distinguish the following pairs of methods: (1) EM vs.  $k$ -means, (2) SPADE vs. PrefixSpan, (3) Hidden Markov Model (HMM) vs. Bayesian Believe Networks (BBN).
- (6) (Selection of clustering algorithms) Name or outline one clustering algorithm that best fits each of the following applications: (1) user-guided clustering across multiple relations, (2) clustering evolving, dynamic data streams, and (3) clustering microarray data with over 10000 dimensions.

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

- (7) (*Retrieval Evaluation*) When comparing the overall ranking accuracy of two retrieval models, is it better to use the Mean Average Precision (MAP) or Precision at  $K$  documents (e.g.,  $K=10$ ), or both? Briefly explain why?
- (8) (*Retrieval Model*) In retrieval with the query likelihood scoring function with Dirichlet prior smoothing, we score a document  $D$  with respect to query  $Q$  based on the likelihood  $p(Q|D)$ . If we add an additional term to query  $Q$  to obtain a new query  $Q'$ , can we tell for sure whether this would increase or decrease the score of document  $D$ ? That is, how is  $p(Q'|D)$  compared with  $p(Q|D)$ ? Which is larger?

- (9) (*Retrieval Techniques*) Use two or three sentences to explain what is Rocchio.

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

- (10) (*Alignment*) Maximum parsimony is one (the most popular) approach to biological sequence alignment, and Smith-Waterman or Needleman-Wunsch algorithms belong to this paradigm. Name one other broad class of approaches used in sequence alignment.
- (11) (*Markov chains*) A Markov chain is often used to model background sequences in motif finding methods. Given a DNA sequence to be used as “background”, and the order  $k$  of the Markov chain to be trained, what is the maximum likelihood approach to learning the parameters (transition probabilities) of the Markov chain. (You only need to write down the appropriate

formula, without deriving it.) *Hint: this is the standard method of learning Markov chain parameters.*

(12) (*Phylogeny*) What is an “outgroup” species, in the context of phylogenetic analysis?

# Data Mining and Data Warehousing: Problem 1

## Part 1

Data cube has become an essential information component. Reason on the correctness or incorrectness of each of the following statements.

- (a) Computing iceberg cube (with iceberg condition  $count \geq 2$ ) costs less than computing the full data cube for the following algorithms: (i) Multi-Way Array Cubing, (ii) BUC, and (iii) StarCubing. (1.5 point)
- (b) If one constructs a datacube on sampling data (e.g., survey data), one cannot support quality drill-down because some low-level cells could contain empty or too few data for reliable analysis. (1.5 point)

## Part 2

It is desirable to construct an AlbumCube to facilitate multidimensional search through digital photo collections, such as by date, photographer, location, theme, content, color, etc.

- (a) What should be the dimensions and measures for such a data cube? (1 point)
- (b) What analytical functions you can provide, and (1 points)
- (c) What are the major challenges on implementing AlbumCube, and how would you propose to handle them? (1.5 points)

## Part 3

One may expect to perform multidimensional analysis of traffic data on highways based on the traffic history to identify what segments are likely jammed at what conditions, e.g., during rush hours or bad weather.

- (a) What should be the dimensions and measures for such a data cube? (1 point)
- (b) What analytical functions you can provide, and (1 points)
- (c) What are the major challenges on implementing TrafficCube, and how would you propose to handle them? (1.5 points)

## Data Mining and Data Warehousing: Problem 2

### Part 1

- (a) Suppose a dataset contains only 5 transactions as follows:

$\{a, b, c, d\}$ : 2

$\{a, c, e\}$ : 3

Let the minimum support be 3. Write out (together with the support information) all the (i) frequent patterns, (ii) closed patterns, and (iii) max-patterns. (1.5 point)

- (b) Present one example to show *cosine* is a better measure than *lift* at disclosing interesting patterns in large transaction databases. (1.5 point)

### Part 2

- (a) What is the worse-case computational complexity of mining sequential patterns? Why do many good sequential pattern mining algorithm claim that it can usually find the complete set of sequential patterns efficiently under a reasonable *min\_support* threshold? (1 point)
- (b) Show max-gap constraint is antimonotonic for sequential pattern mining. Discuss how to refine a typical sequential pattern mining algorithm, such as PrefixSpan, to push “*max-gap* = 5” deep into the mining process. (1 point)
- (b) Why that a typical sequential pattern mining algorithm encounters difficulty at finding rather long patterns (such as of size 100)? Can you propose a method that may find such patterns efficiently? (2 point)

### Part 3

- (a) Explain why discriminative frequent patterns may lead to better better classification quality than a typical associative classifier. (1 point)
- (c) DDPMine is an efficient discriminative, frequent pattern-based classification method. However, it does not perform well with high-dimensional data. Explain why. Can you propose a discriminative, frequent pattern-based classification method that is likely to perform well with high-dimensional data? (2 points)

## Data Mining and Data Warehousing: Problem 3

### Part 1

- (a) Explain why RAINFOREST is more scalable than C4.5 in decision tree induction in large datasets. (*1.5 point*)
- (b) Explain why SVM performs well on high-dimensional data but does not perform well on large volume of data. (*1.5 point*)

### Part 2

- (a) What are the major challenges of classifying dynamic evolving data streams? (*1 point*)
- (b) What are the major challenges of classifying highly skewed data sets? (*1 point*)
- (c) Outline a method that effectively classifies highly skewed data sets in dynamic evolving data streams. (*1 point*)

### Part 3

- (a) Take two typical classification methods,  $k$ -nearest neighbor (KNN) and naïve Bayes as examples, explain what are the major differences between lazy classification and eager classification methods. (*1 point*)
- (b) Extend the KNN classification method to make it work well in data stream environment. (*1.5 point*)
- (c) Many classification tasks assume all of the training data are labeled. In reality, only a small set of training data are (positively) labeled (e.g., a small set of webpages are given class labels). Extend the KNN classification method to make it effective for partially labeled data. (*1.5 points*)

## **Information Retrieval: Problem 1. Retrieval Models**

[Xu & Akella 08] refers to the following paper:

Z. Xu and R. Akella, A New Probabilistic Retrieval Model Based on the Dirichlet Compound Multinomial Distribution, Proceedings of ACM SIGIR 2008.

### **Part 1**

- a Briefly explain what is Maximum Likelihood estimation. If we estimate a multinomial document language model  $p(w|\theta_D)$  for document  $D$  using the Maximum Likelihood estimator, what would be the probability of word  $w$  according to the estimated model? (*1 point*)
- b The paper [Xu & Akella 08] addresses a deficiency of using multinomial distribution in a probabilistic retrieval model. What is this deficiency? (*1 point*)
- c [Xu & Akella 08] proposed to use the Dirichlet Compound Multinomial (DCM) model to replace the multinomial distribution in a probabilistic retrieval model. Does the DCM model have more or fewer parameters to estimate than the multinomial distribution? (*1 point*)

### **Part 2**

- a In [Xu & Akella 08], the authors use  $\beta_R$  and  $\beta_N$  to denote the parameters of the DCM distribution for relevant documents and non-relevant documents, respectively. What text information have the authors used to estimate  $\beta_R$  and  $\beta_N$ ? (*2 points*)
- b Once  $\beta_R$  and  $\beta_N$  are estimated, how did the authors compute the score of a document w.r.t. a query? Sketch the scoring formula. (No need to write any detail about the DCM model.) (*1 points*)
- c In [Xu & Akella 08], the authors also presented some method for pseudo feedback. Use no more than three sentences to briefly explain how the proposed feedback method works. (*1 point*)

### **Part 3**

Suppose we would like to extend the probabilistic retrieval model proposed in [Xu & Akella 08] to support cross-language retrieval, where we would like to use a query in French to retrieve documents in English. What DCM distributions do we need to estimate? How do we score a document according to the method proposed in [Xu & Akella 08]? Can you suggest a different way to score a document using DCM distributions? (*3 points*)

## Information Retrieval: Problem 2. Forum Crawling

[Wang et al. 08] refers to the following paper:

Y. Wang et al., Exploring traversal strategy for web forum crawling, Proceedings of ACM SIGIR 2008.

### **Part 1**

- a According to [Wang et al. 08], what is a skeleton link and what is a page-flipping link in a forum? (1 point)
- b The crawling system proposed in [Wang et al. 08] consists of two parts corresponding to the two steps to be followed. What are the two steps? (1 point)

### **Part 2**

- a How did the authors of [Wang et al. 08] compare two crawlers in terms of effectiveness and efficiency? (2 points)
- b In the algorithm for finding skeleton links proposed in [Wang et al. 08], the function *GetSkeletonLinks* takes as input a set of candidate links  $\{L_1, \dots, L_m\}$ . If without pruning, in order to generate the desired output, what is the maximum size of the search space that *GetSkeletonLinks* has to consider? Why? (2 points)

### **Part 3**

- a In [Wang et al. 08], the *informativeness* of a vertex on the sitemap is defined as follows:

$$Info = -\frac{1}{\log(N)} \sum_{i=1}^K \frac{\|D_i\|}{N} \log \frac{\|D_i\|}{N}$$

- . What role does the term  $\frac{1}{\log(N)}$  play in the formula? Is there any concern if we drop this term? (2 points)
- b Suppose we want to only crawl pages in a forum that are relevant to a particular topic. What changes can you make to the algorithm for finding skeleton links proposed in [Wang et al. 08] to make it more suitable for such focused crawling of a forum?  
(2 points)

## **Information Retrieval: Problem 3. Information Extraction**

[Cong et al. 08] refers to the following paper:

G. Cong et al., Finding question-answering pairs from online forums, Proceedings of ACM SIGIR 2008

### **Part 1**

- a In [Cong et al. 08], the authors used a graph propagation algorithm. What are the nodes in the graph? What does an edge mean in the graph? (*1 point*)
- b List two specific reasons why extracting question-answering pairs from online forums is challenging. (*1 point*)
- c The authors of [Cong et al. 08] decomposed the task of extracting question-answering pairs into two steps. What does each step do? (*1 point*)

### **Part 2**

- a According to [Cong et al. 08], on what kind of stream data is the sequential pattern mining algorithm applied? Can you give an imaginary example pattern that might be discovered? (*1 point*)
- b The graph propagation algorithm proposed in [Cong et al. 08] works in a similar way to the PageRank algorithm. Besides the difference in the graph involved, can you point out any other differences between the propagation algorithm in this paper and the propagation algorithm of PageRank? (*2 points*)

### **Part 3**

- a The authors of [Cong et al. 08] compared their method for question detection with two simple baseline methods (i.e., question-mark and 5W1H words), and showed that their method is better. This is an unfair comparison in that their method has used some additional information. What is this “additional information” used by their method that has made the comparison unfair? Can you suggest a stronger baseline to be compared so that the comparison would be more fair? (*2 points*)
- b Briefly explain intuitively why we would expect the graph propagation algorithm proposed in [Cong et al. 08] to be beneficial. Can you think of any other retrieval technique to achieve a similar effect? (*2 points*)

## Bioinformatics: Problem 1. Segal et al, 2008.

### Part 1

Which two physical/biochemical factors are most important for modeling gene expression from regulatory sequence? The regulatory sequences analyzed in Segal et al. (2008) are related to which important biological process in *Drosophila*? (2 points)

### Part 2

Briefly describe how the authors calculate the probability of a configuration of transcription factor molecules occupying binding sites, i.e.,  $P(C)$  as per their notation. In what form does the position-weight-matrix (PWM) of a transcription factor figure in this probability? (3 points)

What prevents the authors from being able to calculate the probability of expression,  $P(E)$  as per their notation, efficiently? (2 points)

### Part 3

ChIP-on-chip data measures the binding affinity of a transcription factor to different regions of a genome. You have the position-weight-matrix of a transcription factor, and you also have ChIP-on-chip data for that factor. Can you suggest a way to model the ChIP data, borrowing ideas from Segal et al.? In other words, suggest a way to predict the overall binding affinity of a probe sequence, as measured in ChIP experiments, using the entire sequence of that probe. (3 points)

## **Bioinformatics: Problem 2. Siepel and Haussler, 2003.**

### **Part 1**

What is the most novel aspect (modeled by an HMM) in Siepel & Haussler's approach to evolutionary analysis of genomic sequences? (*2 points*)

### **Part 2**

(a) How does the PhyloHMM of Siepel & Haussler allow for different rates at different positions? (*2.5 points*)

(b) Explain the role of the "autocorrelation parameter  $\lambda$ " in the PhyloHMM. (Also provide the mathematical expression that uses  $\lambda$  in setting up the model.) (*2.5 points*)

### **Part 3**

What, in your opinion, is the weakest link in the PhyloHMM framework? Is it the fact that fixed alignments are used, or the way different evolutionary rates are modeled, or the substitution models themselves? What kinds of evolutionary divergences would be most problematic for the framework (due to this weakness)? Suggest how you would go about fixing the weakness you identified above. (*3 points*)

## **Bioinformatics: Problem 3. Narlikar et al., 2006.**

### **Part 1**

In the title of the Narlikar et al. (2006) paper, what does the term “informative priors” refer to? In other words, explain what a prior is in general, and what kind of information is used as “prior” in the paper? (*2 points*)

### **Part 2**

How do the authors use binary classification algorithms to obtain a “prior probability” to be used in motif finding? In other words, explain how the output of the classifiers is used to compute a prior probability on the variable  $Z_i$ . (*4 points*)

### **Part 3**

Before you set about including some type of information as “prior” in motif finding, what statistical test do you need to perform to see if that kind of information may be useful? Suppose you have a function that returns a quantitative measure of the DNA’s accessibility (i.e., uncompact state) at each position. Propose a scheme to include such information as prior for motif finding. (*4 points*)