

DAIS Qualifying Examination
Fall 2008

Department of Computer Science
University of Illinois at Urbana-Champaign

Sep 22, 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) (*Normalization*)

Consider relation $Course(Number, Name, Area, Faculty)$ where each course is taught by just one faculty member all the time, each course has a unique number, and no two different courses have the same combination of name and area. Provide a 3NF version of this relation and determine if your answer is in BCNF and why it is or it is not in BCNF.

(2) (*Query Optimization*)

To optimize a query whose selection clause mentions a view, you'll get the best result if you optimize the view definition and the query separately, and then paste the two of them together. Briefly explain your answer.

(3) (*Access Methods*)

Why is B-tree important for realizing the relational model of DBMS? Explain in three sentences.

Note: Please start with a new answer sheet.

(4) (data preprocessing) Name **four** typical methods that are effective in *dimensionality reduction*. And name **four** typical methods that are effective in *numerosity reduction* (i.e., reducing the large amount of data to a smaller amount).

(5) (data mining methods) Use one sentence to distinguish the following pairs of methods: (1) *PrefixSpan* vs. *gSpan* algorithms, (2) *k-means* vs. *k-medoids* algorithms, (3) *decision tree induction* vs. *rule induction*.

(6) (selection of data mining methods) Name or outline one data mining method that best fits each of the following tasks: (1) *finding snakes in bushes*, (2) *finding computer network anomalies in real time*, and (3) *determining whether a (muscle) tissue is cancerous*.

Note: Please start with a new answer sheet.

(7) (*Vector Space Model*) Name two differences between the TF-IDF formula used in modern retrieval functions such as Okapi and the naive raw TF-IDF formula $\sum_t f_t \log \frac{N+1}{n_t}$ where the sum is over the matched query terms in a document, f_t is the raw frequency of term t , N is the total number of documents in the collection, and n_t is the number of documents containing term t .

(8) (*Language Model*) In retrieval with the query likelihood scoring function, should a long document be smoothed more or a short one? Why?

(9) (*PageRank*) Knowing that a search engine is using PageRank scores in its ranking function, one can take advantage of this knowledge to potentially spam the search engine by promoting a particular page through links. Briefly explain what the spammer can do to increase a page's PageRank score.

Note: Please start with a new answer sheet.

- (10) (*Alignment*) “For the local alignment algorithm to work, the expected score for a random match must be negative.” Why ?
- (11) (*Hidden Markov models*) What does the Viterbi algorithm for an HMM maximize ? How does the running time of this algorithm scale with the number of states in the HMM? (Assume first order transitions.)
- (12) (*Phylogeny*) What does the term “reversibility” mean in the context of evolutionary models?

Database Systems: Problem 1. Pay-as-You-Go Information Integration

The paper “iTrails: Pay-as-you-go” advocates the concept of “Pay-as-You-Go” information integration.

Part 1

- a) What does “Pay-as-you-go” mean? Define it in no more than three sentences. (1 points)
- b) Give an example integration scenario that is *not* pay-as-you-go. (2 points)

Part 2

For building an information integration system, there are several problems (e.g., about data, schema, and queries) that need to be addressed, such as:

- *Source Discovery*: Finding sources to integrate in the system.
- *Wrapper Construction*: Building a wrapper for each source, to interact with the source in a uniform way, including executing queries and extracting result data.
- *Schema Matching*: Determining the correspondence of attributes among sources.
- *Source Selection*: Determining, for a user query, which sources are to be queried.
- *Query Translation*: Mapping a user query into queries for the selected data sources.
- *Data Transformation*: Mapping data from each source to the integrated format of the system.

What problems among the above are addressed by this paper? For each problem that is addressed, briefly explain the key ideas of the solution. (3 points)

Part 3

Considering the integration problems that must be addressed, as outlined in Part 2, since this paper does not address all of them, we ask your opinions on the practical deployment of the proposed techniques:

- a) Identify and describe a complete information integration scenario– the data sources and the objective of integration. Describe how the techniques can be applied. (1 points)
- b) For your specific scenario, do you think the techniques realize pay-as-you-go in a “practical” way? Explain why or why not. (3 points).

Database Systems: Problem 2. Approximate Query Processing

The “Scalable Approximate Query Processing” paper proposes a framework for approximate query processing.

Part 1

- a) Why is approximate query processing important for *aggregate queries*? (1 points)
- b) What are the desired properties to achieve for such online approximate query processing? (2 points)

Part 2

- a) Give an example query that can be processed by the proposed techniques. (1 points)
- b) Using this example, explain how the techniques work, and identify the key components of the solutions. Be concrete and concise. You only need to explain, at a high level, the framework of processing and the major steps. (3 points)

Part 3

The core of query processing is in the handling of joins. The paper develops *levelwise* steps, where each step consists of a *scan* and a *merge* phase for concurrently processing a set of joins.

Do you think the techniques can apply to any arbitrary *join conditions*? Explain why or why not. (3 points)

Database Systems: Problem 3. Scalable Semantic Web Data Management

Part 1

- a) What is RDF? (*1 points*)
- b) Why does the RDF data model incur inefficient query performance? (*1 points*)

Part 2

The paper “Scalable Semantic Web Data Management” proposes vertical partitioning for query processing of RDF data. We ask you to illustrate the ideas: Create an example RDF dataset and an example query, and briefly explain how the mechanism works. (*4 points*)

Part 3

We wonder how the *RDF model* for the Semantic Web is different from the *relational model* used in RDBMS. Identify *two* major differences. For each difference, suggest (even if you have to guess) why the Semantic Web community made the particular choice. (*4 points*)

Data Mining and Data Warehousing: Problem 1

Part 1

Suppose the base cuboid of a data cube contains only two cells

$$(a_1, a_2, a_3, \dots, a_{10}), (b_1, b_2, b_3, \dots, b_{10}),$$

where $a_i = b_i$ if i is an odd number; otherwise $a_i \neq b_i$.

- (a) How many nonempty cuboids are there in this data cube? (*1 point*)
- (b) How many nonempty aggregate cells are there in this data cube? (*1.5 point*)
- (c) If we set minimum support = 2, how many nonempty aggregate cells are there in the corresponding iceberg cube? (*1 point*)

Part 2

Multidimensional analysis is essential for studying many data sets. We want to design and build a “data cube” for multidimensional analysis of DBLP (digital bibliographic library project for computer science publications) database for powerful and flexible analysis. Notice that DBLP contains entries for almost every recognized CS research conference or journal publication entries, with the information like authors, paper title, publication venue, location, and time.

- (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*)
- (b) What are the major challenges in implementation and how would you propose to handle them? (*1.5 points*)

Part 3

Most data cube build on the whole population of data. However, in many cases the collected data are just samples (such as surveys). A major problem for such data is when drilling down, some cells could contains empty or few data for reliable analysis.

- (a) Design a sampling cube such that reliable prediction can still be done despite some cells contains very few or empty data values, (*1.5 point*)
- (c) Discuss how such a cube can handle high-dimensional OLAP. (*1.5 points*)

Data Mining and Data Warehousing: Problem 2

Part 1

- (a) List five major challenges at clustering different kinds of data. (*1.5 point*)
- (b) Outline an efficient algorithm that can effectively cluster (high-dimensional) micro-array data. (*1.5 point*)

Part 2

- (a) Links contain rich information for effective data mining. Take DBLP data as an example, illustrate how links can be used for effective clustering. (*1 points*)
- (b) Is the method you described efficient in large databases? Outline a method that is scalable and efficient in cluster analysis for large sets of linked data. (*1.5 points*)
- (c) In real datasets, data may not be clean (e.g., same person may have different names and different people may share the same name). Outline a method so that link-based clustering can still work well with the existence of such data. (*1.5 points*)

Part 3

- (a) Information networks have become an important target in data mining. Information networks themselves may often need to be clustered. Give an example and convincing argument that such clustering may lead to the discovery of interesting knowledge. (*1.5 point*)
- (b) Outline one effective method for clustering information networks (*1.5 point*)

Data Mining and Data Warehousing: Problem 3

Part 1

- (a) Frequent pattern mining has been studied extensively in data mining research. However, it is not easy to mine large patterns (such as pattern size ≥ 100) in large data sets. What are the difficulties at mining such large patterns? (*1.5 point*)
- (b) Can you design an effective method that mines frequent large patterns effectively in large datasets? (*1.5 point*)

Part 2

- (a) Frequent pattern methods are designed for mining precise patterns. However, the real world patterns are largely approximate. What are the major difficulties at mining approximate frequent patterns? (*1 point*)
- (b) Design an efficient algorithms that can mine approximate patterns efficiently. (*2 point*)

Part 3

- (a) Frequent patterns have been used in classification. What are the major problems in traditional frequent-pattern-based classification methods? (*1.5 point*)
- (b) Outline an efficient method that can perform efficient discriminative frequent pattern-base classification. (*1.5 point*)
- (c) Explain why the above describe method leads to high efficiency and high accuracy in large datasets (*1 points*)

Information Retrieval: Problem 1. Web Search

[Bao et al. WWW 07] refers to the following paper:

S. Bao and others, Optimizing Web Search Using Social Annotations, Proceedings of WWW 2007.

Part 1

- a What are the input and output of the SocialSimRank algorithm proposed in [Bao et al. WWW 07]? (1 point)
- b How are the SocialSimRank (SSR) score and SocialPageRank (SPR) score combined to rank web pages? (1 point)

Part 2

- a In [Bao et al. WWW 07], the authors have designed experiments to compare a baseline method with several other proposed methods. What is this baseline method? What other methods have been compared with this baseline method? Use no more than three sentences to describe the main findings made by the authors based on these experiment results. (2 points)
- b When describing the experiment procedure, the authors said “[We] ... created five different random splits of 40 training and 10 testing queries on MQ set, and 2,400 training and 600 testing queries on AQ set.” Why did they need the training data? Does the fact that the baseline method didn’t use the training data raise any concern about their experiment results and thus conclusions? (2 points)

Part 3

- a Suppose every page is annotated with a *disjoint* set of tags (i.e., annotations). What kind of output values would SocialSimRank give? Why? (2 point)
- b Can you propose a way to improve SocialSimRank so that it can better handle the case when every page is annotated with a disjoint set of tag? (2 points)

Information Retrieval: Problem 2. Retrieval Model

[Metzler & Croft 07] refers to the following paper:

D. Metzler, W. B. Croft, Latent Concept Expansion Using Markov Random Fields, Proceedings of ACM SIGIR 2007.

Part 1

The Markov Random Field (MRF) retrieval model would rank a document D w.r.t. query Q using the following formula:

$$P_{G,\Lambda}(Q, D) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \phi(c; \Lambda)$$

- What is $\phi(c; \Lambda)$ called? Assuming that $f(c)$ is a feature value defined on c , give a commonly used definition of $\phi(c; \Lambda)$ using an exponential function. (1 point)
- Compared with the vector space model or query-likelihood language model, the MRF model is conceptually more general. Why? (1 point)
- In [Metzler & Croft 07], how did the authors set the parameters of the MRF? (1 point)

Part 2

- The latent concept expansion approach proposed in [Metzler & Croft 07] involves the use of a new graph G' in the MRF model to achieve an effect of query expansion. How is this new graph constructed? If we expand a query with *three new two-term concepts*, how many more nodes and how many more edges would we add to the original graph? (2 point).
- In [Metzler & Croft 07], the authors claimed that their latent concept expansion approach is a generalization of the *relevance models*. In what sense does the latent concept expansion approach generalize relevance models and other unigram language models? (1 point)

Part 3

The features used in [Metzler & Croft 07] all contain a logarithm function. For example,

$$f_{T_D}(q_i, D) = \log\left[(1 - \alpha) \frac{t f_{q_i, D}}{|D|} + \alpha \frac{c f_{q_i}}{|C|}\right]$$

- What if we drop all the logarithms and define a feature like $f_{T_D}(q_i, D)$ simply as follows?

$$f_{T_D}(q_i, D) = (1 - \alpha) \frac{t f_{q_i, D}}{|D|} + \alpha \frac{c f_{q_i}}{|C|}$$

Do you think this will affect the performance of the MRF model? Do you expect the new feature definition to work better or worse? Why? (2 points)

- Do you think the MRF model does (or can do) IDF-weighting implicitly? If yes, explain how; if no, explain why. (2 points)

Information Retrieval: Problem 3. Sentiment analysis

The questions refer to the following two papers:

[Eguchi & Lavrenko 06] Sentiment Retrieval using Generative Models, K. Eguchi and V. Lavrenko, Proceedings of EMNLP 2006, pp. 345-354.

[Popescu & Etzioni 05] Extracting Product Features and Opinions from Reviews, A. Popescu and O. Etzioni, Proceedings of HLT/EMNLP 2005, pp. 339-346.

Part 1

- a What is an implicit (product) feature? Can you give an example of it? (1 point)
- b In both [Eguchi & Lavrenko 06] and [Popescu & Etzioni 05], the authors attempted to automatically determine the sentiment polarity of opinions. But the granularity of opinions addressed in these two studies is different. What is the difference? (1 point)
- c PMI is used in [Popescu & Etzioni 05] for multiple purposes. What does PMI stand for? Name two different purposes that the authors have used PMI for. (1 point)

Part 2

The following questions are about the relaxation labeling (RL) algorithm used In [Popescu & Etzioni 05].

- a The RL algorithm has been used in [Popescu & Etzioni 05] to assign a label to each word in a graph. What do these labels mean? How many possible values can a label have? (1 point)
- b The support function of RL is as follows:

$$q(w, L) = \sum_{k=1}^{3^{|N|}} p(l(w) = L|A_k) * p(A_k).$$

What is A_k ? What is N ? How is $p(l(w) = L|A_k)$ computed? (2 points)

Part 3

In Web search, an important challenge is to identify spam pages. Given a basic content-based spam recognizer C (a trained binary classifier that would classify a web page as either spam or non-spam). Can you think of a way to apply relaxation labeling to extend this basic classifier and leverage relations between web pages to potentially improve spam detection accuracy? You may introduce notations and give a sketch of the algorithm. (4 points)

Bioinformatics: Problem 1. Segal et al, 2008.

Part 1

Name the two most important physical/biochemical factors that influence the mapping between regulatory sequence and gene expression. How are these two factors quantified in the approach of Segal et al. ? In other words, how do the authors obtain the values of these two quantities during their analysis? (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. (2 points)

What prevents them from being able to calculate the probability of expression, $P(E)$ as per their notation, efficiently? (1.5 points)

Why do the authors need to compute the gradient of the function $P(E)$? Briefly explain whether computing this gradient, with respect to the expression contribution w_{tf} or the basal expression w_0 , is more or less complex than computing the function value itself. (2.5 points)

Part 3

One of the interesting exploratory analyses performed in the paper is where the authors synthetically mutate the cis-regulatory modules and examine the effect on expression, thereby investigating the robustness of the transcriptional network involved in segmentation. If you had their entire implementation in hand, what are the two biological investigations you would be most interested in conducting? (2 points)

Bioinformatics: Problem 2. Siepel and Haussler, 2003.

Part 1

Very briefly, explain the main purpose of using an HMM in Siepel & Haussler's work. (That is, what important aspect of the data would not be accounted for if the HMM was not used?) (1 points)

Part 2

(a) How does the PhyloHMM of Siepel & Haussler allow for different rates at different positions? How does this affect time complexity? (2 points)

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

(c) How is missing data handled in the PhyloHMM of Siepel & Haussler? (1.5 points)

Part 3

Imagine running the PhyloHMM program on a genome twice: once with the HMM going from left to right and once going from right to left. What do you expect to see in terms of the output of the program from the two runs? How would you go about reconciling any possible differences in the two outputs? What might be a possible direction of research that addresses such "left-right asymmetry" of the PhyloHMM approach? (3 points)

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

Part 1

Describe the basic idea of using “informative priors based on transcription factor structural class”. Your description should be short and at a high level, without the need for mathematical expressions. (2 points)

Part 2

(a) 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 . (3 points)

(b) Is this an unconditional probability? If not, what is it conditional on? (1 points)

(b) Is there a simplification (potentially unrealistic assumption) made in the formulation of the above prior probability on Z_i ? (1 points)

Part 3

Suppose you have multiple types of informative priors (not just based on transcription factor structural class). How would you go about designing a motif-finding framework (similar to the one in Narlikar et al., using Gibbs sampling) that exploits these priors. Explain only the “prior” part of your design, not the “Gibbs sampling” part. Would you consider allowing different types of priors to have different levels of contribution to the analysis? (3 points)