DAIS Qualifying Examination

Spring 2015 (March 9, 2015)

Department of Computer Science
University of Illinois at Urbana-Champaign

Time Limit: 180 minutes

Please read the following instructions carefully before starting the test:

- The examination has 15 questions, including one basic concept question and 14 topic questions. You are required to answer the basic concept question and any **5** of the 14 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 **18** subquestions. You are required to answer **9** of these 18 subquestions. If you answer more than 9 subquestions, the committee will randomly select 9 to grade.
- The 14 topic questions are distributed by areas as follows:
  - **Database Systems**: 3 questions
  - **Data Mining**: 3 questions
  - **Information Retrieval**: 3 questions
  - **Bioinformatics**: 5 questions
- 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. You may find it useful to periodically assess your time usage, and adjust your time allocation dynamically as needed to avoid running out of time on questions that you could have answered.

Required Question (Basic Concepts): Problem 0

You are required to answer 9 out of the following 18 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) (*Data Models*) Identify one major point of how "NoSQL" DBMS differes from Relational DBMS. Explain it with an example.

(2) (*Indexing*) What are the key requirements for an index structure for DBMS? Identify two points.

(3) (*Query Languages*) Is relational algebra a query language? Explain why.

Note: Please start with a new answer sheet.

Short answers: The answer of each of the following 3 questions is expected to be no more than two lines (i.e., no explanation nor reasoning is needed)

(4) (*Data cube algorithms*) Name one algorithm that efficiently computes data cubes for each of the following cases: (i) dense data sets of a small number of dimensions, (ii) computing data cube of a small number of dimensions with a given minimum support threshold (also known as *iceberg cubes*), and (iii) assisting query answering in a high-dimensional cube-space.

(5) (*Comparing data mining concepts/methods*) Use one or two sentences to distinguish the following pairs of concepts or methods: (1) *k-Means* vs. *Expectation-Maximization (EM)*, and (2) *Naive Bayes classifier* vs. *probabilistic graphical model.*

(6) (*Selection of data mining methods*) Name or outline one data mining method that best fits each of the following tasks: (1) *clustering of high-dimensional micro-array data*, and (2) *find frequent items and their "approximate" frequency counts for online data streams.*

Note: Please start with a new answer sheet.

(7) (*TF-IDF weighting*) Use no more than 3 sentences to briefly explain what is TF-IDF weighting.

(8) (*Inverted Index*) Which of the following strategies is more effective for reducing the size of an inverted index? (1) reduce k common words; (2) remove k rare words. Use no more than two sentences to explain why.

(9) (*Evaluation*) Let $R = (+, +, -, -, -, -, -, -)$ be the relevance status of 8 documents in a ranked list of retrieval results with "+" indicating a relevant document and "-" a non-relevant document. (Thus, the first two documents are relevant, while the rest are non-relevant). Suppose there are in total 10 relevant documents in the whole collection. Compute the precision, recall, and average precision. It's sufficient to give an expression; there is no need to reduce the expression to a value.

Note: Please start with a new answer sheet.

(10) (*Sequence alignment*) What does an affine gap penalty model mean in the context of pairwise sequence alignment ?

(11) (*Gene prediction*) When an HMM is used for gene prediction, what do its states represent ?

(12) (*Phylogenetic tree*) I am given a phylogenetic tree T1. I multiplied every branch length by a factor of 2 to get a new tree T2. The probability of a nucleotide x changing to nucleotide y on any branch b is twice as much in tree T2 than in T1. True or false ?

(13) (*Phylogenetic tree reconstruction*) What is the maximum parsimony principle for evolutionary tree building?

(14) (*Gene expression analysis*) You have 10 samples of microarray data from mice with a specific disease and 10 samples from healthy mice. Name a statistical test that you might use to determine differentially expressed genes in the disease (versus healthy).

(15) (*Clustering of biological data*) Name a clustering method where you do not have to decide upon the number of clusters you want ahead of time (i.e., before running the clustering algorithm).

(16) (*Motif finding*) The advantage of using position weight matrices (PWMs) as a motif model is that you can calculate the probability of a site s having been sampled from a PWM W, i.e., the value of $\Pr(s|W)$. However, it is easy to see that a longer PWM will typically give smaller probability values, even to its best sites, than a shorter PWM gives to its best sites. How is this apparent bias addressed in scoring sites for matches to PWMs?

(17) (*DNA Sequencing*) Name a graph algorithm that can be used for assembling sequencing reads.

(18) (*Regulatory Genomics*) Give one reason why a transcription factor ChIP peak near a gene may not be indicative of that transcription factor regulating that gene.

Database Systems: Problem 1

The paper "Incremental and Accuracy-Aware Personalized PageRank through Scheduled Approximation" addresses computing PPV efficiently– through incremental approximation.

Part 1

What does *accuracy-awareness* mean? Why is it important? (*2 points*)

Part 2 The accuracy, in terms of L1 error, is given in the paper.

1) Give the derivation of this error. (*2 points*)

2) Identify the intuition behind the equation. (*2 points*)

Part 3

The technique requires selecting a set of *hubs*, for preprocessing to build precomputed PPV "segment" values.

The paper suggests a scheme considering the factors of *decaying power* and *overall popularity* of a node to select good hubs.

We ask you to disagree with this scheme. Give two *different* factors of considerations for hub selection. (*4 points*)

Database Systems: Problem 2

The paper "Incremental and Accuracy-Aware Personalized PageRank through Scheduled Approximation" addresses computing PPV efficiently– through incremental approximation.

Part 1

What is the key insight of this paper for speeding up computation of Personalized PageRank? (*2 points*)

Part 2

How is "Personalized" PageRank different from the "generic" PageRank? Identify their similarities and differences. (*2 points*) Give each one an example application scenario. (*2 points*)

Part 3

PageRank is one type of graph ranking metric for measuring connection between nodes on a graph. There are other metrics proposed in the literature, such as SimRank or Hitting Time. You may think of more examples.

Do you think the general insight in this paper is applicable to other different metrics? Why? (*4 points*)

Database Systems: Problem 3

The paper "Optimization for iterative queries on MapReduce" develops a query optimization approach for iterative queries in distributed environment.

Part 1

Given an example workload of "iterative querying". (*1 points*)

Part 2

Why is standard Map Reduce inappropriate for such iterative querying? Explain, using an example. (*2 points*)

What is the central idea in the paper for optimizing such querying? (*2 points*)

Part 3

The key concept of the paper lies in the identification and handling of *invariant views* and *variant views.*

1) Why are these two concepts important? (*2 points*)

2) How are they treated differently? Why? (*3 points*)

Data Mining: Problem 1

Sequential pattern mining is useful in the analysis of DNA sequences, shopping sequences, web click streams, and text documents.

Part 1

(a) Suppose a customer shopping sequence dataset D contains only two sequences

(s_1)	$\langle a_1, a_2, \ldots, a_{10}, b_1, b_2, \ldots, b_{10}\rangle$
(s_2)	$\langle a_1, b_1, a_2, b_2, \ldots, \ldots, a_{10}, b_{10}\rangle$

where $a_i \neq b_i$ (for any i). Suppose the minimum support threshold is two. How many sequential patterns does this data set D contain? (*2 points*)

(b) Since sequential pattern mining may generate a large number of patterns, it is desirable to derive a compressed set of patterns.

(pattern_id)	⟨sequence⟩	support_count
(p_1)	$\langle a, b, c, d, e, b\rangle$	20239
(p_2)	$\langle a, b, c, d, e, a, b\rangle$	11532
(p_3)	$\langle a, b, c, d, e, f, a, b\rangle$	11520

Suppose the set of patterns derived is like the above. Discuss what should be a good definition of measure for effective sequential pattern compression. (*2 points*)

Part 2

(a) Outline an efficient method that mines long sequential patterns (*e.g.*, length around 100). (*3 points*)

(b) Outline an efficient algorithm for construction of an effective sequential pattern classifier. (*3 points*)

Data Mining: Problem 2

Many real-world datasets form heterogeneous information networks naturally. We examine the DBLP dataset, which consists of the bibliographic information of a large set of computer science research papers, with each entry recording a list of authors, a title, a publication venue, and a year.

Part 1

(a) Describe how ranking-based clustering, such as NetClus, can be performed efficiently to derive clusters for venues, terms and authors as well as their rankings. (*2 points*)

(b) Outline an effective method that extracts hierarchical term relationships from the DBLP dataset. For example, one level-1 node may contain {information retrieval, web search, text mining} and one of its children nodes may contain {text categorization, text classification, document clustering, multi-document summarization}. (*2.5 points*)

Part 2

(a) Design an active learning method for the heterogeneous DBLP network so that one can effectively build a classification model for the network by selecting a small number of nodes in the network to ask for labels from experts. (*2.5 points*)

(b) Outline a network mining scheme that may mine the DBLP network to support expert finding for queries represented by a set of terms, such as {*sensor network mining*}? (*3 points*)

Data Mining: Problem 3

Knowledge Vault (KV) is a web-scale probabilistic knowledge base, developed by Google, that combines extractions from Web content with prior knowledge derived from existing knowledge repositories.

Part 1

(a) Knowledge Vault adopts a *Local Closed World Assumption* (*LCWA*) to determine the labels for their classifiers. Describe what is the Local Closed World Assumption and at what situation this assumption is more appropriate than at other situations. (*2 points*)

(b) For automatic knowledge-base construction, YAGO, DBPedia and Freebase are built on Wikipedia infobox and other structured data sources. What are the strength and weakness of this approach? And what are the differences between the approach adopted in Knowledge Vault and that in Freebase? (*2.5 points*)

Part 2

(a) It is not unusual that conflict information could be extracted from different web pages. However, some information conflicts are due to the different creation times of webpages (*e.g.*, Obama could be a *senator*, a *president candidate* or *the U.S. president* at different times). Outline a method that may resolve such conflicts effectively and distinguish whether the error is due to time conflict or due to data error, without seeking expert help. (*2.5 points*)

(b) Propose an approach on knowledge-base construction than can be potentially more efficient and effective than the current Knowledge Vault does. (*3 points*)

Information Retrieval: Problem 1. Information Retrieval Models

[Luo et al. 14] refers to the following paper:

Jiyun Luo, Sicong Zhang, and Hui Yang. 2014. Win-win search: dual-agent stochastic game in session search, SIGIR'2014.

Part 1

The work [Luo et al. 14] gave a brief introduction to Markov Decision Process (MDP) and Q-learning, and defined an MDP as a tuple $< S, A, T, R >$.

a What does each of S, A, T, R refer to? (one phrase or sentence for each symbol is sufficient.) (*1 point*)

b What does Q-learning learn exactly? (*1 point*)

c Q-learning maximizes an objective function. What does this objective function measure? (*1 point*)

Part 2

a What are the four states proposed by the authors of [Luo et al. 14] in their dual-agent POMDP? (*1 point*)

b What is the message that a user sends to the search engine agent through the communication-level action? What is the message that the search engine agent sends to the user agent? Use one sentence to answer each of these two questions. (*1 point*)

c The optimal action of a search engine is given by

$$a_{se} = \arg\max_a (Q_{se}(b, a) + Q_u(b, a_u)).$$

(1) In order to solve this optimization problem, we must know a_u. How does the system know the value of a_u? (2) On the surface, $Q_u(b, a_u)$ does not contain a, the variable to optimize. Briefly explain how the choice of a for the search engine agent should be affected by $Q_u(b, a_u)$. (*2 points*)

Part 3

In [Luo et al. 14], the authors proposed a dual-agent POMDP, but in the end, we only compute an optimal policy for the search engine agent. Briefly explain how we might be able to use the standard (single-agent) POMDP with a more complicated state representation to solve the same problem. Sketch the idea using some formulas if possible. (*3 points*)

Information Retrieval: Problem 2. Text Mining

[Chen & Liu 14] refers to the following paper:

Zhiyuan Chen, Bing Liu, Mining topics in documents: standing on the shoulders of big data, KDD 2014

Part 1

a One motivation of the work [Chen & Liu 14] is "Given a small number of documents, the classic topic model LDA generates very poor topics." Briefly explain why this is so. (*1 point*)

b The authors of [Chen & Liu 14] proposed to mine knowledge of two relations, i.e., "must-link" and "cannot-link". What entities are these two relations defined on (i.e., what kind of entities cannot or must be linked)? (*1 point*)

c According to [Chen & Liu 14], how did the authors use the knowledge of "must-link" and "cannot-link" to address the limitation of LDA in performing poorly when there are a small number of documents? Use no more than 2 sentences to answer this question. (*1 point*)

Part 2

a When mining "must-link" knowledge from the output of a set of prior topics generated by a topic model, multiple thresholds must be applied. List all the thresholds that must be set. (*2 point*)

b In their experiments, some baseline methods that used must-link and cannot-link knowledge actually performed worse than LDA. What was the authors' explanation? (*1 point*)

Part 3

a What's your strongest criticism of the work [Chen & Liu 14] if you were a reviewer of this paper? (*2 points*)

b The idea of incorporating must-link and cannot-link knowledge into a topic model can also be potentially implemented by adding some regularization constraints to the Probabilistic Latent Semantic Analysis (PLSA) model. Sketch a possible objective function based on the likelihood function and some specific forms of constraints to incorporate must-link and cannot-link knowledge. (*2 points*)

Information Retrieval: Problem 3. Evaluation

[Dincer et al. 14] refers to the following paper:

B. Taner Dincer, Craig Macdonald, Iadh Ounis, Hypothesis testing for the risk-sensitive evaluation of retrieval systems, SIGIR 2014.

Part 1

a According to [Dincer et al. 14], what is risk-sensitive evaluation? What is exactly the “risk” here? (*2 point*)

b Briefly explain how U_{risk} is defined. (*1 point*)

Part 2

a Briefly explain what are the main contributions of the work [Dincer et al. 14]. (*2 points*)

b The authors of [Dincer et al. 14] proposed to compute the t-score of a raw U_{gain} measurement, i.e., T_{gain}. The authors then claimed that T_{gain} differs from the raw U_{gain} measurement in two important aspects. User no more than 4 sentences to explain what are these two aspects. (*2 points*)

Part 3

a The authors said “the direct application of T_{risk} in LambdaMART to attain risk-sensitive optimisation cannot offer marked improvements on the resulting learned models than U_{risk}.” Briefly explain why. (*1 points*)

b It is often desirable to perform utility-sensitive optimization of ranking where the goal is to emphasize more on improving accuracy on “high utility” topics (i.e., topics representing queries frequently typed in by users) than on rare topics (queries). Briefly explain how we can adapt the idea of FARO and SARO proposed by the authors of [Dincer et al. 14] to achieve this purpose. Sketch some formulas if possible. (*2 points*)

Bioinformatics: Problem 1. Stormo and Fields 1998

(a) The authors consider the binding constant of a site X_i for a transcription factor T:

$$K_{eq}(X_i) = \frac{[T \cdot X_i]}{[T][X_i]}$$

They then define a normalized version of this binding constant, denoted by $K_s(X_i)$. Upon this normalization, the probability that a TF molecule binds to X_i, given that it has the possibility of binding anywhere in the genome, is given by what function of $K_s(X_i)$? (*1 points*)

(b) Does a gel shift assay allow direct measurement of the binding constant $K_{eq}(X_i)$? If not, does it allow measurement of some quantity (formula) related to $K_{eq}(X_i)$? (*2 points*)

(c) How are the quantities ΔG_s (binding energy) and K_s (binding constant) related to each other? (*1 points*)

(d) *Fill in the blank at the end of this sentence*: Given a collection of binding sites and assuming these to be strong binding sites, one can show that the "weight matrix" given by the formula $W(b,j) = \log_2 \frac{f(b,j)}{p(b)}$ (where $f(b,j)$ is the frequency of base b at position j of the sampled sites and $p(b)$ is the frequency of base b in the genome) maximizes ... (*1 points*)

(e) The authors propose that given the ΔG_s values of a set of sites (e.g., single nucleotide mutants of the consensus site), one can predict the ΔG_s value for any site. What is the key assumption about binding energies that goes into this claim? (*2 points*)

(f) If the assumption alluded to in part (d) above is not true, what changes might you make to model binding site strengths? (*3 points*)

Bioinformatics: Problem 2. Degner et al. 2012

(a) What is an eQTL? (*2 points*)

(b) What is a dsQTL? (*2 points*)

(c) What is a possible mechanism to explain dsQTLs? (*2 points*)

(d) What relationship do the authors note between dsQTLs and QTLs in this paper? (*2 points*)

(e) What relationship do the authors note between dsQTLs and transcription factor binding? (*2 points*)

Bioinformatics: Problem 3. Gutenkunst et al. 2007

(a) The authors measure the change in model behavior with varying parameters θ by the χ^2 score:

$$\chi^2(\theta) = \frac{1}{2N_cN_s}\sum_{s,c}\frac{1}{T_c}\int_0^{T_c}[\frac{y_{s,c}(\theta,t) - y_{s,c}(\theta^*,t)}{\sigma_s}]^2 dt$$

Here, an average is taken over all molecular entities s and all conditions c. Explain the quantity whose average is taken, i.e., the term inside the summation. (*3 points*)

(b) The authors calculate the Hessian matrix of the χ^2 score:

$$H^{\chi^2}_{j,k} = \frac{d^2\chi^2}{d\log\theta_j d\log\theta_k}$$

and compute the eigenvalues of this Hessian matrix. These eigenvalues are what they use to quantify the "sloppiness" of the model around the optimum θ^*. Briefly explain their rationale for doing so. That is, why are they looking at the Hessian matrix? What do the eigenvalues of the matrix reflect intuitively? (*4 points*)

(c) The authors argue that even if parameters of a model are poorly determined, i.e., there is great uncertainty in the values of some of the parameters, this may not necessarily be a problem for the modeler. Why is this so? (*3 points*)

Bioinformatics: Problem 4. Guan et al. 2008

(a) What is the classification problem addressed in this paper? (What is the output label to be predicted, and what data forms the input vectors?) (*2 points*)

(b) The baseline classifier utilized in this work is an SVM with bootstrap aggregation (bagging). What does 'bagging' mean? (*1 points*)

(c) Briefly describe the main idea behind "Bayesian hierarchical combination of SVM classifiers" on the GO hierarchy. You do not need to describe the actual inference algorithm, which was described in an earlier paper. Just describe how the Bayesian network is set up and what are the observables and hidden nodes in the network. (*4 points*)

(d) One of the techniques explored in this work was a naive Bayes classifier to combine multiple classifiers' outputs. What was the difference among these classifiers whose outputs were combined? (*2 points*)

(e) Which of the four methods tested in this work (baseline SVM, Hierarchy:MarkovBlanket, Hierarchy:Tree, naive Bayes) worked best most of the time? (*1 points*)

Bioinformatics: Problem 5. Nourmohammad and Lassig 2011

(a) The first analysis that the authors show in this paper is that of computing an autocorrelation function on cis-regulatory modules. The autocorrelation function is defined by them as $\Delta(r) = c(r) - c_0$. Explain this formula. (That is, what are $r, c(r)$ and c_0?) (*1 points*)

(b) Provide two different sources suggested by the authors for the autocorrelation signal seen by them using the above formula. (They actually list more than two sources.) (*2 points*)

(c) How do the authors define the "similarity information" of a pair of sites? (*3 points*)

(d) What is the maximization problem they solve for a given cis-regulatory module, using dynamic programming, in their attempt to study the extent of local duplications in these modules? (*2 points*)

(e) What is the one message you will take home from this paper about evolution of binding sites in Drosophila cis-regulatory modules? (*2 points*)