DAIS Qualifying Examination
Spring 2014

Department of Computer Science
University of Illinois at Urbana-Champaign

March 4, 2014.
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) (*Extensibility*)
As built-in data types in a DBMS are often restrictive, it is desirable to support *extensible types*. Identify one significant complication for supporting type extensibility for query processing. Explain why.

(2) (*Parallel DBMS*)
In parallel databases, is data partitioning scheme (to spread data across different computing nodes) important for query efficiency? Explain why.

(3) (*Relation Algebra*)
Some may say: *Relational algebra is a better query language than SQL*. Give one argument to support this statement.

Note: Please start with a new answer sheet.

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

(4) (Data mining algorithms) Name three algorithms that efficiently mine frequent patterns, sequential patterns, and frequent subgraph patterns from large datasets respectively.

(5) (Data mining or data cube concepts) Use one or two sentences to distinguish the following pairs of concepts or methods: (1) *k-means* vs. *Expectation-Maximization (EM)*, and (2) *support vector machines (SVM)* vs. *probabilistic graphical models*.

(6) (Selection of data mining methods) Name or outline one data mining method that best fits each of the following tasks: (1) *find rivers and lakes from Google Earth maps*, and (2) *model construction for fraud detection based on online data streams*.

Note: Please start with a new answer sheet.

(7) (*Retrieval Models*) Use no more than two sentences to explain how the query likelihood retrieval function can achieve the effect of IDF weighting.

(8) (*PageRank*) Suggest two different ways to increase the PageRank score of the homepage of the Department of Computer Science at UIUC.

(9) (*Evaluation*) Let $R = (-, +, -, -, -)$ be the relevance status of five documents in a ranked list of retrieval results with "+" indicating a relevant document and "-" a non-relevant document. (Thus, the first document is non-relevant, while the second is 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*) Why is the standard dynamic programming approach, which is guaranteed to find the optimal alignment, used for the problem of multiple sequence alignment?

(11) (*Gene prediction*) Name one important challenge in the goal of prokaryotic gene prediction. (Note that prokaryotes do not have exons and introns.)

(12) (*Gene expression*) Why is observed mRNA level of a gene not a great proxy for protein concentrations?

Database Systems: Problem 1

The paper "Shark: SQL and Rich Analytics at Scale" presents a system that marries query processing with complex analytics on large clusters.

Part 1

What is the purpose of this paper– that is, what is the central thesis that the authors attempt to address? (*2 points*)

Part 2

The paper claims to support the two objectives of *iterative* computation as well as *interactive* querying.

Identify the key technique that makes this possible? (*1 points*)

Describe how it works. (*2 points*)

Explain why it enables the two objectives. (*2 points*)

Part 3

This paper puts in contrast two common data processing: *SQL queries* vs. *sophisticated analytics functions*. We ask you to think their difference.

1) Give an application scenario for each, to illustrate how they are different. (*1 points*)

2) For computing efficiently, how do they impose different requirements? (*2 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 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 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 and Data Warehousing: Problem 1

It is desirable to construct information networks and perform multidimensional analysis over a large collection of text documents (e.g., news articles, computer science publications).

Part 1

(a) Why is it more desirable to use phrases for modeling topics in text data than use single words (i.e., uni-grams)? (*1 point*)

(b) Outline an effective method that may generate high quality phrases from a large collection of text documents. (*2 points*)

(c) Suppose one can successfully model text data by phrases. Take news data as an example, outline how to construct a data- and text-cube that consists of multi-dimensions (e.g., person, organization, location, time, and event) and possibly multiple-levels. (*2 points*)

Part 2

(a) A collection of text documents can be viewed as inter-connected information networks. Draw the schema of such a heterogeneous information network. (*1 point*)

(b) A heterogeneous-network-Cube can be constructed on such a heterogeneous information network. What should be the OLAP operations on such a heterogeneous information network? How is this cube different from Graph-Cube, described in Zhao et al. (2011)? (*2 points*)

Note: Zhao et al. (2011) refers to the paper: *Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han, "Graph Cube: On Warehousing and OLAP Multidimensional Networks", Proc. of 2011 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'11), Athens, Greece, June 2011.*

(c) Given a news article, outline how search for similar news articles can be realized effectively in such a multi-dimensional, multi-level heterogeneous information network. (*2 points*)

Data Mining and Data Warehousing: Problem 2

Part 1

(a) There are often multiple information sources that provide much information on entities. However, the information provided by different sources may not be consistent. Explain why a simple majority voting may not be a good way to find truth and resolve inconsistency. (*1.5 point*)

(b) For truth-finding, what are the major differences on the design ideas of LTM (by Zhao et al. (2012)) in comparison with TruthFinder (by Yin et al. (2008))? (*2.5 points*)

Note:

- Zhao et al. (2012) refers to the paper: *Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han, "A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration", Proc. 2012 Int. Conf. on Very Large Data Bases (VLDB'12/PVLDB), Istanbul, Turkey, Aug. 2012.*
- Yin et al. (2008) refers to the paper: *Xiaoxin Yin, Jiawei Han and Philip S. Yu, "Truth Discovery with Multiple Conflicting Information Providers on the Web", IEEE Transactions on Knowledge and Data Engineering, 20(6):796-808, 2008.*

Part 2

(a) The above methods study truth-finding for categorical data. In reality, there are cases that one needs to find truth for numerical values if multiple sources give conflicting information. For example, different web-sites may give LA's population as 3.9 million, 2.8 million, or 3,858,852. Design an effective method that find quality truth from such conflicting numerical data. (*3 points*)

(b) A heterogeneous information network may help truth-finding. For example, building an information network to link among book sellers, among books, and among entities in both categories, may help consolidate the truth. Outline a method that will explore the power of a heterogeneous information network in truth finding. (*3 points*)

Data Mining and Data Warehousing: Problem 3

Part 1

(a) Many tweets are geo-coded (i.e., their geo-locations are known). Suppose a tweet contains user-id, hashtag, time, location, and message body. Design a *tweet stream cube* and outline how on-line tweets can be incrementally incorporated (i.e., summarized) into such a multi-dimensional stream cube. (*2 points*)

(b) If one wants to find spatiotemporal-related sequential patterns for tweets, outline an efficient sequential pattern mining algorithm for it. (*2 points*)

(c) One may like to cluster tweets based on the contents of their discussion. Outline a scalable tweet clustering method. (*2 points*)

Part 2

(a) One may construct a heterogeneous information network based on the relationships among tweets (e.g., pose, follow, retweet, etc.) and the content information. Outline how such a network should be designed and constructed. (*2 points*)

(c) Discuss how to use such a network to uncover outliers in tweet streams. (*2 points*)

Information Retrieval: Problem 1. Retrieval Models

[Paik 13] refers to the following paper:

Jiaul H. Paik. 2013. A novel TF-IDF weighting scheme for effective ranking, SIGIR'13.

Part 1

Use no more than 4 sentences to briefly explain what are the two aspects of TF (RITF and LRTF) proposed in [Paik 13].

(*2 points*)

Part 2

a According to [Paik 13], why is it desirable to combine the two TF factors (i.e., BRITF(t,D) and BLRTF(t,D))? (*1 point*)

b In [Paik 13], the two TF factors are combined as follows:

$$TFF(t, D) = w * BRITF(t, D) + (1 - w) * BLRTF(t, D).$$

Use no more than 3 sentences to explain how the parameter w is set. (*2 points*)

c Use no more than 4 sentences to explain what are the most important differences between the final ranking function proposed in [Paik 13] and BM25. (*2 points*)

Part 3

Suggest a way to extend the query likelihood retrieval function to implement the idea of combining the two TF factors proposed in this paper. Sketch your idea with formulas if possible. (*3 points*)

Information Retrieval: Problem 2. Text Mining

[Zhu et al. 13] refers to the following paper:

Xingwei Zhu, Zhao-Yan Ming, Xiaoyan Zhu, and Tat-Seng Chua. 2013. Topic hierarchy construction for the organization of multi-source user generated contents, SIGIR '13.

Part 1

a According to [Zhu et al. 13], what is exactly a topic? How is their notion of a topic similar to (or different from) the topic defined in a probabilistic topic model such as LDA? (*2 point*)

b How is the "sub-topic relation" defined in [Zhu et al. 13]? Give an example of such a relation. (*1 point*)

Part 2

a The proposed framework of [Zhu et al. 13] consists of three modules. Use no more than 5 sentences to briefly explain how each module works. (*1 points*)

b The authors of [Zhu et al. 13] proposed to leverage a search engine for both topic set extension and topic relation identification. Give an example query pattern used in each case. (*2 points*)

Part 3

[Zhu et al. 13] motivated their paper by the problem of generating a topic hierarchy for *multiple* sources of information and proposed methods that would also exploit different resources such as Wikipedia and search engine results.

a What are the input and the output of the computation problem that the authors of [Zhu et al. 13] tried to solve? Compared with similar problems solved in the existing literature, is this problem novel? If so, where is exactly the novelty? (*2 points*)

b What do you think about their technical approaches? What is your strongest criticism of their approaches? Based on your criticism, suggest an idea to further improve their approaches. (*2 points*)

Information Retrieval: Problem 3. Evaluation

[Urbano et al. 13] refers to the following paper:

J. Urbano et al. On the measurement of test collection reliability, SIGIR'13.

Part 1

Use no more than 6 sentences to briefly describe how the Cranfield evaluation methodology works (i.e., how evaluation of retrieval algorithms is typically done as described in [Urbano et al. 13]). (*2 points*)

Part 2

a How is the "reliability" of an evaluation result defined in [Urbano et al. 13]? (*1 point*)

c [Urbano et al. 13] provided two Generalization Theory-based indicators of reliability, i.e., Generalizability Coefficient and Index of Dependability. How is each defined? (*2 points*)

d What are the main take-away messages from [Urbano et al. 13]? (*2 point*)

Part 3

The authors of [Urbano et al. 13] mainly examined the reliability from the query perspective. One can ask a similar question about reliability from the perspective of documents. Following the ideas of [Urbano et al. 13], briefly discuss how we might be able to define indicators of reliability from the perspective of documents. (*3 points*)

Bioinformatics: Problem 1. Kantorovitz et al. 2007

Part 1

(a) Provide one important motivation for computing an alignment-free similarity score for regulatory sequences. (*1 points*)

(b) Describe at least one alternative to the D2z score, from the prior literature, that the authors compared their new score to. (*1 points*)

(b) What is the D2 score, whose z-score ('D2z') is computed in this paper? (*2 points*)

(c) Show briefly how the expectation of the D2 score is calculated for an IID and a Markov model background. (*3 points*)

Part 2

Based on the results, what problem do you think the proposed method is good at solving and what problem is it not so good at solving? What might be some reasons why the latter problem (which the method is not so good at) is hard? (*3 points*)

Bioinformatics: Problem 2. Peccoud and Ycart 1995

Part 1

(a) When analyzing what kind of data may you require a Markov model of gene product synthesis ? (*2 points*)

(b) What are the biochemical interactions included in the proposed Markov model? (*1.5 points*)

(c) What is the state space of the proposed Markov model? (*1.5 points*)

Part 2

Explain the derivation of the formula for the time derivative of the probability $p_{0,n}(t)$ that the system is in state $(0, n)$ at time t. (*3 points*)

Why is the set of differential equations difficult to solve when the protein degradation rate is strictly positive? (*2 points*)

Bioinformatics: Problem 3. Zinzen and Papatsenko 2007

Part 1

(a) Explain the title of the paper – 'Enhancer Responses to Similarly Distributed Antagonistic Gradients in Development' – especially the term 'antagonistic gradients'. (*2 points*)

(b) How is the case of one activator site and one repressor site modeled? (*2 points*)

(c) How is the case of cooperativity among two or more binding sites of the activator modeled? (*2 points*)

Part 2

What would you do if you performed a similar modeling exercise and found several distinct optima in the parameter space? How will this affect your reasoning about the system? What precautions will you take in such reasoning? (*4 points*)