DAIS Qualifying Examination

Spring 2016 (March 3, 2016)

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 10 questions, including one basic concept question and 9 topic questions. You are required to answer the basic concept question and any **5** of the 9 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 **9** subquestions. You are required to answer all of these subquestions.
- The 9 topic questions are distributed by areas as follows:
  - **Database Systems**: 3 questions
  - **Data Mining**: 3 questions
  - **Information Retrieval**: 3 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 all of the following 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*) To support indexing of *arbitrary* data types beyond the basic ones (e.g., integers, strings)– such as indexing *sets* of values– identify two key issues.

(3) (*Query Languages*) Can we use a programming language such as Python as a query language? Explain why (or why not).

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 two algorithms for each of the following cases:

(i) *generate quality, arbitrary shaped clusters*, and

(ii) *classify high-dimensional data effectively* (note: "high-dimensional" means higher than 50 dimensions here).

(5) (*Comparing data mining concepts/methods*) Use one or two sentences to distinguish the following pairs of concepts or methods:

(i) *k-Means* vs. *k-NN*, and

(ii) *boosting* vs. *bagging*.

(6) (*Selection of data mining methods*) Name or outline one data mining method that best fits each of the following tasks:

(i) *clustering online data streams*, and

(ii) *find frequent substructures from a large set of chemical compounds*.

Note: Please start with a new answer sheet.

(7) (*IDF*)

Use a specific example of query to briefly explain the benefit of IDF weighting in a retrieval function. (Use no more than 3 sentences.)

(8) (*Inverted index*)

Use no more than 3 sentences to explain how an inverted index can be used to quickly retrieve all the documents matching both "computer" and "virus" in the Boolean query "computer AND virus".

(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 document is relevant, while the second is 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.

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. (*1 point*)

Part 3

Do you think this technique can be generalized beyond PPV to other graph *proximity* measures?

1) Identify the key principle behind the techniques in this paper. (*1 point*)

2) Explain how it may be generalized for other proximity measures. Use one such measure as a new problem, and explain how the principle can be extended to the new problem. (*4 points*)

Database Systems: Problem 2

The paper by Low et al. describes GraphLab, a special-purpose graph analysis engine.

Part 1

What are the benefits of specialized graph processing systems over MapReduce systems? (*1 point*)

Part 2

Compare GraphLab with MapReduce. You will be graded by the comprehensiveness and depth of the arguments.

What are the concepts that they have in common? Why? (*1 point*)

What are the concepts that they *significantly* differ? Why? (*2 points*)

Part 3

One interesting notion in GraphLab is consistency models and hence the consistency guarantee.

1) Give one example computing task (a concrete, specific task) that is suitable for each of the three consistency models, and explain why. Use *different* examples from what the paper uses. (*3 points*)

2) How is this consistency notion handed in MapReduce? For the examples you gave above, explain how one would handle it in MapReduce. (*3 points*)

Database Systems: Problem 3

The following questions are about Pavlo et al.: A comparison of approaches to large-scale data analysis.

Part 1

According to the paper, which of these systems (Hadoop, Parallel Databases) performs better on *each* of the following aspects? (*2 points*)

1. Data Loading
2. Selection
3. Compression
4. Code Reuse

Part 2

According to the comparison, there is a large margin of difference in the join task, where parallel DBMS significantly outperforms Hadoop. Why? Give the key reasons. (*1 point*)

Despite the subpar performance of MapRecuce when compared to parallel database systems, it still enjoyes a big success. Why? (*2 points*)

Part 3

As inspired by the comparison and analysis of this paper, we ask you to suggest *three* aspects and sketch concrete ideas to change/improve the MapReduce framework. (*5 points*)

Data Mining: Problem 1

Phrases can be considered as minimal semantically meaningful units in text documents. Thus it is important to develop effective methods for mining phrases from text corpora.

Part 1

(a) Briefly outline a list of phrase mining methods that you know of and comment on their strength and weakness. (*2.5 points*)

(b) Explain why SegPhrase+ mines quality phrases effectively from text corpora and outperforms many previously studied phrase mining methods. (*2.5 points*)

Part 2

(a) SegPhrase+ does not use any natural language processing (NLP) methods in phrase mining. Give an example that NLP may further enhance the quality of phrases generated. Discuss how to integrate some NLP methods to further improve the quality of phrase mining results. (*2.5 points*)

(b) Previous mining of heterogeneous information networks does not explore the power of phrase mining. Take DBLP as a dataset, and one heterogeneous information network mining task as an example, explain how you will explore phrase mining to improve the quality of network mining. (*2.5 points*)

Data Mining: Problem 2

Many real-world datasets, such as news, tweets, scientific literature, contains lots of text data. It is a major challenge to turn such massive data into actionable knowledge.

Part 1

(a) Take a corpus like "2015 New York Times" as an example, explain why entity extraction and typing are essential for construction of quality heterogeneous information networks. (*1.5 points*)

(b) Explain how ClusType (X. Ren et al., KDD 2015) can find quality types for news corpus by distant supervision and reason why such a method can be more effective than many existing typing methods. (*3 points*)

Part 2

(a) The PubMed database contains millions of biomedical research papers generated from biomedical research. Outline your design of a set of methods that may perform step-by-step construction of heterogeneous biomedical information networks from such a PubMed database. (*2.5 points*)

(b) Suppose such a heterogeneous biomedical information network has been constructed, outline how you would like to "uncover" from such a network a set of drugs that may likely to be effective for a certain kind of disease? Reason on the effectiveness and efficiency of your proposed method. (*3 points*)

Data Mining: Problem 3

Social and Information Networks have both node and content information; for example, in Facebook, you have a friend relationship graph as well as attributes associated with each person on the social network.

Part 1

(a) In the CESNA paper (Yang 2013), the authors discover that the combining content with network structure improves the results. In particular, they find that CESNA performs best for information networks (Philosophers), in contrast to social networks (Facebook, Twitter, Google+). Provide an explanation why this might be so. (*2.5 points*)

(b) Construct a counter-example network with content where one can expect CESNA to **not** do well; provide a simple justification. (*2.5 points*)

Part 2

(a) What is assortitative mixing? Give a simple measure to calculate assortitative mixing. (*2 points*)

(b) In the general case of combining network structure and content to extract communities, how will you handle the cases of assortivity equal to zero and the case when assortivity is equal to 1? (*3 points*)

Information Retrieval: Problem 1. Information Retrieval Models

[Zhang15] refers to the following paper:

Yinan Zhang and Chengxiang Zhai. 2015. Information Retrieval as Card Playing: A Formal Model for Optimizing Interactive Retrieval Interface. Proceedings of SIGIR 2015.

Part 1

a What are the two assumptions made in the Probability Ranking Principle (PRP)? (*1 point*)

b Which of the two assumptions made in the PRP was relaxed in the IIR-PRP (PRP for Interactive Information Retrieval) proposed by Fuhr? (*1 point*)

c According to [Zhang15], what is an interface card? (*1 point*)

Part 2

The main formal framework of the interface card model for interactive retrieval proposed in [Zhang15] is the following:

$$\begin{array}{lll} maximize_{q^t} & & E(u^t|c^t,q^t) \\ & = & \sum_{a^{t+1}\in\mathcal{A}(q^t)} p(a^{t+1}|c^t,q^t)\, u(a^{t+1}|c^t,q^t) \\ \text{subject to} & & f_c^t(q^t) \leq 0 \end{array}$$

a Briefly explain what each of the following symbols in the optimization framework shown above donotes: t, q^t, u^t, a^{t+1}, $\mathcal{A}(q^t)$. (*2 points*)

b What is the meaning of the conditional probability $p(a^{t+1}|c^t,q^t)$? What constraint must this probability distribution satisfy? (*1 point*)

c What is the meaning of $u(a^{t+1}|c^t,q^t)$? (*1 point*)

Part 3

When responding to a query, a search engine such as Google would return a result page consisting of a ranked list of documents with buttons for fetching additional results (e.g., "next page") as well as a query box for a user to potentially enter a new query. To formally model such a search engine using the interface card model proposed in [Zhang15], how would you define $\mathcal{A}(q^t)$ to capture the possible user actions on such a result page? Give at least three different elements that should be included in the set $\mathcal{A}(q^t)$. How can we estimate $p(a^{t+1}|c^t,q^t)$ based on search log data? (*3 points*)

Information Retrieval: Problem 2. Evaluation

[Sebastiani15] refers to the following paper:

Fabrizio Sebastiani. 2015. An Axiomatically Derived Measure for the Evaluation of Classification Algorithms. Proceedings of ICTIR 2015.

Part 1

a Give the formula for computing the F1 measure. (*1 point*)

b How do you define the Accuracy measure used for evaluating classification results? Why is this measure not appropriate when the data set is imbalanced with many more data points in one class than another? (*2 points*)

Part 2

a In [Sebastiani15], the author used $M(D, Y_c, h_c)$ to denote a measure function. What does each of D, Y_c, and h_c denote? (*1 point*)

b In [Sebastiani15], the author has given multiple axioms. One of them is **Strict Monotonicity (MON)**. Give the definition of the **Strict Monotonicity** axiom. (*1 point*)

c In [Sebastiani15], the author has shown that F1 does not satisfy multiple axioms and thus may behave in an unreasonabl way. Give a specific example of a case where the F1 measure would behave in an unreasonable way (i.e., violate one of the axioms). (*2 points*)

Part 3

a Give the formula for computing the proposed K measure in [Sebastiani15] based on TP (true positives), TN (true negatives), AP (all positives), and AN (all negatives). (*1 point*)

b The proposed K measure satisfies many desirable properties, but is the K measure suitable for all kinds of binary text classification problems? If not, can you give a specific binary classification problem where K-measure isn't suitable? (*2 points*)

Information Retrieval: Problem 3. Application

[Shih10] refers to the following paper:

Shih, B., Koedinger, K.R., and Scheines, R. (2010). Unsupervised discovery of student strategies. In Proceedings of the Third International Conference on Educational Data Mining, pp. 201-210

Part 1

a What do we need to specify in order to fully define a hidden Markov model (HMM)? (*2 points*)

b HMMs are used to learn "learning tactics" from student behavior data. What are the data modeled using HMMs composed of? (*1 point*)

Part 2

a If an HMM has k states and the data to be modeled has m distinct symbols. How many transition probability parameters are there in total? How many emission probability parameters are there in total? (*2 points*)

b Briefly explain how the HMM-Cluster algorithm proposed in [Shih10] works? Use no more than 4 sentences to answer this question. (*1 point*)

c Give one limitation of HMM-Cluster that is addressed in Stepwise-HMM-Cluster. Use no more than 3 sentences to briefly explain how exactly Stepwise-HMM-Cluser addresses this limitation. (*1 point*)

Part 3

a The learning tactics discovered by the HMMs in [Shih10] are very simple ones (e.g., Aaaaaa), which are not so interesting from education perspective. One reason is due to the nature of the data set used. What specific characteristics of the data set have limited the complexity of the patterns that can be discovered by using HMMs? (*1 point*)

b If we are to use the Stepwise-HMM-Cluster on text data (e.g., English text), can we also expect the algorithm to return meaningful patterns? If so, what kind of patterns would you expect it to generate? (*2 points*)