DAIS Qualifying Examination
Fall 2016 (October 7, 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) (*Query Language*)
What is the most important property of a query language? Explain why.

(2) (*Query Processing*)
Join is expensive to process. Name two techniques for speeding up join processing and briefly explain how each of them helps.

(3) (*Indexing*)
Indexing helps query processing. Other than speeding it up (making it more efficient), give one more major advantage that indexing enables for query processing.

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)

1. *Name two algorithms that cluster high dimensional data sets effectively*, and
2. *Name two algorithms that mine frequent/closed subgraph patterns efficiently.*

(5) (Data mining concepts) Use one or two sentences to distinguish the following pairs of concepts or methods:

1. *KL-divergence* vs. *cosine similarity*, and
2. *Expectation-Maximization (EM)* vs. *K-means.*

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

1. *Scalable and efficient decision tree induction in large datasets*, and
2. *Online incremental clustering of massive data streams.*

Note: Please start with a new answer sheet.

(7) (*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.

(8) (*Inverted index*)

Imagine a computer science literature search engine ranks research papers using the BM25 (also called Okapi) retrieval function based on an inverted index. Which of the following four queries do you think would most likely take the longest time to answer? Use no more than two sentences to explain why.
(A) “computer software (B) “health care (C) “health software (D) “computer

(9) (*TF-IDF weighting*)

Suppose we use TF-IDF weighting to compute the term vector for a sports news article about baseball in a collection of general news articles. Which of the following words do you expect to have the highest weight in the term vector for this article? Use no more than two sentences to explain why.
(A) the (B) baseball (C) computer

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 is the key insight of this paper for speeding up computation of Personalized PageRank? (*1 points*)

Part 2

How is "Personalized" PageRank different from the "generic" PageRank? Identify their similarities and differences. (*1 points*) Give each one an example application scenario. (*2 points*)

Part 3

This paper focuses on evaluating Personalized PageRank. Do you think the general insight in this paper is applicable to the "generic" PageRank as well? Please make observations and speculate for both positive and negative positions:

1) Argue that this application *is* feasible. Why? (*3 points*)

2) Argue that this application is *not* feasible. Why? (*3 points*)

Database Systems: Problem 2

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

Part 1

What is the main purpose of GraphLab and why is such a system necessary? (*1 point*)

Part 2

The authors argue that "the MapReduce abstraction fails when there are computational dependencies in the data". Explain why. Give an example of *your own* to support the argument. (*3 points*)

Part 3

Compare the GraphLab framework to a parallel database (e.g., the Gamma system). How are they similar? (*3 points*) How are they different? (*3 points*) Use examples to illustrate and explain.

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*)

While both for parallel data processing, MapReduce differs quite significantly from parallel DBMS. Name two key differences– in the order of their importance– and explain why such differences are necessary. (*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. Use examples to support your arguments. (*5 points*)

Data Mining: Problem 1

Google researchers have recently been working on a project (called *Biperpedia*) that could discover entity-attribute structure for a "universe" (e.g., around 100,000) of attribute names.

Part 1

(a) What are the major difficulties at mining long and heavy tails of attribute names for a large number of entities from massive datasets? (*2 points*)

(b) Outline the major technical innovations of Google Biperpedia researchers that may overcome those difficulties. (*2.5 points*)

Part 2

(a) Google researchers have been exploring massive query log data at mining the universe of attribute names. However, most other researchers may not be able to access such massive query log data. Outline your proposed method that may effectively mine attribute names from massive data sets, such as news corpora or scientific corpora, without using query log data. (*3 points*)

(b) Entity-associated attribute structures mined from massive text data may contain noises and inconsistencies. Take news corpus or scientific corpus as an example, outline possible mechanisms that may help resolve inconsistencies and reduce noises from such mining. (*2.5 points*)

Data Mining: Problem 2

Smartphones and other smart devices have been generating tremendous amount of mobility data.

Part 1

(a) A lot of mobility-related data are very sparse (e.g., data generated from animal motion sensors). Outline an efficient method that mines (periodic or sequential) movement patterns from such very sparse data effectively. (*2.5 points*)

(b) Scattered movement patterns may not be interpretable or semantically meaningful (e.g., people may go to different restaurants). Outline an efficient method that may mine *semantically meaningful* movement patterns effectively (e.g., students may like to go for lunch on the Green street but this group of professors may like to go to dinner together in Japanese restaurants Fridays after work). (*2.5 points*)

Part 2

(a) Smartphone users may generate a lot of location-associated tweets (called *geo-coded tweets*). Design a stream data mining algorithm that can uncover *interesting bursty events* (e.g., not about jammed downtown traffic since it happens every day), finding the likely locations and the themes of the bursty events in real time. (*2.5 points*)

(b) In the future, sensors and mobile devices will be connected to the Web and communicate via messages. One may like to build a heterogeneous information network based on the types, locations, and communications of the devices. Outline your design on how such an information network should be constructed and mined to facilitate situation understanding. (*2.5 points*)

Data Mining: Problem 3

Consider a standard Erdos-Reyni Graph denoted by $G_{n,p}$, where n refers to the number of nodes in the graph and where p is the probability that two nodes are connected by an edge. Let $p = \frac{c}{n-1}$.

Part 1

(a) What is the probability that a node is disconnected from the rest of the network? (*2.5 points*)

(b) What is the probability that at least one node of the graph is disconnected? (*2.5 points*)

Part 2

Using the results in part 1, show that when $c = \log n$, the probability that the network is connected tends to 1, when $n \to \infty$ (*5 points*).

Information Retrieval: Problem 1. Information Retrieval Models

[Zhang16] refers to the following paper:

Yinan Zhang and Chengxiang Zhai. 2015. A Sequential Decision Formulation of the Interface Card Model for Interactive IR. Proceedings of SIGIR 2016.

Part 1

a If we use the Interface Card Model described in [Zhang16] to model a Web search engine such as Google, how many different interface cards can Google potentially play in respond to a user's query? (*2 points*)

b According to [Zhang16], what is a "plain card" and what is a "navigational card"? (*1 point*)

Part 2

The main formal framework of the sequential formulation of the interface card model proposed in [Zhang16] is the following recursive equation:

$$E(u^t|d^t) = \max_{q^t} \sum_{a^{t+1} \in \mathcal{A}(q^t)} \Big(p(a^{t+1}|d^t, q^t) \cdot \big(u_0(d^t, q^t, a^{t+1}) + E(u^{t+1}|d^{t+1})\big) \Big)$$

$$\text{subject to } f_c^t(q^t) \leq 0.$$

a What does $E(u^t|d^t)$ mean? (*0.5 point*)

b What is d^t? (*0.5 point*)

c What is $u_0(d^t, q^t, a^{t+1})$? (*1 point*)

d Suppose the user has only a very limited amount of time to interact with the search engine and thus would stop the interaction after taking just one action to interact with the search engine. What does the equation look like in this special case? (*2 points*)

Part 3

A search engine such as Google often includes a query box in every result page to offer the user an opportunity to reformulate the query and search using the new query. This is intuitively a good idea since the query box does not occupy much space on the screen but can potentially be very useful to a user if the user is not satisfied with the results returned. Briefly describe how we can apply the Interface Card Model proposed in [Zhang16] to analyze this component in the interface and the corresponding actions that a user can perform due to the inclusion of this component so as to figure out under what conditions it is beneficial to include such a query box and when the inclusion may not be beneficial. [**hints**: you may assume that the interface card contains 10 search results, a "next" button", and possibly a query box, and consider what actions a user can potentially take. Then try to formally analyze the expected surplus of an interface card with a query box and the expected surplus of the same interface card without the query box with consideration of a user's

potential actions. Introduce as many formal symbols as needed and do not worry about whether you can actually compute/estimate the values of those symbols; just use them to help analyze when it's a good idea to include a query box on a result page.] (*3 points*)

Information Retrieval: Problem 2. Evaluation

[Sakai16] refers to the following paper:

Tetsuya Sakai. 2016. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015, Proceedings of ACM SIGIR 2016.

Part 1

a Use no more three sentences to briefly explain how to compare two information retrieval methods/systems quantitatively using a test collection. (*1 point*)

b Use no more than two sentences to briefly explain why we need to do statistical significance test when evaluating a retrieval system. (*1 point*)

c Briefly explain how to do paired t-test when using Mean Average Precision to compare two retrieval systems. (*1 points*)

Part 2

According to [Sakai16], both "overpower" and "underpower" can be problematic. What is the problem of "overpower"? What is the problem of "underpower"? What can we do in each case to address the problem? Give an example of "overpower" and another example of "underpower", and use these examples to illustrate what you think we can do in each case to address the problem. (*3 points*)

Part 3

Since a Web search engine must respond to a user's query quickly, a scientist came up with an idea to speed up the search engine. The idea was to score only a subset of the "most promising" documents heuristically selected based on the term weights stored in the inverted index, thus avoiding touching all the documents that match at least one query term. The hypothesis was that such a new approach would be much faster than the baseline method (which scores all the documents matching at least one query term) without compromising much the retrieval accuracy.

a What kind of experiments do you think this scientist needs to do in order to test the hypothesis? (*2 points*)

b The scientist experimented with 20 queries to compare the new algorithm with the baseline in terms of the time required to respond to a user's query, and in every case, the new algorithm was faster. However, the scientist did not do statistical significance test. Is this a concern? (*2 points*)

Information Retrieval: Problem 3. Word embedding for query expansion

[Diaz16] refers to the following paper:

Fernando Diaz, Bhaskar Mitra and Nick Craswell, Query Expansion with Locally-Trained Word Embeddings, Proceedings of ACL 2016.

Part 1

a From the perspective of computation, what is the input and what is the output of a word embedding algorithm such as word2vec? (*1 points*)

b A word embedding algorithm such as word2vec solves an optimization problem. What is the objective function of the optimization problem (i.e., what does it attempts to optimize)? (*2 point*)

Part 2

a A main contribution of [Diaz16] is to propose a local word embedding. What is the key difference between local word embedding and global word embedding? Why is local word embedding expected to be better than global word embedding for information retrieval? (*2 points*)

b Query expansion with word embedding is achieved by interpolating the original query language model $p_q(w)$ with another expansion language model $p_{q^+}(w)$. Explain how to compute p_{q^+} exactly. (*2 point*)

Part 3

While word embedding has so far been mostly used for text analysis, similar ideas (i.e., the idea of embedding) can be applied to analyze social networks where we can learn a vector representation of each person involved in a social network. Propose some specific ideas for using an algorithm similar to word2vec for analyzing social networks such as Twitter or Facebook where we do not necessarily have text data, but we do have data with associations of people (e.g., A is a friend of B, or A follows B). Sketch the objective function to be optimized and discuss potential applications of the learned vector representations of people in a social network. (*3 points*)