DAIS Qualifying Examination

Fall 2015 (October 6, 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 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) (*Locking*)
When would you prefer two-phase locking over optimistic concurrency control and vice-versa?

(2) (*Query Processing*)
What are interesting orders and why are they important for Selinger's dynamic programming algorithm?

(3) (*Indexing*)
Inserts and searches in B+ trees are O(?) and O(?) respectively, where n is the number of keys. Fill in the ?.

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) (Pattern mining algorithms)

1. *name two algorithms that mine frequent itemsets effectively*, and
2. *name two algorithms that mine sequential patterns effectively.*

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

1. *RAINFOREST* vs. *random forest*, and
2. *active learning* vs. *transfer learning.*

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

1. *finding cancer patients that are best to be treated by a particular therapy*, and
2. *finding contaminated regions along the coast.*

Note: Please start with a new answer sheet.

(7) (*BM25*)

Suppose a search engine uses BM25 as the scoring function to rank all the documents in response to a query Q which has two terms t_1 and t_2. Suggest two distinct ways to increase the score of a particular document D by adding terms to D or deleting terms from D.

(8) (*Inverted index*)

Use no more than 3 sentences to explain what is an inverted index.

(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 four 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.

Database Systems: Problem 1

The following questions are about CrowdScreen: Algorithms for Filtering Data Using Humans.

Part 1

Which quantities does the paper optimize for when filtering data using the crowd? (*1 points*) Succinctly describe how worker errors are modeled in the paper. (*1 points*)

Part 2

Can you succinctly provide the intuition for why for every shape, there is a ladder shape that is at least as good or better? (Not looking for a detailed proof, just some intuition.)

(*2 points*)

Part 3

The paper makes the following assumptions about crowdsourcing. Identify potential solutions for each of these assumptions that retain a similar state space and probabilistic model to the one the paper adopts. Justify that your solutions work. Please state any assumptions you make clearly.

- The paper assumes a single worker type. What if you have two types of workers: expert workers and novice workers, each with different error rates and abilities, known in advance. (*2 points*)
- The paper assumes a single type of task. What if you have two different types of tasks: easy and hard, known in advance. That is, a computer can automatically tell if a task is an easy or a hard task. (*2 points*)
- The paper assumes a single type of task. What if you have two different types of tasks: easy and hard, but not known in advance. That is, a computer cannot automatically tell if a task is an easy or hard task. (*2 points*)

Database Systems: Problem 2

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

1. Data Loading
2. Joins
3. Compression
4. Failures
5. Transactions
6. Code Reuse

Just provide the answer — no explanation necessary. (*3 points*)

Part 2

Why does Vertica, which organizes data in columns, have faster query execution than DBMS-X, which organizes data in rows? (Two lines) (*2 points*)

Part 3

Design an efficient Map-Reduce algorithm for:

1. Joins between two relations, where both relations are large and distributed (*2.5 points*)
2. Joins between two relations, where one relation is small, and one is large (*2.5 points*)

Provide high-level pseudocode for the Map and Reduce functions. Syntax is not our concern here. (Hint: Borrow from ideas in parallel databases.)

Database Systems: Problem 3

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 Map-Reduce systems? (Two lines) (*1 points*)

Part 2

Prove that if the full consistency model is used, then GraphLab guarantees sequential consistency. (*3 points*)

Part 3

The consistency guarantee of sequential consistency provided by GraphLab is remarkably similar to serializability, in that the parallel execution is guaranteed to be similar to a sequential execution of any order on the vertices.

Let us consider a stronger consistency model: that of Bulk-Synchronous-Parallel (BSP), where all vertices perform an operation, and then globally synchronize, and then all vertices perform another operation, then globally synchronize, and so on. So communication occurs in "rounds", and no vertex is allowed to go to round i if some vertex is still in round (i - 1). Note that it is easy to implement BSP within the Map-Reduce paradigm.

In which cases would sequential consistency give the same eventual outcome as BSP? Answer with short justification (2 lines) is sufficient.

1. Pagerank computation, where each vertex accumulates the pagerank of all of its incoming edges and propagates it to its outgoing edges in each round. (*2 points*)

2. Connected components computation, where each vertex accumulates the counts of the smallest node id that it is connected to via any path, and then propagates it to its outgoing edges in each round. (*2 points*)

3. Shortest path computation, where each vertex accumulates its shortest path to all other vertices via other neighboring vertices, and then propagates it to its outgoing edges in each round. (*2 points*)

Data Mining and Data Warehousing: 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) Describe why ToPMine can mine phrases effectively from text corpora without any training data. (*2.5 points*)

(b) Compare the strength and weakness of two methodologies that generate phrase-based topic models: (1) first performing topic modeling (such as LDA) and then generating phrases for each topic, and (2) first performing phrase mining and then generating phrase-based topic models (*2.5 points*)

Part 2

(a) ToPMine does not use any labeled data in phrase mining. In reality, users may have some labeled data (e.g., phrases provided in a domain-specific dictionary, such as from a biomedical knowledge base). Describe how to extend ToPMine or develop a new algorithm to incorporate labeled data in the phrase mining process. (*2.5 points*)

(b) TopMine does not use any natural language processing methods in phrase mining. It may generate undesirable results such as short sentences. Discuss how to integrate methods derived from natural language processing to improve the phrase mining results. (*2.5 points*)

Data Mining and Data Warehousing: Problem 2

People nowadays move around with smart-phones, generating tremendous amount of mobility data.

Part 1

(a) Mobility sequential patterns can be generated, such as "many go to offices in the morning and then go to local restaurants for lunch, and then go back homes in late afternoon." Outline an efficient method that mines such patterns efficiently. (*2.5 points*)

(b) Suppose only a small portion of people keep their smart-phones on, and thus data so collected is sparse. Discuss how your algorithm should be revised in such a situation. (*2.5 points*)

Part 2

(a) Suppose text messages can be associated with smartphones. Outline an efficient method that may make credible friendship recommendation to smart-phone users based on their interests and their mobility patterns. (*2.5 points*)

(b) Suppose smartphone users will generate a lot of location-associated tweets (called *geo-coded tweets*). Design a stream data mining algorithm that can uncover *true bursty events* (e.g., not about jammed downtown traffic since it happens every day) happening around some locations in a city in real time. (*2.5 points*)

Data Mining and Data Warehousing: Problem 3

A biological network can be considered as a *heterogeneous information network*. It may contain information about genes connected with proteins, diseases, drugs, and biological pathways. Suppose one would like to construct such a network from a set of research papers from a biological literature database, such as PubMed Central, which contains a good set of research papers.

Part 1

(a) Discuss how such a heterogeneous biological information network can be constructed from research literature. (*2.5 points*)

(b) Discuss how to define *similar genes* effectively by exploring the concept of *meta-path*, and outline an efficient mechanism to implement it in large heterogeneous networks. (*2.5 points*)

Part 2

(a) If one wants to cluster genes and proteins, based on their relationships with diseases, outline an efficient method to implement it. (*2.5 points*)

(b) Discuss how to design an effective method to predict new functions of a drug based on the information provided in such a heterogeneous information network. (*2.5 points*)

Information Retrieval: Problem 1. Information Retrieval Models

[Paik 15] refers to the following paper:

J. H. Paik, A Probabilistic Model for Information Retrieval Based on Maximum Value Distribution, Proceedings of ACM SIGIR 2015.

Part 1

a According to [Paik 15], what is the elite set of a term? (*1 point*)

b The work [Paik 15] proposed a new model to improve existing retrieval models such as BM25. What limitation of the existing models does the new model address? (*1 point*)

c What is Maximum Value Distribution? (*1 point*)

Part 2

a The two building blocks of the new model proposed in [Paik 15] are

$$\begin{aligned} ritf(t,d) &= \frac{\log(1+tf(t,d))}{\log(k+mtf(d))} \\ lrtf(t,d) &= tf(t,d)\log(1+\frac{adl}{l(d)}) \end{aligned}$$

What is $mtf(d)$? (*1 point*)

b The final ranking function proposed in [Paik 15] is based on the following new normalized TF:

$$tff(t,d) = \alpha G(ritf(t,d)) + (1-\alpha)G(lrtf(t,d)).$$

where $G(ritf(t,d))$ has a probabilistic interpretation. (1) What probability does $G(ritf(t,d))$ capture? (2) Does $lrtf(t,d) > lrtf(t',d)$ always imply $G(lrtf(t,d)) > G(lrtf(t',d))$? (3) Does $lrtf(t,d) > lrtf(t,d')$ always imply $G(lrtf(t,d) > G(lrtf(t,d'))$? (*3 points*)

Part 3

a $ritf(t,d)$ and $lrtf(t,d)$ are related because of the connection between $mtf(d)$ and document length $l(d)$. Can you suggest an equation to connect $mtf(d)$ and $l(d)$? Suppose two documents $d1$ and $d2$ have identical length, but $d1$ has a higher number of unique words than $d2$. Assume a query term t occurred the same times in these two documents. Which document tends to have a higher $ritf$? That is, which is larger? $ritf(t,d1)$ or $ritf(t,d2)$? Why? (*2 points*)

b Briefly explain what you would store in the inverted index if you are to implement the proposed ranking function in [Paik 15] so as to minimize the time taken to respond to a query. (*1 point*)

Information Retrieval: Problem 2. Text Mining

[Yang et al. 15] refers to the following paper:

Z. Yang, A. Kotov, A. Mohan, S. Lu, Parametric and Non-parametric User-aware Sentiment Topic Models. Proceedings of ACM SIGIR 2015.

Part 1

a In a topic model, how is a topic usually represented? (*1 point*)

b What does "User-aware" mean in "User-Aware Sentiment Topic Models"? (*1 point*)

Part 2

a In [Yang et al. 15], the authors proposed 4 different topic models, USTM-FT(W), USTM-FT(S), USTM-DP(W), and USTM-DP(S). Briefly explain their similarities and differences. (*2 points*)

b What additional random variables have been introduced in USTM-FT(W) to extend PLDA (Partially Labeled Dirichlet Allocation) that was proposed in the previous work? (*2 points*)

Part 3

a The two main baselines used for comparison are Aspect-Sentiment Unification Model (ASUM) and Joint Sentiment-Topic Model (JST). However, it is unfair to compare the proposed models with these two baselines since they use different amounts of information. Elaborate the difference in the data used by the baseline models and the proposed models, and why this makes the comparison unfair. (*2 points*)

b The application problem of discovering segment-specific opinions can also be solved by simply using a basic topic model such as LDA multiple times to model different subsets of data. Briefly suggest a specific way of doing this to attempt to obtain similar results to what we can obtain using USTM-FT(W). Use no more than 5 sentences. (*2 points*)

Information Retrieval: Problem 3. Evaluation

[Bailey et al. 15] refers to the following paper:

P. Bailey, A. Moffat, F. Scholer, P. Thomas, User Variability and IR System Evaluation, Proceedings of ACM SIGIR 2015.

Part 1

a Briefly explain the process of using Cranfield evaluation method (i.e., test collection-based evaluation) to evaluate an information retrieval (IR) system. (*2 point*)

b According to [Bailey et al. 15], what is “external validity”? (*1 point*)

c According to [Bailey et al. 15], what is “user-generalizability”? (*1 point*)

Part 2

a Crowdsourcing was leveraged in the study of [Bailey et al. 15]. What exactly was a crowd worker asked to do? (*2 points*) Use no more than 5 sentences.

b What is the main finding about user variation made in [Bailey et al. 15]? (*1 point*)

Part 3

Suppose you are to evaluate a social media search engine to enable users to find interesting news and explore related people to an event. Based on the findings and recommendations made in [Bailey et al. 15], how would you design the test collection? In particular, how would you design the queries? (*3 points*)