DAIS Qualifying Examination

Spring 2018 (1pm-4pm, February 26, 2018)

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.

DB - Problem 1 + 3
DM - Problem 1 + 2
IR - Problem 2

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) (*Relational Algebra Understanding*) Consider a relation $R(A, B)$ that contains r tuples, and a relation $S(B, C)$ that contains s tuples; assume $r > 0$ and $s > 0$. Make no assumptions about keys. For each of the following relational algebra expressions, state in terms of r and s the minimum and maximum number of tuples that could be in the result of the expression.

1. $R \cup \rho_{S(A,B)} S$
2. $\Pi_{A,C}(R \bowtie S)$
3. $\sigma_{A>B} R \cup \sigma_{A<B} R$

(2) (*Updates to Views*) Updates to materialized views may or may not be automatically translatable to the underlying base relations. Provide one example query when it can be directly translated by the database system, and one where it can't.

(3) (*Query Processing*) Which of the following features would you least expect to find in a NoSQL system, and why?

- Extreme scalability
- Simple processing based on key values
- Serializable multi-statement transactions
- Fault-tolerance

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 concepts) Use one or two sentences to distinguish the following pairs of concepts or methods:

1. *active learning* vs. *lazy learning*, and
2. *transfer learning* vs. *distantly supervised classification.*

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

1. *effectively and incrementally cluster dynamic data streams*, and
2. *mining quality phrases in a large corpus with minimal human labeling effort.*

(6) (Concepts on deep learning)

1. *Name two key points that deep learning distinguishes from traditional machine learning methods.*

2. *Name three typical deep learning frameworks (or architectures).*

Note: Please start with a new answer sheet.

(7) (*TF weighting*)

Suppose we compute the document vector for a technology news article in a collection of general news articles using **only TF weighting** without stop-word removal. Which of the following words do you expect to have the highest weight? Why? (A) "computer" (B) "the" (C) "science"

(8) (*PageRank*) Briefly describe how PageRank can be used to analyze Twitter social network where the nodes are users and edges indicate a "follow" relationship between users. What kind of users do you expect to have the high PageRank scores? Why? (Use no more than three sentences to answer this question.)

(9) (*Evaluation*) Let $R = (-, +, -, +, -, -)$ be the relevance status of 6 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, and the second document 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.

Database Systems: Problem 1

The paper "A Comparison of Approaches to Large-Scale Data Analysis" performs a comparative study of MapReduce with traditional parallel database systems.

Part 1 (*4 points*)

Say we have three relations $R(A, B)$, $S(B, C)$, and $T(C, D)$ stored in a partitioned manner in a data center. We want to perform a natural join of R, S, and T.

How would we do it via Map-Reduce? (Hint: you are free to use multiple rounds of Map-Reduce) Explain what the Map step(s) would do, and what the Reduce step(s) would do. (Pseudocode is not needed, but try to be as clear as possible.)

Part 2 (*2 points*)

Explain two advantages declarative queries (e.g., in a language like SQL) have over imperative code (e.g., in a program like Python), and two advantages imperative code have over declarative queries.

Part 3 (*4 points*)

Explain under which scenarios a row-store would win out against a column-store, when operating on a relation R, and why. R has 100 columns $c_1, \ldots, c_{100}$.

- The only selection query run on R is a `SELECT *` query, with no modifications being done to R.
- The only selection query run on R is a `SELECT` c_1, c_2 query, with no modifications being done to R.
- The only selection query run on R is a `SELECT` c_1, c_2 query. In addition, there are only frequent point-wise modification queries to attribute c_3, based on the value of attribute c_1 (e.g., set c_3 to be 100 when c_1 = "Tom".)
- The only selection query run on R is a `SELECT` c_1, c_2 query, run infrequently. In addition, there are much more frequent modification queries that add or delete tuples.

Database Systems: Problem 2

The "BlinkDB" paper presents a solution for approximate query processing.

Part 1 (*7 points*)

Say you have selected a query column set (QCS) $S = \{State, Month, ProductCategory\}$ with distinct values 50, 12, and 10 respectively for each attribute, to create stratified samples on, on Relation

$R(State, Month, ProductCategory, Product, City, Year, TransactionID, SaleValue, ProfitValue)$.

What is the worst case number of strata you will need to maintain samples for? What is the best case?

Assume that the number of samples per strata that you store is large. For each of the following queries, describe whether or not the samples on S help answer the query satisfactorily, or partially, or not at all, and justify your answer. (The `FROM R` part of the queries is omitted for succinctness.)

- `SELECT Product WHERE State = "MA"`
- `SELECT * WHERE State = "MA" AND Month = "12"`
- `SELECT AVG(SaleValue) WHERE Month = "12" GROUP BY State`
- `SELECT MAX(ProfitValue) GROUP BY ProductCategory`
- `SELECT SUM(SaleValue) WHERE Month = "12" AND State = "MA" AND Year = "2016"`
- `SELECT AVG(ProfitValue), Month GROUP BY State, Month HAVING Month > 2`

Part 2 (*1 points*) The BlinkDB approach of maintaining QCSs may not work for all interactive exploration settings. Explain why.

Part 3 (*2 points*) Explain when the BlinkDB approach would win out against a materialized data cube approach for OLAP, and when the data cube approach would win out against BlinkDB.

Database Systems: Problem 3

The "Pregel" paper introduces a new graph processing framework.

Part 1

Why is the Map-Reduce paradigm ill-suited for graph processing? (*1 points*)

Part 2 Explain how one could compute (undirected) connected components via Pregel? Pseudocode is not needed, a few lines of intuition is sufficient. (*2 points*)

Part 3 (*5 points*)

One way to improve the latency of Pregel is to not insist that all of the workers are proceeding in "lock-step", i.e., one worker can be in super-step i, while the other is in $i - k$, for some $k > 0$. That is, we do not wait until all of the workers are in sync with each other. In which of the following problems (described in the paper) could that lead to (a) results that are different from the synchronous execution, and if different (b) whether the results could be erroneous? Explain why.

- Pagerank
- Shortest Path
- Bipartite Maximum Matching

Part 4 (*2 points*) In assigning vertices to partitions, Pregel uses an arbitrary hash function that is based on the worker ID. Why could this lead to unnecessary network overhead? How would you improve it?

Data Mining: Problem 1

A biological network can be considered as a *heterogeneous information network*, where information about one type of entities, such as diseases, can be connected with other types, such as genes, proteins, drugs, and biological pathways. Suppose one would like to construct such a network from a set of research papers in a biological literature database, such as PubMed Central, which contains a large set of research papers.

Part 1

(a) Discuss how to automatically extract biological entities from biological text, based on some available but incomplete information (e.g., available genes, proteins, diseases, drugs) (*2.5 points*)

(b) One may like to automatically extract relationships among different typed entities. However, there are many different ways to express similar or rather different relations. Outline methods that may lead to high quality extraction of relations among such entities. (*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 clustering. (*2.5 points*)

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

Data Mining: Problem 2

Social media (e.g., tweets) may generate a lot of geo-coded social communications that provide rich information for data mining.

Part 1

(a) In reality, only a small portion of tweets are geocoded. Outline a method that may utilize both geocoded and non-geocoded tweets to effectively discover the true locations of (some) non-geocoded tweets. (*2.5 points*)

(b) With massive geocoded and non-geocoded tweets, one may be able to mine unusual events happening in a local region (e.g., a city center). Outline an efficient method that mines such unusual events from a mixture of geocoded and non-geocoded tweets effectively. (*2.5 points*)

Part 2

(a) In many cases, tweets are short and contain lots of abbreviations or acronyms. On the other hand, news and some other media may contain sufficient information for newcomers to understand. Discuss how news and tweets may complement each other and outline a mining method that may realize an effective joint mining. (*2.5 points*)

(b) Outline an efficient method that may use geocoded tweets to discover *periodic events* in a local region. (*2.5 points*)

Data Mining: Problem 3

Assume that researchers at Facebook plan to analyze activities of Facebook users to identify trends. Assume further, that besides knowing the social graph $G = (V, E)$ of friendships, we have for any user i, access to a set of time stamped actions a_t^i (e.g. create, comment, link) on entities e_j, where user j created the entity e_j. The first goal for the researchers is to identify clusters of similar users, and then inspect these clusters to make recommendations to the CEO.

Part 1

Assume that the researchers first analyze a *static* snapshot of the data.

(a) Since the transactions database is huge (over several trillion records), the researchers decide to take a small subset of the database to analyze. How should they go about sampling the database? (*2.5 points*)

(b) Identify two challenges to clustering the sampled data. Outline your plan to tackle these challenges to effectively cluster the data. (*2.5 points*)

Part 2 Once the researchers have identified a suitable sampling and clustering algorithm, they now turn their attention to streaming data (i.e. the real-time stream of user activities). They want to update the clusters in real-time.

(a) How will you update the sampling algorithm that you proposed in part 1 to deal with the stream? You are still limited by the problem of finite sample size; that is, your clustering algorithm needs a bounded sample for analysis. (*2.5 points*)

(b) Update your proposed clustering algorithm to handle the streaming case keeping in mind the two challenges to clustering identified earlier. (*2.5 points*)

Information Retrieval: Problem 1. Information Retrieval Models

[Dehghani et al. 17] refers to the following paper:

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In Proceedings of ACM SIGIR 2017.

Part 1

a According to [Dehghani et al. 17], what does "weak supervision" mean? Based on this definition, what do you think "strong supervision" would mean in the case of using machine learning for information retrieval? *(1 point)*

b Training a neural network generally involves solving an optimization problem. How is the objective function for this optimization problem usually defined? *(1 point)*

c How many parameters are there in a typical neural network with n input nodes, m output nodes, and one hidden layer with h nodes? *(1 point)*

Part 2

a In [Dehghani et al. 17], the authors proposed a way to generate pseudo-training data using a retrieval function such as BM25. Use no more than 3 sentences to explain how this works. *(1 point)*

b According to [Dehghani et al. 17], what is the main advantage of using an embedding vector representation as compared with a regular word vector representation? *(1 point)*

c In [Dehghani et al. 17], the authors exerimented with three different ways to compute the embedding vectors. What are those three ways? Which one worked the best? Can you explain why this particular method worked better than other methods? *(2 points)*

Part 3

The authors of [Dehghani et al. 17] compared their neural ranking models with BM25 and found that their models outperform BM25 substantially. However, one may argue that their models used some extra information that the baseline BM25 did not have access to, thus in some sense, this comparison is not really fair. Can you explain what this extra information is? Can you suggest another stronger baseline than the simple BM25 that would also use this extra information and thus make the comparison more fair? (Hint: what do you think about pseudo feedback based on BM25?)

(3 points)

Information Retrieval: Problem 2. Text Mining

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

Lau, Raymond YK, Yunqing Xia, and Yunming Ye. A probabilistic generative model for mining cybercriminal networks from online social media. IEEE Computational intelligence magazine 9.1 (2014): 31-43 .

Part 1

a What are the advantages of representing a topic as a word distribution over representing a topic by just one term? Use no more than 3 sentences to answer this question. (*1 point*)

b The paper [Lau et al. 14] proposed an algorithm to mine cybercriminal networks from online social media data. What are the input and output of this algorithm, respectively? (*2 points*)

Part 2

a Latent Dirichlet Allocation (LDA) is used as a baseline method in [Lau et al. 14]. Briefly explain how LDA works. What output can LDA generate as a text mining algorithm? (*2 points*)

b The authors of [Lau et al. 14] proposed an improved version of LDA called Context-Sensitive LDA (CSLDA) and the experiment results show that CSLDA improves over LDA. Why do you think CSLDA has outperformed LDA? Can you see any disadvantage of CSLDA as compared with LDA? (*2 points*)

Part 3

The complete algorithm proposed in [Lau et al. 14] includes many steps and multiple parameters to be set.

a Name at least two major parameters and briefly explain how each is set in the experiments. (*1 points*)

b Give at least two reasons why a method with too many parameters to be manually set is not as good as one that has fewer parameters. (*2 points*)

Information Retrieval: Problem 3. Evaluation

[Zhang et al. 17] refers to the following paper:

Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating Web Search with a Bejeweled Player Model, Proceedings of ACM SIGIR 2017.

Part 1

a According to [Zhang et al. 17], what is a "system-oriented test" (i.e., Cranfield evaluation methodology)? Briefly describe how it works and how to construct a test collection for evaluating a retrieval system. (*2 points*)

b According to [Zhang et al. 17], what is a "user-oriented study"? Point out one advantage of a user-oriented study over a system-oriented test. (*1 point*)

Part 2

a According to [Zhang et al. 17], the game Bejeweled and the task of Web search are similar because they both have two final states. What are those two final states in Web search? Suppose a user is searching for information to answer the question "what does SVM stand for?". Describe exactly what is each of the two possible final states. (*2 points*)

b The proposed dynamic BPM metrics in [Zhang et al. 17] have two parameters h_B and h_C. For what kind of users should we set each to a higher value? (*2 point*)

Part 3

The authors of [Zhang et al. 17] mentioned the following direction for future work: "We make some simplied assumptions for Static BPM Metrics and Dynamic BPM Metrics. In the future work, we plan to explore more complex situations for them. For instance, we can consider upper limits as latent variables and estimate them with large scale user logs." Suggest some specific ideas for estimating the upper limits using search logs. Sketch your ideas with formulas if you can.

(*3 points*)