

DAIS Qualifying Examination
Spring 2005

Department of Computer Science
University of Illinois at Urbana-Champaign

Feb. 14, 2005

Time Limit: 180 minutes

- You are required to answer *six* questions. If you answer more than six questions, the committee will randomly select six to grade.
- Out of the six questions answered, at least *two* must be from the area of *Database Systems*. The remaining questions will be your choice. The distribution of the questions by areas is as follows.
 - **Database Systems:** 5 questions
 - **Data Mining:** 3 questions
 - **Information Retrieval:** 3 questions
- For each question, there will three subquestions. The first is to test the basic knowledge related to a topic, the second is to test current research related to the topic, and the third is to test research capability.
- To ensure the fairness of the grading, all the exams will be kept anonymous. Your answer sheets should 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. *Simplicity does count!*

Database Systems: Problem 1

Part 1

It is said that effective support of *access methods* is key to moving the implementation of DBMS from navigational access to query-based access.

- (a) What does an access method provide?
- (b) Explain why access methods are critical for supporting query-based access, as compared to navigational access.
- (c) Name three different access methods for a relational DBMS, and briefly explain what they are.

Part 2

B-tree is a common access method used since the early days of RDBMS implementation.

- (a) Give two reasons why B-tree is attractive in RDBMS.
- (b) Can B-tree help in processing a *join* operation? Explain.

Part 3

In the development of object-oriented or object-relational DBMS such as Postgres, queries with “nested dot” notation are supported, e.g., the query below find all the employees managed by Joe. Such queries are processed using *object ID* or *OID* as a new access method.

retrieve (EMP.name) **where** EMP.manager.name = “Joe”

- (a) Explain why using *OID* is desirable for efficient query processing.
- (b) What are the complications of using or maintaining *OID* as an access method? Give two.
- (c) How can you ask this query in a *relational* DBMS? How will you process it? Explain concretely by 1) giving an equivalent SQL query and 2) describing how it can be processed, and identifying what access methods are used.

Database Systems: Problem 2

Part 1

Query optimization has been considered as a key technique for the realization of the relational model.

- (a) Why does the relational DBMS (in particular) need query optimization? Why this was not an issue for earlier DBMS (e.g., of the network model).
- (b) System R has established the “Selinger-style” query optimization. What are the main techniques in this framework?

Part 2

However, the research on query optimization did not stop at the “Selinger-style” optimizer. Name *two* query optimization *ideas* that were not considered in the System R framework.

Part 3

System R query optimization employs the *Principle of Optimality*. In essence, the principle means that “*components of a globally optimal solution are themselves optimal.*”

- (a) How does this principle help in optimization?
- (b) Can you design a query optimizer without relying on this principle? If *no*, briefly explain why. If *yes*, briefly explain how.

Database Systems: Problem 3

Part 1

1. Briefly compare undo logging, redo logging, and undo/redo logging. When should one use which scheme?
2. List all possible sequences of states through which a transaction may pass (until it commits or aborts). Explain why each state transition may occur.

Part 2

1. Briefly explain the notions of conflict serializability and view serializability. Since every conflict-serializable schedule is view serializable, why do we emphasize conflict serializability rather than view serializability?
2. Briefly describe a polynomial algorithm that has been developed for testing conflict serializability? How about a polynomial algorithm for testing view serializability?

Part 3

1. Briefly discuss how transaction management and concurrency control change when applied to database systems that manage XML data. Discuss what new issues may arise and sketch your proposed solutions. Give brief examples if appropriate.

Database Systems: Problem 4

Part 1

1. Compare the global-as-view (GAV) and local-as-view (LAV) approaches to data integration. Describe a brief data integration example and use it to illustrate your comparison.
2. In practice do you think when people would use which of the above two approaches? Can you use both?

Part 2

1. Discuss the basic ideas behind adaptive query processing. Why is it necessary in data integration contexts? Name at least two groups that have worked on adaptive query processing.
2. Briefly compare and contrast the following three notions: federated database systems, data integration systems, and peer-to-peer data management systems.

Part 3

1. Given two relational tables $T(a,b,c)$ and $S(x,y,z)$, where a, b, c, x, y, z are attributes. Assume each table has 100 associated tuples. Describe a simple scheme for matching the schemas of the two tables: that is, to find semantic correspondences between the attributes of T and S .
2. Consider again the above two tables T and S . Now assume you have 50 more tables U_1, U_2, \dots, U_{50} in the same application domain (e.g., real estate), each of which has 100 tuples. Can you exploit these tables to improve the accuracy of matching between tables T and S ? If yes, briefly describe a scheme to do this.

Database Systems: Problem 5

Consider an XML document and the query (or query fragment) 'chapter1/*/table', which asks for all tables that are located in any part of chapter 1.

Part 1

Propose a simple algorithm for processing queries of this form, and give its time complexity. What are the shortcomings of your algorithm likely to be in practice?

Part 2

In an XML tree, a node n_1 is an ancestor of a node n_2 if and only if n_1 occurs before n_2 in the preorder traversal of the tree, and occurs after n_2 in the postorder traversal of the tree. Based on this insight, propose a more sophisticated algorithm for processing queries of this form, and give its time complexity. Your new algorithm should overcome some or all of the shortcomings of the algorithm that you presented earlier. What are the strengths and weaknesses of your new algorithm? Give its time complexity.

Part 3

Generalize your algorithm from Part 2 to allow queries of the form ' $p_1/p_2/ \dots /p_n$ ', where each p_i can be either an element name or a wild card. What are the strengths and weaknesses of your new algorithm? Give its time complexity.

Data Mining and Data Warehousing: Problem 1

Part 1

Assume a base cuboid of 10 dimensions contains only five (nonempty) cells in the base cuboid. The measure of the cube is *count*. Answer the following questions and provide short reasoning on your answers.

- (a) What is the possible **maximum** number of nonempty cells that the whole data cube may contain (excluding the base cuboid)?
- (b) What is the possible **minimum** number of nonempty cells that the whole data cube may contain (excluding the base cuboid)?

Part 2

There are three kinds of measures that a data cube may compute: *distributive*, *algebraic*, and *holistic*.

- (a) Give one example for each such category.
- (b) Suppose one would like to compute a sales iceberg cube of 10 dimensions, with the following iceberg condition:

$$avg_price(*) > 50 \wedge count(*) \geq 100$$

Outline an efficient algorithm that may compute such an iceberg cube efficiently.

Part 3

Data cube facilitates online analytical processing of multi-dimensional data. Such kinds of analysis may also be desirable in data streams, where data may flow in and out indefinitely.

- (a) What are the major challenges for multi-dimensional analysis of data streams?
- (b) Proposed a design that may facilitate multi-dimensional analysis of data streams efficiently.

Data Mining and Data Warehousing: Problem 2

Part 1

For mining transactional or transaction-sequence databases, scalable algorithms have been developed for mining

- (a) *frequent itemsets*,
- (b) *closed (frequent) itemsets*, and
- (c) *(frequent) sequential patterns*.

Name three algorithms for each of the above three mining tasks.

Part 2

Association rules can be derived from frequent itemsets. However, not all of the strongly associated itemsets are positively correlated.

- (a) *Lift* has been used as a correlation measure. Present one concrete example to show that *lift* can be used to identify some strongly associated itemsets that are not positively correlated.
- (b) However, in general *lift* may not be a good measure to test correlation among itemsets. Present an example to show this point. What correlation measure do you suggest then?

Part 3

Association can be used for the classification task.

- (a) Propose a method that performs effective classification by association analysis.
- (b) In comparison with traditional decision-tree induction method, discuss efficiency and effectiveness of your proposed approach.

Data Mining and Data Warehousing: Problem 3

Part 1

For effective clustering, different data types may need to use different similarity (distance) measures. For each of the following types of data, present one distance measure that is often used for the generation of high-quality clusters.

- (a) numerical data
- (b) asymmetric binary data
- (c) vector objects (e.g., text documents)
- (d) categorical data

Part 2

Micro-clustering has been used for scalable clustering of massive data sets.

- (a) Taking one clustering task as an example, explain how micro-clustering can be used for effective clustering of large data sets.
- (b) If a data set is partitioned and distributed at multiple sites, propose a k -means like algorithm that can perform efficient and effective clustering of the whole data set without moving all the data sets to one site.

Part 3

In biological data analysis, many data sets, such as microarray data sets, contain a large number (such as thousands) of dimensions.

- (a) State what are the major challenges at clustering high-dimensional data?
- (b) Propose an efficient algorithm that performs efficient and effective clustering of the high-dimensional data in micro-array data analysis.

Information Retrieval: Problem 1. Text Categorization

Part 1 Decide whether the following statements regarding the multi-variate Bernoulli model and the multinomial model for the Naive Bayes classifier are true. Circle your answer.

- a true false
The number of parameters in both models grows linearly to the vocabulary size of the language.
- b true false
Removing a word from a document will change the probability of the document according to a multinomial model, but not necessarily if the model is multi-variate Bernoulli.
- c true false
In the paper “A comparison of event models for naive Bayes classification” (McCallum & Nigam 1998), the multinomial model has been shown to outperform the multi-variate Bernoulli model when the vocabulary is large.

Part 2

- a Text classification literature often reports a breakeven value as a measure of classification accuracy. What is the precision-recall breakeven point? How do we interpret a breakeven value of 0.8?
- b Suppose we have k documents $\{d_1, \dots, d_k\}$ known to be in category C , and the vocabulary size is $|V|$. Assume that we use a multinomial event model. Use V to denote the set of words in our vocabulary and $c(w, d_i)$ to denote the counts of word w in document d_i . Write down the formula for estimating the probability of a word w given by the multinomial model for category C (i.e., $p(w|C)$) with Laplacean smoothing. In this case, how much probability mass in total has been given to unseen words (i.e., the words that have never occurred in any of our k documents)?

Part 3

- a Both the multivariate Bernoulli model and the multinomial model make an *independence* assumption about word occurrences. Explain what is the difference between the assumptions made in these two models with two sentences. For each of the two models, give a concrete example of English words to show that the Independence assumption made is not really true in reality. Again use no more than two sentences for each example.
- b Can you sketch a categorization method that uses Gaussian distributions (i.e., normal distributions) to model the frequency of words? Use $N(\mu, \sigma^2)$ to denote a Gaussian distribution. If we have n categories and $|V|$ words in the vocabulary, how many parameters would we need to estimate in total? Do you expect such a model to perform well? Why?

Information Retrieval: Problem 2. Language Models for Retrieval

Part 1

A common way of using language models for retrieval is to rank documents based on $p(Q|D)$, where Q is a text query, D is a text document, and $p(Q|D)$ is computed based on a unigram language model. Given a particular query Q , what kind of documents would give the highest and the lowest value for $p(Q|D)$, respectively? Why?

Part 2 According to one translation model proposed in [Berger & Lafferty SIGIR 99] (i.e., Model 1), the score of document d w.r.t. query q can be computed as

$$p(q|d) = \psi(m|d) \prod_{j=1}^m \left(\frac{n}{n+1} p(q_j|d) + \frac{1}{n+1} t(q_j | < null >) \right)$$

where $p(q_j|d) = \sum_w t(q_j|w)l(w|d)$.

- Briefly explain the meaning of $t(q_j | < null >)$ and how to estimate it.
- How should we estimate $l(w|d)$?
- $t(q_j|w)$ is the so-called translation model. Briefly explain how this model can help solve the problems of polysemy and synonym in retrieval.

Part 3

Ranking documents based on $p(Q|D)$ can be justified using the following formula

$$p(D|Q, U) \propto p(Q|D, U)p(D|U)$$

where U is a user variable. That is, our original goal was to rank documents based on $p(D|Q, U)$, the probability that D is relevant given that the user is U and the query is Q . If we assume that $p(D|U)$ is uniform, we essentially will rank documents based on $p(Q|D, U)$. Normally, without any further information available, we would simply estimate $p(w|D, U)$ based on the relative frequency of word w in document D , and then use $p(w|D, U)$ to compute $p(Q|D, U) = \prod_{w \in Q} p(w|D, U)$.

Now suppose we have a lot of interaction history of the user U in the form of (Q, D, U) , which means a user U issued a query Q and viewed document D in the results of Q . (Imagine we have a set of tuples $H = \{(Q_i, D_i, U_i)\}$.) We can exploit such history information to potentially improve retrieval performance.

- Instead of assuming $p(D|U)$ to be uniform, we can estimate $p(D|U)$ based on the interaction history H of the user U , which presumably can help improve retrieval performance. Suggest one way to do this. Give a specific formula for $p(D|U)$ if possible.
- Another possibility to improve the performance is to improve our estimate of $p(w|D, U)$ by exploiting the history information H . Suggest a formula to estimate $p(w|D, U)$ based on both the relative frequency of w in D and the history information H .

Information Retrieval: Problem 3. XML and PageRank

Part 1

- a List the names of at least 4 different XML query languages.
- b In the paper “Querying XML Data” (Deutch et al 1999, IEEE Bulletin on Data Engineering), the authors argue that five query operations should be supported by an XML query language. Write down the names of these five query operations.

Part 2

The PageRank algorithm can be described as follows:

$$\vec{R} = (1 - \alpha)M\vec{R} + \alpha\vec{p}$$

where \vec{R} is a vector of importance values for all pages (i.e., rank values), and M is a link matrix defined based on the web graph.

- a What kind of pages can be expected to have the highest PageRank scores? Suppose you are allowed to add precisely one link to web page A from another web page B . How do you choose page B to maximize the PageRank score of page A ?
- b Suggest *two* distinct methods to make the standard PageRank algorithm described above query-sensitive. For each method, sketch the modified PageRank formula.
- c Suppose a user is willing to make relevance judgments on some initial retrieval results (i.e., labeling some documents as relevant and some others as non-relevant). Suggest one way to modify PageRank so that we can exploit such feedback information to improve PageRank scoring.

Part 3

Discuss the difference between an XML data collection and an ordinary unstructured text collection. How does this difference affect our design of XML query languages? Explain why neither the standard SQL nor the simple keyword text query is ideal for searching an XML data collection. A brief explanation (about two sentences) is sufficient for each point.