uc berkeley school of information management & systems

SIMS 202 Information Organization and Retrieval

Final Exam Preparation Guide, Fall 2004

The SIMS 202 final exam will take place Tuesday December 14 from 9:30 am - 12:30 pm in 202 South Hall. This will be an open-book, open-note and open-computer exam, you may use your own laptop, or one of the machines in the computer lab (Room 210). You can also write the exam by hand if you wish, so please bring your own pen/pencils (It's OK if your handwriting isn't great). We'll supply the paper as part of the exam itself. Each person will work individually. The exam period is three hours; you will likely need the entire time. If you use network-accessed material for any part of the exam be sure to cite your sources.

The exam is comprehensive, meaning it will cover both parts of the class. However, the emphasis will be on materials covered in the second half of the course.

Each question will be worth an indicated number of points. Partial credit will be awarded. In your answers, please balance conciseness with illustration of all of the requested information. (In other words, don't write a lot of things that aren't asked for, but try to address **all** of what is asked for.) Make sure to answer all questions since we do award partial credit.

To study for the exam:

- Be sure you understand the material that was covered in lecture and have read and absorbed the corresponding material in the readings.
- Be sure you can do activities similar to what was done in the homework assignments.
- We will have questions that require you to generalize from what you've learned and to synthesize ideas. So be sure you have thought about the ideas covered in lecture, readings, and homework assignments.

These ideas and abilities should be at your fingertips. There won't be time during the exam to do a lot of catch-up reading on topics you haven't studied.

Below are shown the major topics we've covered in the course, and some example questions. Please note that these are *examples of the types of questions we will ask*. The second half of the course will be emphasized in the exam, but there will also be questions related to materials from the first half. The example questions are (probably) **not** the *exact* questions we will ask. We will

probably ask some other types of questions too, in particular the kind where we give you an example of some information and ask you to do something with it (design an ER diagram, convert to a hierarchy, etc.).

Information Retrieval Topics and Example Questions (1st half – *same contents as midterm study guide*)

- *Topic:* **Information**
- *Example Questions:*

  What is the information life cycle?

  What are different ways of measuring information? What are different ways of defining information?

- *Topic:* **Document Representation and Statistical Properties of Text**
- *Example Questions:*

  What is the significance of Zipf's law for weighting of terms in information retrieval?

  What kinds of errors can a stemming algorithm produce?

- *Topic:* **Queries, Ranking, and the Vector Space Model**
- *Example Questions:*

  What is the difference between a search engine that uses the vector space ranking algorithm on natural language queries and a system that uses Boolean queries?

  What is the role of coordination level ranking in a faceted Boolean system?

  Describe the following information need in terms of a faceted Boolean query. What kinds of weighting algorithms can be applied to a faceted query like this? “I would like to find articles about the effects of the passage of the independent investigator statute by Congress on how the U.S. president chooses an attorney general."

  Why do different web search engines return different sets of documents for the same query?

  Redo the computations of Assignment 3 part 3 using different values for TF.

- *Topic:* **IR systems and Implementation**
- *Example Questions:*

Draw and label a diagram that shows the major components of an IR system.

What are the special features of the Cheshire II information access system?

What is the purpose of an inverted index? How is it used to generate answers to Boolean queries?

Convert the contents of a set of documents (short texts) into an inverted index representation.

- *Topic:* **Evaluation of IR Systems**
- *Example Questions:*

Define precision. Define recall. Define relevance. How are the three interrelated?

Under what circumstances is high recall desirable? Under what circumstances is high precision?

What is the main purpose of TREC? How does it differ from earlier evaluation efforts?

- *Topic:* **The Search Process and User Interfaces**
- *Example Questions:*

Search and retrieval is part of a larger process. Name some other components of that process.

How/why doesn't the Bates berry-picking model fit with the standard information retrieval model?

How (fundamentally) does search on a directory system like Yahoo differ from search on Altavista or Google?

- *Topic:* **Relevance Feedback**
- *Example Questions:*

What is main the difference between relevance feedback as defined in the literature and the more current web-based notion of "more like this"?

Given a query, three documents marked as relevant, and the Rocchio formula for relevance feedback given in class, compute the vector for the new query that results.

The Koenemann & Belkin study found results in three conditions for relevance feedback: opaque, transparent, and penetrable. Consider the different ways people have recently implemented systems for predicting which web page to show the user next. How do the differences in these systems correspond to the different relevance feedback

- *Topic:* **Database Design**
- *Example Questions:*

  How is a database different than a file system?

  What are the benefits of a database system?

  What do we mean by data independence?

  What are the benefits/drawbacks of the primary database models?

  Entity-Relationship Diagrams -- what are they for, how do you create them?

  How do you normalize a relational model database?

  What is a join?

Information Organization Topics and Example Questions (2nd half)

- *Topic:* **Categorization**
- *Example Questions:*

  What is the definition of class membership in traditional categorization? How does traditional categorization have difficulty describing certain phenomena, like games (give 1 other example besides games)?

  What is the "basic level" in categorization and how is it psychologically primary? How might the use of basic level categorization affect the design and use of information systems?

- *Topic:* **Knowledge Representation**
- *Example Questions:*

  What limitations in standard information retrieval do knowledge representation technologies try to overcome? What challenges do they face in the attempt?

  What are the similarities and differences between commonsense knowledge representation systems like CYC and facetted metadata

classifications like the Art and Architecture Thesaurus or the facetted classification you built (give three examples)?

- *Topic:* **Lexical Relations and WordNet**
- *Example Questions:*

  What are three lexical relations in WordNet that would be useful in an information retrieval task (explain how and give examples)?

  Where are the meanings of the words in WordNet? How would assuming the conduit metaphor vs. the toolmakers' paradigm of communication lead you to different answers to this question?

- *Topic:* **Controlled Vocabularies**
- *Example Questions:*

  What does Svenonius consider to be the primary difficulties with using controlled vocabularies?

  What is the purpose of authority control? Is this a type of controlled vocabulary? Why or why not?

- *Topic:* **Semantic Web and RDF**
- *Example Questions:*

  What are the different basic topological structures of XML and RDF? What benefits and problems do these respective structures offer for information organization and retrieval?

  What is the Semantic Web effort trying to accomplish? What challenges does that effort face and how might they be overcome?

- *Topic:* **Facetted Classification and Thesaurus Design and Construction**
- *Example Questions:*

  What are the differences between classical and faceted classification and how do these differences affect the design and use of information systems?

  How is a classification scheme or a thesaurus designed?

- *Topic:* **Metadata Standards**
- *Example Questions:*

  What are the motivations behind creating and using metadata systems like Dublin Core, MARC, AACR II, etc.?

How do metadata standards come about and how might their provenance affect their adoption?

- *Topic:* **Multimedia Information Organization and Retrieval**
- *Example Questions:*

  What is the "Kuleshov Effect" and how might it affect the design of metadata for multimedia data?

  What are the "semantic gap" and the "sensory gap" and what challenges do they present for the design of information systems for multimedia data?

- *Topic:* **Metadata for Motion Pictures: Media Streams and MPEG-7**
- *Example Questions:*

  What limitations do keywords pose for multimedia information retrieval and how might those limitations be addressed?

  What aspects of multimedia content description is MPEG-7 attempting to standardize?

- *Topic:* **Mobile and Context-Aware Multimedia Information Systems**
- *Example Questions:*

  How are cameraphones distinguished from traditional digital cameras in their technological capabilities and use (give 5 examples)?

  What and how could contextual metadata be useful in describing and retrieving information (give 4 examples)?

- *Topic:* **Looking Backward Looking Forward: Future of Information Systems**
- *Example Questions:*

  How are Bush's vision of the Memex and the current World Wide Web similar and different (explain two similarities and two differences)?

- *Topic:* **Project Presentations**
- *Example Questions:*

  In revising your facetted metadata ontology how did you increase its expressiveness and reusability (give 3 examples)?

  How well would the ontology you and your partner group designed support one of the other mobile media metadata applications presented by your classmates?