

Kristin Yiotis  
Final Exam  
202 Information Retrieval  
Apring 2004 - Weedman

Please type or cut-and-paste your answers directly into this e-mail rather than sending as an attachment. Thanks!!!

For all questions, be careful to express your answer in your own words rather than simply quoting your notes, handouts, or the text. It is NOT expected that you will need to go to any other reference books beyond the required readings for the class.

Part A.

Short answer. Answer 3 of the following 5 questions. Answers should be 2-4 sentences in length (take this limit seriously!). Each question is worth 20 points. Don't answer more than 3 questions; any additional answers will not be graded.

1. Define recall and precision and explain why they matter.

Precision is the function of discrimination. It is a measure of the degree to which the database is able to discriminate documents that are relevant from the ones that are not. Recall is a function of aggregation. It is a measure of the degree to which the database is able to aggregate all the documents that are relevant to your search.

2. Why is relevance an important concept in information retrieval?

What are the problems associated with using relevance as a means of evaluating the performance of an information system? Relevance seems to be the best way to determine the effectiveness of a search--that the search is effective if the documents retrieved were relevant. Relevance is not always a straightforward determination but is subjectively determined based on the questions of the searcher. In the Cranfield studies no one could agree on exactly which documents were relevant because it could only be determined by the person with the question.

3. What are the one or two major differences between a classification system and a controlled vocabulary as information storage and retrieval systems?

Classification systems are logic structures that are used to organize objects into categories based on logical similarities such that all objects of the same kind or type are located near each other, for example you organize your closet by type of clothing. Controlled vocabularies are alphabetical lists of attributes and characteristics that are assigned to object surrogates in order to access and retrieve the objects from a database. Unlike classification systems, all vocabularies that apply to the object's attributes are assigned to the object surrogate, thus enabling more than one possible access point or pathway to retrieving the object.

4. Explain the pearl growing technique in analytic searches.

Pearl growing technique means that the searcher starts with a logical term that she thinks will retrieve appropriate information objects and tries the search. The seacher then checks the results to determine the accuracy of her terms and

looks in the results for subject terms and descriptors that better describe her information need trying another search with these new terms and again checking the subject headings and descriptors in the results to see if even more precise subject headings are available. The analogy to pearl growing is that the search goes in layers using better terms each time, terms that are already there in the database and not in the user's head, and terms that gets the searcher better recall and precision.

5. What does it mean to say that information seeking is a process of construction?

Information seeking is basically a trial-and-error process. Unless you have access to the controlled vocabulary, such as the lists of Descriptors used in the ERIC database and Dialog databases, or a Library of Congress Subject Headings, you simply start with words or phrases that best represent your information need and try them. Based on the results you modify your terms using synonyms and adding words using Boolean operators to broaden or narrow your search until you are satisfied with the number of information items recalled and the relevance of the items, always balancing the two factors of precision (narrowing a search) with recall (broadenins a search).

Part B.

Slightly less short answer. Answer 3 of the following 5 questions. Answers should be no more than 5-7 sentences (take this limit seriously too). Each question is worth 30 points. Don't answer more than 3 questions; any additional answers will not be graded.

6. Examine a directory from a search engine such as google or Yahoo! and look for types of hierarchical relationships. List 3 genus/species relationships, 3 associative relationships, and 3 instance relationships. (Remember you have to show both of the levels in the relationship - for instance, Cats and Siamese cats as an example of a genus/species relationship.)

Genus/species relationships: (works better without the plural s)  
social sciences/psychology  
games/puzzles  
sports/basketball

Associative relationships:  
computers/internet  
shopping/by region  
kids and teens/pre-school

Instance relationships:  
recreation/camping  
dog clubs/International Kennel clubs  
primitive technology equipment/blowguns

7. What are the four most important ideas contained in the Marchionini text, and why are they the most important?  
I did not answer this question.

8. Take one of the 15 articles you read from the Supplemental Readings and summarize it in NOT MORE THAN 3 sentences. Then explain how it fits in with the

texts and lectures in this class - how does it extend or enrich the material? (Note: the most frequent problem people have in answering questions like this is that they continue to describe the contents of the article rather than answering the final question.)

In "Subject Access in Online Catalogs: A Design Model" Marcia Bates makes the point that the current model of subject access in online catalogues is impoverished and does not make use of the online catalog's technological capability for enabling searches that are more attuned to users mental processes. She proposes a new design for online access of existing Library of Congress subject headings while expanding the capabilities of the search to include significant components that make use of natural language syntax and intelligent feedback information from the system that assists the searcher. In class we learned that Library of Congress subject headings are assigned to library holdings to enable searching or browsing a controlled vocabulary by subjects and retrieving information that is relevant to the subjects searched. We learned that the subject headings are assigned based on the attributes and characteristics of the document, the "aboutness" of the document, and that at least 20% of the document should be about that subjects. We also learned how to create a controlled vocabulary based on the concepts we found in documents. Bates' article extends the classwork in that she presents an analysis of user psychological needs that are not met by the current design of online catalogs subject searching. So this article enhanced my understanding of current online catalog subject searching capabilities and made me consider their limitations.

9. What is requirements elicitation? Why is it so important, and why is it so hard?

Requirements elicitation is studying how users will use a product, usually a software product, and should take place before the product design phase begins. The idea is to gather information about how people in the real world do their jobs and built these user needs into the product. Requirements analysis, or user studies, are important particularly with job automation, where a machine or software will perform a job now done manually. The point of analyzing how humans function at their jobs is to capture in the machine design all the possible human acts that people do without even being necessarily aware of them. This is why often ethnographic anthropologists do the user studies because they are trained to observe humans in contact with the environment and to ask pertinent questions. Another purpose of requirements analysis is to explore ways to improve upon present procedures in the way a process is carried out. This requires a correct analysis of the present procedures and brainstorming new possibilities.

TOTAL POINTS:  
(possible: 150)