

Kristin Yiotis
LIBR 202.01
Information Retrieval
Midterm Exam
Spring 2004
Judy 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 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 5 of the following 7 questions. Answers should be 2-4 sentences in length (take this limit seriously!). Each question is worth 15 points. Don't answer more than 5 questions; any additional answers will not be graded.

1. What is a surrogate? How do they function in information retrieval systems? Why does David Blair say (Meadow text, p. 3) that how to represent documents is the "central problem" of information retrieval?

Surrogates are representations of items or objects that are searched and retrieved by a database. The function of surrogates is to represent the items in the collection, so that the surrogates and not the items themselves are searched by the database. The central problem of creating databases is deciding what qualities or attributes of the items will be represented as surrogates. You want to choose attributes that allows the users to aggregate like items and discriminate unlike items.

2. Explain the concept of an inverted file and its use in information retrieval.

Databases store records in two ways: one is as files of ID and field values associated with that ID; the other is as files of fields called inverted files. The database builds an inverted file of each field that connects the field's values to the unique record ID. These are the files that the database searches when queries are submitted. The files live permanently in the database and are updated as new records are entered.

3. What is the importance of structures such as the MARC record and the Dublin core?

MARC records and the Dublin core are metadata schemes that enable data about data to be organized and represented according to established standards. MARC, or machine readable cataloging records, provides a schema for encoding data about books and other media that follows AACR2 descriptive and subjective cataloging rules. The Dublin Core is an effort to standardize metadata, such that all metadata schemes would include the same elements and could therefore be interchangeable and readable by each others' database software.

4. What does it mean to say that key functions of an information retrieval system are to discriminate and aggregate? How does a controlled vocabulary affect the user's ability to discriminate and aggregate?

A retrieval system works in two ways; it must enable the user to aggregate or gather together all records with similar attributes, and allow the user to select out or discriminate the records with unique attributes. A controlled vocabulary limits the number of possible field values. The controlled vocabulary fields that have validation lists enable aggregation of records. Free text fields allow for greater discrimination but must be accompanied by rules and explanatory notes to reduce error.

5. In what way are precoordinate and postcoordinate vocabularies related to specific technologies for information retrieval?

I did not answer this question.

6. Describe an example of an interface design causing cognitive friction (your own example; don't use one of Maria's) and explain what it is about the design that causes the friction. (Don't let this turn into a long answer; stay within the sentence limitation.) Be sure to provide the url in your answer.

I did not answer this question.

7. What does the term "navigation" mean as applied to interfaces?

Navigation is the term used to describe how users move around Web sites and still knows where they are. Web design standards have emerged that lead users to expect a common look and feel to each page in a site, such as a horizontal banner with page name and a return home logo or site ID at the upper left. Users expect a navigation bar at the left with button links to other pages or across the top with tab links to other pages. Good sites leave trails that show where the user is and where they have been, showing sections and subsections. Example of a good site: <http://www.fermanaghherald.com/index.html>

Part B.

Slightly less short answer. Answer 3 of the following 4 questions. Answers should be no more than 5-7 sentences (take this limit seriously too). Each question is worth 25 points. Don't answer more than 3 questions; any additional answers will not be graded.

8. Some retrieval systems provide relevance ranking for documents retrieved; others do not. Give a brief explanation of the two kinds of system (note: NOT of each system, just of the two kinds) that shows how they differ. (You don't need algorithms in this answer; I'm looking for general concepts that explain the difference.)

I did not answer this question.

9. In terms of how well they allow a user to discriminate and aggregate:
9a. How is a multiple access point system an improvement over a single point access system?

In the single access point system more than one document may be assigned a single term allowing for aggregating documents that are assigned the same term. In the multiple access point system more than one doc. may be assigned multiple terms but still you can only search one term at a time. Documents may share terms allowing for more than one way to get to a doc, but still aggregating docs under a single term.

- 9b. How is a conjunctive system an improvement over a multiple access point system?

Multiple point system and conjunctive system both assign multiple terms to multiple docs. But the conjunctive system can combine search terms in a Boolean AND format, allowing for a more granular aggregation of like docs.

- 9c. How is a coordinate system an improvement over a conjunctive system?

Like the conjunctive system, the coordinate system assigns multiple docs multiple terms, but the inclusion of the Boolean OR allows result to be rank ordered according to number of terms in the search request that were also in the document surrogates. This provides more granular aggregation and permits discrimination.

10. Select a website not discussed by Maria Brahme and analyze its interface in terms of her "Billboard 101" criteria. Do this within the 5-7 sentence range; you don't need to analyze every detail of the interface, just pick out highlights that meet or fail to meet her criteria. Be sure to provide the url in your answer.

I'm critiquing <http://lore.fhda.edu/lcen/50/index.html>, a site designed by my instructor at Foothill College. Maria's Billboard 101 criteria require a clear visual hierarchy. Pauline has used this aspect very subtly, using slightly larger, colored font for section headings, enhanced by small visuals like clip art consistently throughout the site. The standard conventions apply: similar look and feel to each page, identifying banner and logo ID, left navigation bar, color scheme consistent throughout. I would like the De Anza logo to be a link to De Anza College Home page. Pages are broken into blocks of text using background color. Each page has a specific purpose and topic. All underlined text is clickable links except for Library Vocabulary page where terms are underlined but not clickable, which is counterintuitive. Current page is clearly indicated on the nav bar. You can't tell where you have been, visited links don't change color, which I think keeps the site fresh. Minimum of noise, contributed to the fact that she built the site by hand and not with any Web design software.

11. Write an explanation of the diagram in Handout #1 with which we began this class, the Typology of an Information System (showing the document, document surrogate, need surrogate, etc.). Be sure you cover each element in the graphic.

A user has an information need and goes to a database to fetch documents, which are anything that are information bearing entities. The documents determine the indexing language, or controlled vocabulary, that the indexer used to create document surrogates and the searcher uses to retrieve records. The user creates the need surrogate, which is the search string that searches the inverted files. The database looks for an exact match between the user's search string and the document surrogates, assigning retrieved to documents that match and giving them retrieved status. In some databases items retrieved are rank ordered by relevance. The user's information need is met or the process starts over.