. **Conclusion** ***April 25***

Lectures 14.1-14.2

Final review. Answers

Numbers at left margin indicate number of minutes = number of points.

The answers are given on the page following the question so you can make your own notes before looking at the answer. The answers are generally far more complete than would be expected on the exam.

For easy reference, the question is repeated in a box on top of the first answer page.

****

Some key ideas you need to be able to apply whether or not there is a sample question

Entity-relationship data modeling, developing a conceptual data schema
Facets
Frames and document templates
Chained searching
Constructing a hierarchy from two facets
Applying KOS to search: synonym expansion and hierarchic expansion

****

15 **1. Develop a typology of Web documents**.

There are a wide variety of "documents" on the World Wide Web ("Web pages" and "Web sites"). In a catalog of Web documents, it might be useful to include an indication of the type of document in the catalog record. **Develop a typology of Web documents for this purpose**. (A typology is a list or classification of types).

Elaboration on the question. It would be useful if in a Web search engine one could filter the results by type of website or Web page. In Google this is possible in some cases (for example by searching Shopping rather than all of Google).

To start thinking about this task, ask yourself: How would I want to filter documents?

Question 1. Develop a typology of Web documents

There are a wide variety of "documents" on the World Wide Web ("Web pages" and "Web sites"). In a catalog of Web documents, it might be useful to include an indication of the type of document in the catalog record. **Develop a typology of Web documents for this purpose**. (A typology is a list or classification of types).

Elaboration on the question. It would be useful if in a Web search engine one could filter the results by type of website or Web page. In Google this is possible in some cases (for example by searching Shopping rather than all of Google).

Key ideas:

- **Facets**,
- Facets expressed as relationship types

Answer

The key idea is: Web documents can be classified from many different points of view (facets), so take a faceted approach. This is the principle used by Amazon and other shopping sites.

The box below lists some examples. There are more facets and certainly many more values under each facet; after the box, a few more examples are listed but not elaborated.

By overall purpose **Document** <*hasPurpose*> **Purpose**

. get information
. . home page
. . search engine
. . subject directory / reference
. . question-and- answer site or forum
. . just plain document (e.g., a journal article)
. . blog
. . give data (substantive data) or give pointer data (directory)
. social networking
. shopping
. entertainment
. games

By type of legal entity responsible

Chain: **Document** <*underResponsibiltyOf*> LegalEntity
LegalEntity <*isa*> **LegalEntityType**

. Personal
. Organization
. . Business
. . Non-profit
. Educational institution
. Government agency

By type of intellectual creator

Chain: **Document** <*createdBy*> LegalEntity
LegalEntity <*isa*> **LegalEntityType**

Sample values: Teacher, Student, Scientist, Journalist, Businessperson

By Web domain (correlated with previous, but not exactly the same)

Values: .com, .org, .net., .edu, .gov, .mil; .it[aly], .fr[ance], etc.

By medium **Document** <*usesMedium*> **Medium**

. Text
. Image
. Sound
. Multimedia

By format / file type **Document** <*isInFormat*> **Format**

Sample values: html, pdf, docx, pptx, xlsx, gif, jpeg, wave, mp3, **...** (hundreds of formats)

By language

By subject

Here are few more examples, not elaborated

(1) Factual versus fictional content.

(2) Physical Size (indication of scope): the amount of content can be quantified counting words, computer bites, number of images, etc..

(3) Intended Target audience: including age, education level, and geographic location, nationality, language, are items in this area.

(4) Quality/Authority of Website within a given field: This information may not be very objective and using popularity alone will be misleading.

(5) Suitability for certain audiences, access restrictions. If the content is not suitable for minors, there should be a filter to keep them out of reach of minors.

Conclusion. This is an example of the general principle that few topics can be adequately represented by a one-dimensional classification. It usually requires several dimensions or facets. Applying this simple principle leads, in most cases, to much better classifications and understanding of a topic.

30 2. **Home page design**

A naturalist organization keeps an inventory of (rare) plants and animals in Western New York.

They provide access to several general databases on plants and animals.

They maintain a database of locations where (rare) plants are found and of sightings of (rare) animals. They collect these data from individual members (or anybody, for that matter, with some check of credentials). Each occurrence/ sighting report is recorded by place, date, and time. Some of this information is kept confidential so as to not make poaching easy.

They also have events (lectures, excursions) they want to announce and make registrations for

Design a home page for such an organization.

Lectures 5.2a - 6.1 provide useful principles

Question 2. Home page design

A naturalist organization keeps an inventory of (rare) plants and animals in Western New York.

They provide access to several general databases on plants and animals.

They maintain a database of locations where (rare) plants are found and of sightings of (rare) animals. They collect these data from individual members (or anybody, for that matter, with some check of credentials). Each occurrence/ sighting report is recorded by place, date, and time. Some of this information is kept confidential so as to not make poaching easy.

They also have events (lectures, excursions) they want to announce and make registrations for

Design a home page for such an organization.

Key ideas mostly from document design

- Form follows function - determine the purpose of a document to be designed and adapt document structure to the function
- Express internal structure of information through the external structure – meaningful arrangement
- Follow common layouts so people can use what they already know
- Adapt document structure to the user's background
- Use the full repertory of means of expression.
- Use good contrast to make text easy to read and graphics easy to see

Answer

Include a search box and provide for navigation

Start with a generally known and used layout (Web layout 1 below). Use that layout for all pages on the site so the user does not need to look around and find things in different places on every page
Note: Some people who suffer from the disease of wanting to be different for the sake of being different put the vertical detailed navigation pane on the right. But in most Western writing we move from right to left, so that is what people are used to, and most websites put the navigation pane on the left. To just reinforce this: Think about a book put up on the Web. Such a book often has the option of showing the table of contents in a separate pane. Would you ever put the table of contents on the right? One argument for placement on the right is this: it avoids people having to skip over the navigation pane when looking at the main pane. Also, when the reader

reads line after line, she needs a clear stopping point upon returning to the beginning of the next line; the left edge of the screen provides a clear stopping point. The solution to this problem is to make the left-hand navigation pane a dark color with white font, effectively making the light-colored main display area the screen people look at.

Then look at the functions,

(1) Input of plant and animal sightings,

(2) Access to databases, and

(3) Events.

That needs three places in the main pain. These areas can be demarcated by a very light pastel background color.

For data input we need a form or template. To avoid clicking, that can be on the home page, say on the left side. There should be enhanced functionality, such as

- Once the user gets to the place field, a map opens up so the user can just click on a spot in the map.
- Once the user gets to the field for entering the species, a click will open a database that assists with identification
- There might be a data field that asks for the type of environment in which a plant was found, such as *meadow*, *swamp*, *forest*; the web page would have a drop-down menu with a meaningful arrangement of these. Same for soil types.

For database access, we may have a drop-down box of databases where the user can select one or more databases to search in. For displaying the results, this area would expand over the events area.

The events area lists events, newest first (events past are searchable on a separate page but not shown on the home page). Clicking on an event will show details.

There also is a button that opens a registration screen.

There might be some other elements to be fit onto the home page

A mission statement the organization wants to "push" on the reader. Perhaps best: a mission slogan, clicking on that will open the mission statement.

A welcome statement.

Pictures of rare plants and animals to attract readers; perhaps rotate.

A button to make contributions.

Could have a button for reporting construction projects that endanger rare species.

Perhaps under events: Encourage advocacy against such construction projects

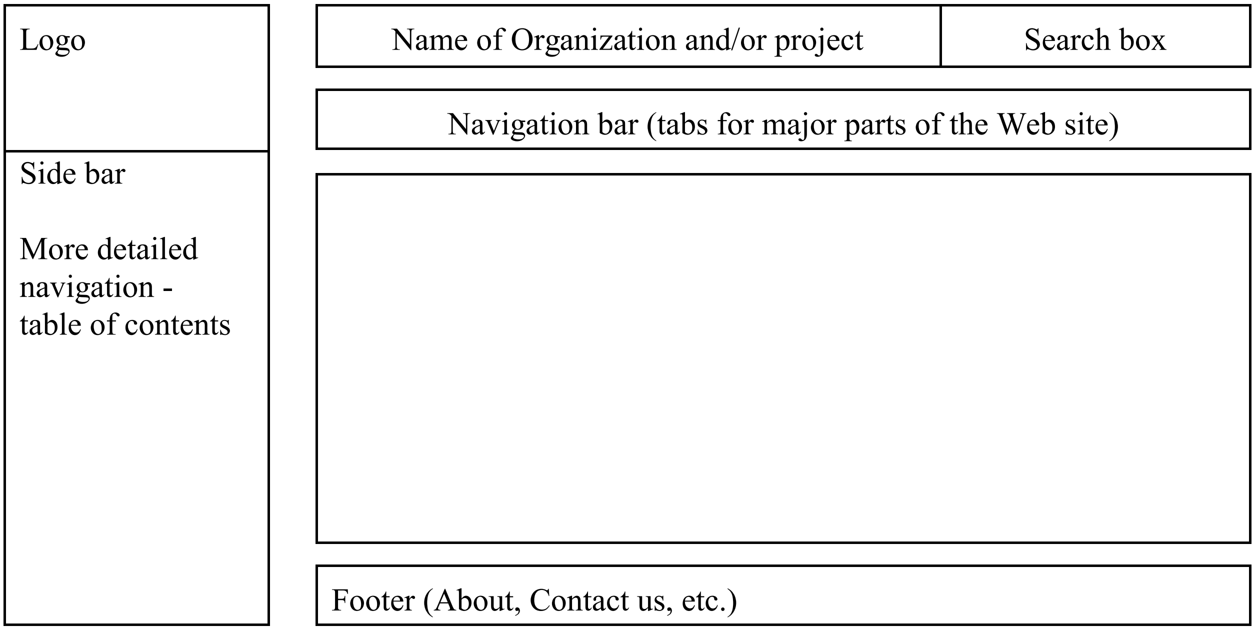
Think about colors for backgrounds and fonts. Contrast is all-important. Never use textured background for text

Other ideas

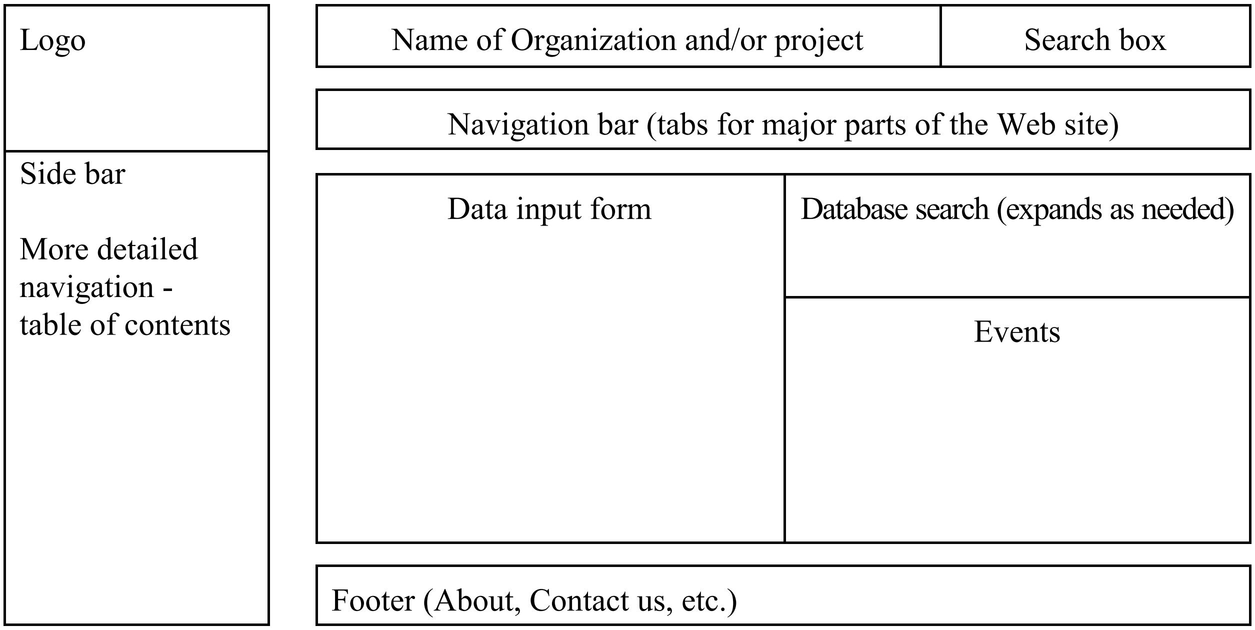
Tabbed sub-pages: About Us; Lectures; Excursions; Calendar and Registration; Rare Plant Sightings; Rare Animal Sightings; Contact Us, Log-in

- **About Us** Page: Brief history and purpose of the organization; Administrative/functional structure; past events and attendance
- **Lectures** Page (information only): List of up-coming lectures with description of Date; Time; Speaker with title and introduction; title; and a short abstract; total number allowed.
- **Excursions** Page (information only): List of up-coming excursions with destination; purpose; leader; Date; Time; Notes; and total number allowed.
- **Calendar and Registration** Page: with a monthly calendar format, clickable area for specific Lecture or Excursion leading to the registration subpage for name, address, phone number, email entrance and collection. Counter return "Full" signal when reached maximum capacity.
- **Rare Plant Sighting**s page: photos of rare plants, on clicking, sent viewer to log-in page.
- **Rare Animal Sighting**s page: same as above.
- **Contact U**s Page: list address, direction, phone number, email, and office hours.
- **Log-in** Page: viewers can become a member by registration with two member recommendation and gain access to secretive information in the database.

Web layout 1


Logo
Name of Organization and/or project
Search box
Navigation bar (tabs for major parts of the Web site)
Side bar
More detailed navigation - table of contents
Footer (About, Contact us, etc.)

Web layout 2


Logo
Name of Organization and/or project
Search box
Navigation bar (tabs for major parts of the Web site)
Side bar
More detailed navigation - table of contents
Data input form
Database search (expands as needed)
Events
Footer (About, Contact us, etc.)

A useful list of Web design principles that could be applied to this task.

- Use a **good title on the home page** so users understand at a glance what the site is about.
- Use **consistent page format** (layout, color scheme, fonts) across the site – use a template. There may be several specializations of the format for different kinds of information.
- **Chunk information**; show chunks in graphic blocks, using frames or colors. Divide chunks into still smaller points, use bulleted list.
- **Group information.** Arrange things often needed together or that logically belong together on the same page or in the same group of pages. Devise a logical top-down hierarchical structure for the entire site and provide a table of contents to the site.
- Use **consistent, clear, understandable titles or labels** (and language in general) for blocks of text.
- **Arrange by importance.** Arrange links and options from most important / most frequently used to least important. Draw attention to important items through color or font.
- **Easy access to frequently used functions and information** – on the top page, through a direct link from the home page (one click) or with few clicks. Examples: If 20% of all visitors to a university's home page want to search the faculty and staff directory, put the directory search box on the home page. Yahoo puts often-searched second level categories on its home page.
- **Clear labels for links** so that users always know why the link is made and what to expect once they get there.
- **Omnipresent navigation bar** for one-click access to the major subdivisions of the site and to the home page on every page. On specific pages have button *Back to* that leads to a major page in the user's most likely path, for example, a search page.
- **Use color and fonts consistently to express function**.
- Use an **appropriate mix of media** (text, images, sound). Make sure images serve a function (often in conjunction with text)
- **Avoid distractions**.
- **Ensure legibility**: contrast, no distracting background, font size, font.
- **Keep in mind color-blind and visually handicapped users**. May design separate pages and have clearly visible buttons leading to them (just as sites in different languages).

15 3. **Reorganize thesaurus information to take less reading and less storage space.**

The ERIC Thesaurus has the following entries (RT = Related Term)

Autoinstructional aids
RT Audiovisual aids
RT Computer assisted instruction
RT Courseware
RT Individualized instruction
RT Learner controlled instruction

P**rogrammed instructional materials**
RT Audiovisual aids
RT Computer assisted instruction
RT Courseware
RT Learner controlled instruction
RT Workbooks

Teaching machines
RT Computer assisted instruction
RT Courseware
RT Learner controlled instruction
RT Pacing

Hint: Compare the cross-references under each of the three terms.

Other example:

free participation
RT health care delivery and administration
RT health care economics

payment-based participation
RT health care delivery and administration
RT health care economics

subsidized payment
RT health care delivery and administration
RT health care economics

f**ull cost-recovery payment**
RT health care delivery and administration
RT health care economics

Question 3. Reorganizing thesaurus information to take less reading and less storage space.

Key ideas:

- **Hierarchical inheritance**,
- Grouping items that share the same information so that shared information needs to be listed only once under a new node. (Remember the restaurant menu)

Answer

The ERIC Thesaurus has the following entries:

Autoinstructional aids
RT Audiovisual aids
RT Computer assisted instruction
RT Courseware
RT Individualized instruction
RT Learner controlled instruction

Programmed instructional materials
RT Audiovisual aids
RT Computer assisted instruction
RT Courseware
RT Learner controlled instruction
RT Workbooks

Teaching machines
RT Computer assisted instruction
RT Courseware
RT Learner controlled instruction
RT Pacing

Reorganized entries (Relationships inherit down from the newly created top node)

The descriptors listed have three **RT** (Related Term) cross-references in common. The cross-reference information can be recorded more efficiently by creating a new node, a new descriptor and list the common information (the common cross-references) under it. The common cross-references are inherited down and need not be repeated under each of the original descriptors. So under Autoinstructional aids there is no need to repeat, for example,
RT Courseware since it is already at the higher level.

This takes less space (system perspective) and is easier on the reader (user perspective). Also the new descriptor is useful for searching. A computer program could go through the ERIC thesaurus and find groups of descriptors that have many cross-references in common; for such a group, a new descriptor one level above in the hierarchy would be created, improving the ERIC hierarchy. Finding good names for the new descriptors is a challenge.

Instruction/learning tool [newly created]
  RT Computer assisted instruction
  RT Courseware
  RT Learner controlled instruction

Autoinstructional aids
  RT Audiovisual aids
  RT Individualized instruction

Programmed instructional materials
  RT Audiovisual aids
  RT Workbooks

Teaching machines
  RT Pacing

Other example:

free participation
  RT health care delivery and administration
  RT health care economics

payment-based participation
  RT health care delivery and administration
  RT health care economics

subsidized payment
  RT health care delivery and administration
  RT health care economics

full cost-recovery payment
  RT health care delivery and administration
  RT health care economics

Answer

type of payment [newly created]
  RT health care delivery and administration
  RT health care economics

- **free participation**
- **payment-based participation**
  - **subsidized payment**
  - **full cost-recovery payment**

Keeping this idea in mind you can create documents that avoid redundancy and are easier to read and make it easier for the reader to remember the information the document conveys.

20 4. **Design an controlled vocabulary IR system**

> You have to **design an IR system that uses a controlled-vocabulary as its index language and human indexing**. An important requirement is that the system give the searcher the option of emphasizing either discrimination (one factor determining precision) or recall. List the features that are important for achieving this flexibility.

Questions to think about:

Assume users want to search for very specific topics without finding material even on closely related topics. Example:

> The user wants to find documents on *Atrial Fibrillation* but the only descriptor available is the broad descriptor *Arrhythmias, Cardiac*. This descriptor includes many other specific diseases such as *Pre-Excitation Syndromes* or *Ventricular Fibrillation ,* that the user is not interested in.

How should the index language be changed to satisfy this user? In general, what is the requirement on the index language when users want to search for very specific topics?

Another user may want to find all documents on the broad topic *Arrhythmias, Cardiac* without necessarily knowing all the different kinds. How can the index language be structured to help this user?

Some users looking for a topic may want to find just the key documents while others my want to leave no stone unturned. How can we design a system that will help both?

Question 4. Design an controlled vocabulary IR system

You have to **design an IR system that uses a controlled-vocabulary as its index language and human indexing**. An important requirement is that the system give the searcher the option of emphasizing either discrimination (one factor determining precision) or recall. List the features that are important for achieving this flexibility.

Key ideas:

- Exhaustivity and specificity of indexing and their effect on retrieval;
- indexing with weights;
- inclusive searching.

Answer

The basic principle is that the system should allow the user to express the search topic perfectly, whether the search topic is specific or broad and whether the user wishes to emphasize discrimination or recall.

Brief

Use a hierarchy with very specific descriptors to enable very specific indexing (allowing high precision for specific queries) and inclusive searching (allowing high recall for broad queries).

Use high exhaustivity indexing with weights (Organizing Information, p. 336, top). This will allow the searcher to emphasis recall (using descriptors with weight 1) or precision (using descriptors with weight 2). High exhaustivity can also be to advantage in a high precision search where it might allow the searcher to add a restrictive criterion to the search.

Provide for Boolean searching

More elaborated

1 **Very specific topic, user wants high discrimination**. The actual specificity of indexing must be high: The controlled vocabulary (CV) must include a descriptor for the specific user topic and the indexers must use that specific descriptor in indexing. For example, if a user is looking for something on type II diabetes, then type II diabetes must be a descriptor within the CV and indexers must assign it when applicable. If the most specific descriptor were just diabetes, the search would find many documents not relevant to the user, lowering discrimination.

2 **Broad topic, user wants high discrimination**, such as a search for all types of diabetes, a broad descriptor will deliver high discrimination because now all items retrieved by the broad descriptor diabetes are relevant; so for that case, low-specificity indexing would do. But remember that we want to build as system that can deliver high discrimination in all situations.

3 **Broad topic, user wants high recall**. Low specificity indexing would support this, but not 1. In a system that uses highly specific indexing (to support 1), a search for a broad topic must include all narrower terms. If the system does not offer help with that, the user might forget specific descriptors, resulting in lower recall. A system that uses highly specific indexing should provide inclusive searching to achieve high recall for a broad search.

4 **Broad or narrow topic, user wants to find even marginally relevant documents with high recall**. For some searches, especially searches on a single concept, this requires high exhaustivity in indexing, otherwise marginally relevant documents are not indexed with the concept and will not be found.

5 **Broad or narrow topic, user wants to find only documents that are right on target, with high discrimination**. For some searches, especially searches on a single concept, this requires low exhaustivity indexing, so marginally relevant documents are not indexed with the concept and correctly excluded

6 To **satisfy both 4 and 5,** the system needs to use high exhaustivity of indexing (to satisfy 4) with weights (to satisfy 5 by searching for the concept assigned with weight 2, thereby excluding marginally relevant documents).

Note that for multiple concepts ANDed high exhaustivity of indexing may allow formulating a query that achieves better discrimination.

Summary: **To give maximal flexibility and control to the user, a system should**

- use high specificity of indexing with inclusive searching;
- use high exhaustivity of indexing with weights.

5. **Query formulation for free text.**

A user needs information on the following topic:

Validity of the evaluation of instructors through undergraduate students in social science courses. (The user wants to search for publications that discuss whether the *evaluation of instructors by undergraduate students in social science courses* is valid, or not valid, or somewhere in between.)

A free-text search for this topic is to be made in two bibliographic databases

(1) in Database 1, searching is by terms occurring in the **title** of the document,

(2) in Database 2, searching is based on terms that occur in the **title and/or the abstract of the document.**

20 a. For each database give the **conceptual query formulation** that you would use (do not worry about terminology at this point). Give your rationale.

10 b. **Give the free-text query formulation for Database 2.** Assume that a search for this topic is to be made in Database 2 that searches on **titles and abstracts**. Any word or phrase (multi-word term) occurring in the title or abstract can be used as descriptor for searching. Briefly describe how you would go about developing the query formulations in terms of descriptors (words and phrases) (3 min.) Start doing it (7 min.)

To think about question a.

First note all the concepts in the query statement

Next, think about which of these concepts are likely to appear in the title? Is the title likely to mention all these concepts? How about the abstract? So would you include the same concepts into the query formulation for the title-based system and for the system based on title and abstract?

For Database 1, consider all the words in the title index terms.
For Database 2, consider all the words in the abstract index terms.
How does the indexing in databases 1 and 2 differ?

To think about question b.

What words may the title use other than *instructor*?

Question 5. Query formulation for free text.

A user needs information on the following topic:

Validity of the evaluation of instructors through undergraduate students in social science courses. (The user wants to search for publications that discuss whether the *evaluation of instructors by undergraduate students in social science courses* is valid, or not valid, or somewhere in between.)

A free-text search for this topic is to be made in two bibliographic databases

(1) **Database 1**. Searching is by terms occurring in the **title** of the document,

(2) **Database 2.** Searching is by terms occurring in the title and/or the **abstract** of the document**.**

a. For each database give the **conceptual query formulation** that you would use (do not worry about terminology at this point). Give your rationale.

Key idea

- Query formulation depends on the exhaustivity and specificity of indexing

Answer

- **All concepts**: *Validity, evaluation, instructors, undergraduate students, social sciences*
- Concepts represented in the **title**: *evaluation, instructors, student*
  Note: Titles often express content at a broader hierarchical level than the document itself
- Concepts represented in the **abstract**: *evaluation, instructors, student* and, in addition, *validity* (a concept in the study), *social science* (this is the environment of the study)

Thus, on the conceptual level, the query formulations should be as follows:

Database 1 Query: evaluation AND instructors AND students

Database 2 Query: evaluation AND instructors AND students
AND validity AND social science AND undergraduate students

Indexing by words in the title only is low exhaustivity, title + abstract is high exhaustivity;
the query formulation depends on the exhaustivity of the indexing.

Often the title indicates a concept that is broader than what is actually in the document; for example, the title may say student instead of undergraduate student. In such cases, indexing by words from the title is less specific than indexing by words from the abstract.

b. **Give the free-text query formulation for Database 2.** Assume that the search for the topic is to be made in Database 2 that searches on **titles and abstracts**. Any word or phrase (multi-word term) occurring in the title or abstract can be used as descriptor for searching. Briefly describe how you would go about developing the query formulation in terms of descriptors (words and phrases) (3 min) Start doing it (7 min).

Key idea

- Synonym expansion and hierarchic expansion of query terms for free-text searching

Answer

For each concept to be searched, find all its terms and narrower and broader concepts and their terms. Put differently, for each search term find synonyms, narrower terms, and broader terms

- **All terms for the concept *instructor*:** instructor, professor, teacher, educator, faculty (all ORed)
- **All terms for the concept *evaluation*:** evaluation, assessment, survey (maybe)
- **All terms for the concept *undergraduate student* plus all narrower concepts and their terms:** undergraduate, freshman, frosh (CALTECH for freshman), sophomore, junior, senior, underclassman, upperclassman [English terms]
- **Synonyms for validity**: none
- **Synonyms and narrower terms for social science**: behavioral science , sociology, anthropology, political science, etc.;

In Google, replacing *instructor* by (instructor OR professor OR teacher OR educator OR faculty) would lead to better results.

20 6. **Retrieval in Archives**

Background. This question deals with **retrieval in archives**; sufficient background is provided so that you can answer it even if you are not familiar with archives. Archives are a collection of documents (letters, memoranda, reports, etc.) produced by an organization, its various units, and the persons working in the units. (Assume an organization of the complexity of the Federal Government with many organizational units interrelated hierarchically and otherwise.) The organization of archives usually allows for easy retrieval of all documents produced by an organizational unit or a person; a document is linked to its producer at its creation so that the archivist need not do additional indexing to provide this type of access. Date when created, receiving organizational unit or person, and often related documents are also known for each document. **It is usually too expensive to assign subject descriptors to individual documents, yet subject searches are frequent.** The archivist doing a subject search uses her - more or less - complete knowledge of organizational units and persons and the subjects they have been dealing with at certain times to find relevant documents to look under appropriate units and persons.

Question. Sketch a conceptual data schema for a computerized retrieval system for archives that implements in a formal way what the archivist does informally. Describe how the system performs searches for a subject/topic.

To think about:

Formalize the steps that the archivist does in her head as described in the background statement.

What does the archivist know in her head?
What knowledge is embedded in the organization of the archives?

How does the archivist use the knowledge in her head and the knowledge embedded in the organization of the archives to find documents (records) on a subject even though documents are not indexed by document?

Question 6. Retrieval in Archives.

Sketch a conceptual data schema for a computerized retrieval system for archives that implements in a formal way the approach described. Describe how the system performs searches for a subject/topic. (Background not repeated)

Key ideas

- Entity-relationship modeling
- Chained searching. Putting information from multiple sources together to answer a question

Simple answer (good as an exam answer)

Devise a conceptual data schema that captures the type of data that are in the archivist's head. With such a schema we can build a database that stores the same knowledge.
Reminder: LegalEntity can be an Organization (organizational unit) or a person

Subject	<*coveredBy*>	LegalEntity	[In the archivist's head]
LegalEntity	<*produced*>	Document	[Through the organization of the archives]
Document	<*producedOn*>	Date	
LegalEntity	<*receivedBy*>	Document	
Document	<*isRelatedTo*>	Document	

Note 1: **Document <*dealsWith*> Subject cannot be used** since, as stated in the question, archives do not (usually) index by subject.
Note 2: If you listen to the recording from Spring 2011: I turned the direction of the relationships around so that the chained search is more natural. Instead of LegalEntity <*dealsWith*> Subject I used Subject <*coveredBy*> LegalEntity. No more need to follow relationships backward

Here is the key: The system needs to emulate the search process of the archivist. Give a search for documents on a given subject, the archivist first makes a (mental) list of the people and organizations (LegalEntities) he knows that dealt or deal with this subject. Then he looks in the records produced by (or received by) these LegalEntities:

Search: Subject --> LegalEntity --> Document

A chained search, two statements chained together (remember Lecture 1.2)

Can also add: Person <*belongsTo*> Organization (both of these are LegalEntities), make more complex chains

***We do searches that combine different types of data (chaining)

***Once you understand the process, you can design a system for it

***General note: Look up each piece of information you need and piece it together logically - it doesn't all need to be from the same source, but you do need to know to look for different sources if it's not all there.

More complex answer

(1) Sketch of conceptual schema

(Organizational unit, Time span)	<*dealsWith*>	Subject
(Person, Time-span)	<*dealsWith*>	Subject
Organizational unit	<*partOf*>	Organizational unit
(Person, Time-span)	<*belongsTo*>	Organizational unit
(Person, Time-span)	<*heads*>	Organizational unit
Document	<o*riginatedIn*>	Organizational unit
Document	<*authoredBy*>	Person
Document	<*receivedBy*>	(LegalEntity, Time)
Document	<*producedAt*>	Time
Document	<*refersTo*>	Document

Note: Time dependency of information is a complicating factor here. You could omit all the time elements and still have a good answer.

(2) Retrieval process for subject queries

Starting from a subject, find an (Organizational unit, Time-span) combination. Then look for documents originated in or received by the organizational unit during the time span. Same for person. One can also look for persons that belong to an organizational unit, especially the person that heads the unit (always within the time span) and then look for documents authored by these persons. Likewise, one can go up and down the hierarchy of organizational units. Having found a relevant document, one can look for documents it refers to and for documents that refer to it, or one can identify receiving organizational unit or person (or originating organizational unit or authorizing person for documents bound based on their recipient) or coauthors, and use these as starting points in new searches.

An organizational unit or person may lead to related subjects. Reading a document may lead to related subjects and to organizational units or persons that can serve as new starting points.

Other example

To find out whether a certain kind of chemicals poses a problem in the environment.

(1) Find out all chemicals in a certain class, consult a chemical structure data base.
(2) For each of the chemicals, find the toxicity / harmful effects in a chemical effects database.
(3) In a database of industrial, agricultural and other uses of chemicals, find information on how and how much each chemical is used.

Putting all this information together will answer the question.

40 7. **Index design**

You are appointed as head of a medium-sized IR-system (about 200,000 documents) that uses three different systems for subject access:

(1) an alphabetical subject catalog of books;
(2) shelving books by subject;
(3) an independent classification scheme for filing newspaper clippings

Your analysis shows that the subject heading list and the shelving classification are both far from satisfactory. The subject headings have grown without control and no listing is available. The users have difficulty finding the right descriptors in these index languages. **But a cost-benefit analysis rules out major changes or revision, like introducing new schemes, especially in view of the large costs for re-indexing the old collection**. On the other hand, the cost-benefit analysis also shows that some costs would be justified to improve the usability of the IR-system. What do you suggest should be done? How would you implement your suggestions?

To think about

What is the main problem of the users?

How could the users be assisted in solving that problem.

Question 7. Index design

You are appointed as head of a medium-sized IR-system (about 200,000 documents) that uses three different systems for subject access:

(1) an alphabetical subject catalog of books;
(2) shelving books by subject;
(3) an independent classification scheme for filing newspaper clippings

Your analysis shows that the subject heading list and the shelving classification are both far from satisfactory. The subject headings have grown without control and no listing is available. But a cost-benefit analysis rules out major changes or revision, like introducing new schemes, especially in view of the large costs for re-indexing the old collection. On the other hand, the cost-benefit analysis also shows that some costs would be justified to improve the usability of the IR-system. What do you suggest should be done? How would you implement your suggestions?

Key ideas

- Descriptor-find index (see Organizing Information, Section 15.5.1, p. 313 - 317)
- Semantic factoring - expressing compound concepts through a combination of elemental concepts
- User-friendly system front-end: A user-friendly system that helps the user with formulating a query and then runs the query against a hard-to-use system (a system that makes it hard to formulate good queries, which includes free-text systems)

Answer

A user wanting to search for a given topic has difficulty finding descriptors for one of the systems, let alone for all 3 (for a cross-system search). To help the user find descriptors, **build a descriptor-find index**.

Have a small classification of elemental concepts (the core classification, preferably a faceted classification); ideally, you could use an existing classification as a starting point. Express each of the precombined descriptors (from all three schemes) through a combination of elemental concepts. Then, through a faceted search the users can find precombined descriptors and then they can use these descriptors to find documents.
When the core classification does not contain all elemental concepts needed to express a precombined descriptor, add the needed elemental concept(s) as you go.

Constructing a descriptor-find index is cheaper than revamping all three index languages and re-index all three collections. The descriptor-find index is a front-end that makes the existing system more usable.

40 8. **How much money should be spent for indexing?**

You are given the task to design an IR system. One problem is to determine **how much money should be spent for indexing**. Discuss the data you need/the considerations on which you would base your decision.

This is a big question. If you are very pressed for time you may want to skip it.

Question 8. How much money should be spent for indexing?

You are given the task to design an IR system. One problem is to determine **how much money should be spent for indexing**. Discuss the data you need/the considerations on which you would base your decision.

Key ideas

- Cost-benefit analysis
- System performance
- Effect of indexing on system performance

Answer

Note: This answer is way more elaborate than would be expected on an exam.

The basic consideration is as follows:

1. Each alternative strategy for indexing has a certain cost.

2. Each alternative method of indexing has an effect on the performance of the system, specifically7 recall and precision of the answers.

3. Recall and precision result in benefits that can be derived from the ISAR system.

Therefore, we need to know the following:

1. What are the costs associated with each alternative strategy for indexing?

2. What are the effects of each alternative strategy on recall and precision for the type of search requests likely to be asked? (There user needs come in!)

3. What are the savings in search and processing cost and the benefits to be derived from better recall and precision (which also affect the speed with which the information ultimately needed can be obtained)?

Question 1 can be answered through cost studies.

Question 2 relates to the internal functioning of an ISAR system and may be answered through retrieval experiments (having in mind all the limitations of such experiments).

Question 3 relates to the purpose of the organization at hand, and how this purpose can be fostered through more information. It can be answered through user and use studies.

In this methodological framework, we can look at some concrete points. Indexing strategies can be characterized by the following parameters:

Type of indexing language used
Use of checklist technique vs. extraction and translation method
Exhaustivity
Specificity
Pre-combination vs. post-combination
Use of weights
Use of roles and links

Indexing costs arise:

(a) from the construction of the indexing language

(b) from the indexing process itself

(c) from subsequent processing and storage

Points 1 and 2 of the following discussion focus on exhaustivity as an example.

1 Costs

(a) Exhaustivity of indexing does not affect the construction of the indexing language.

(b) Higher exhaustivity means more time spent for indexing, and therefore higher costs for (b). (An exception occurs at very low levels of exhaustivity; it might often be easier to assign five descriptors than decide which three to choose to fit into a rule max three descriptors per document.)

(c) If the degree of pre-combination is the same, higher exhaustivity means more descriptors per document; therefore, more costs for processing and storage. In a card catalog, this factor severely limits exhaustivity.

(d) Specificity of indexing affects both the cost of constructing the index language (many more specific descriptors need to be included) and the cost of indexing

2 Performance

The effect of exhaustivity on recall and precision depends on the search situation as discussed in text chapter16. Furthermore, exhaustivity may only be required for certain subject areas.

The effect of specificity on precision depends on the nature of the search requests. For a highly specific search request the system can achieve high precision only with highly specific indexing.

3 Cost savings and benefits from recall and precision

This discussion is general; it does not matter by what indexing device recall and precision are achieved

Savings in search costs. When precision is higher, less time is needed for screening results. This is all the more important in screening of search results is expensive (for example, if the documents are audio or video passages or if it is expensive to obtain the documents for examination).

Benefits from search results. Recall is important if complete information is required. Precision is important if responses have to be timely. For example, there is a big bonus for timeliness in medical or other emergencies when information is needed as a basis for quick action). This timeliness requirement may justify expenditure to obtain high precision even if there are only a few questions (there is no other way to achieve timeliness).

Two obvious parameters were not discussed so far:

Number of documents. The more documents in the collection, the more expensive is it to index them all. On the other hand, the more documents in the collection, the more important is it to pinpoint the documents actually relevant in order to limit the costs for search result screening.

Number of search requests. The more search requests, the more the savings in screening search results add up. Also, the higher the total benefits (summed over all requests). As is clear from the discussion above we also need to know about the nature of the requests, for example broad vs. highly specific.

40 9. **Developing an index language and thesauru**s

You are given the task of **developing an index language and thesauru**s for
(1) a newly set up information center in a company, or

(2) a public information center in the inner city (choose **one)**.

What are the main points you have to take into consideration in performing this task?

To think about

How do you know what information would help people solve their problems?

Question 9. Developing an index language and thesaurus

You are given the task of **developing an index language and thesaurus** for

(1) a newly set up information center in a company, or

(2) a public information center in the inner city (choose **one)**.

What are the main points you have to take into consideration in performing this task?

Key ideas

- User orientation, studying true user needs
- Request-oriented indexing

Answer (with focus on (2) a public information center in the inner city, but method general)

Apply the approach of request-oriented indexing. Find out the problems of the users and the information needs that come from those problems. Analyze users problems and information needs arising from these problems (Chapter 7). Need to use sources beyond asking users, for example a study done of the demographics and the economic situation of the community. Answer the following questions:

1. Who are the users? What are their information needs?
2. What terminology do the patrons use in their searching?
3. Do they know what information they need? What information will be useful to them?
   a. Ask them, of course, but also analyze the problems of the community and see what information you can make accessible to help them.
   b. Use census data
   c. Learn about the community, what is available (e.g., if people are looking for jobs, need to give them information about careers, kinds of jobs available within a reachable area, job training material. The index language to be developed needs to support this)
4. Is the information system meant to be used by librarians or by the patrons themselves?
   a. This will determine the kind of language you can use – more layman terms for patrons, more specialized terms for librarians

Another aspect is what kind of retrieval system is currently available? If it is a computer retrieval system we can use an index language of elemental concepts that can be combined at search times to express the (usually compound) search topics. On the other hand, if we want to organize a career file for browsing (on paper or on the Web), we need and index language of precombined descriptors. We may need both.

Note on index language (taxonomy) for a company or other organization

You might think that all searches are done on Google anyway, so why bother. You would be wrong on two accounts:

(1) A large company or other organization has much internal information, often made available through a secure intranet. (You could further argue that the intranet could be searched using the Google search appliance internally, but see (2))

(2) Google is very useful for simple searches. When used expertly it can also do conceptually complex searches, but there are limits. For complex searches for concepts of importance for the company's or organization's mission, where high precision and reasonably high recall is required, searching with a controlled vocabulary organized into a hierarchy works much better. Indexing can be done manually or computer-assisted or completely automated.

This is why many organizations build hierarchically structured index languages (often called taxonomies). Usually these are expanded to include many synonyms to become full-fledged thesauri. They use their thesaurus to index internal documents and sometimes relevant external documents to support efficient search focused on what is important for the organization. The thesaurus is also useful to devise better free-text queries, including queries to Google.

Lastly, many organizations collect an array of statistical data. The very definition of these data is determined by the concepts in the taxonomy.

20 10. **Assist users in coping with large Web search results**.

A search in a Web directory, such the Open Directory Project (http://dmoz.org/about.html) or the Wikipedia directory, or a search engine, such as Google or Bing, often returns hundreds or thousands of documents. What could the system do to help the user to cope with these large numbers?

Question 10. Assist users in coping with large Web search results.

A search in a Web directory, such the Open Directory Project (http://dmoz.org/about.html) or the Wikipedia directory, or a search engine, such as Google or Bing, often returns hundreds or thousands of documents. What could the system do to help the user to cope with these large numbers?

Key ideas

- Meaningful arrangement

Answer

Note: Simply suggesting a meaningful arrangement would be a good answer.

Ranking results by system-predicted relevance is the predominant method. It works well when the user needs specific pieces of information that can be found in the three top-ranked Web pages. But if the user needs many documents, for example for writing a paper on a complex topic, ranking provides little help. Grouping the top 500 - 1,000 Web pages by topic would enable the user to eliminate some whole groups immediately without looking at each Web page individually. This is particularly useful if the key search term has multiple meanings, such as *canal*. In a search for *canals for transportation* it would be useful to eliminate right off the bat groups of Web pages dealing with *ear canal* or *root canal*. The remaining groups may represent useful subtopics of the paper topic and suggest a first draft of an outline for the paper. They contain similar documents that are best read together. Groups can be shown in a linear arrangement or in a 2-D concept map.

Groups can be formed automatically in two ways:

1 One can build a classifier that uses the words occurring in Web pages to assign Dewey classes or Yahoo classes. For example, a preponderance of medical words or phrases in a Web page would cause it to be grouped under *root canal* or *ear canal*. Groups will then be displayed according to the classification chosen.

2 The system can form groups from scratch by computing the similarity of every pair Web pages based on the words they contain and then group Web pages together that are more similar to each other than to Web pages outside the group. This method is called *clustering*.

See, for example, www.folden.info/searchengineclustertechnology.shtml

Another approach is to refine the relevance scores of web pages by applying more criteria (a process that takes more computer time than is practical in searching the Web at large). The answer to Question 1 provides some ideas.

15 11. Discuss **exhaustivity** in the context **of hyperlinks** made on a web page.

Things to think about.

How many hyperlinks?

Are all hyperlinks equally important?

Question 11. Discuss **exhaustivity** in the context **of hyperlinks** made on a web page

Key ideas

- Exhaustivity of indexing
- Weights in indexing
- Taking a concept from one context and applying it in another context

Answer

Low exhaustivity: Only the most important links are included, which means using few links. There is high importance threshold. The user of the linking page must derive a reasonable benefit from viewing the linked-to page

High exhaustivity: Links are included even if the association to the other document is only tenuous, Results in many links. For many users, following many of the links might not be a benefit at all.

One could weight the links, for example by expressing their importance through the link anchor text, such as using different color, put asterisks around them, or explicitly label them *important*. Eventually mark-up languages might permit encoding link weight, giving the user the option of showing only the most important links or all links.

Optional

15 12. **A large subject index is to be put on microfiche. How to arrange?**

The system has two parts:

(1) The actual index on microfiche. This is an ordinary index: Under each descriptor the entries for the documents (or other retrieval objects) indexed by that descriptor are listed.

(2) To help the user find the appropriate microfiche, there is a hard copy "index to the index." This is simply a list of all descriptors, giving for each the microfiche number and the frame number on the microfiche.

Question: Should the subject index on microfiche be arranged in classified or in alphabetical order? How should the hard-copy "index to the index" be arranged?

Assume a microfiche reader where the user must manually insert the fiche and find the frame.

Note: This question is clearly obsolete technologically but still useful to illustrate a principle that applies generally.

A microfiche is like microfilm on a 4x6 cards with, for example, 96 pages of text, that in this example contain catalog records.

Question 12. Subject index on microfiche. How to arrange?

A large subject index is to be put on microfiche. The system has two parts:

(1) The actual index on microfiche. This is an ordinary index: Under each descriptor the entries for the documents (or other retrieval objects) indexed by that descriptor are listed.

(2) To help the user find the appropriate microfiche, there is a hard copy "index to the index." This is simply a list of all descriptors, giving for each the microfiche number and the frame number on the microfiche.

Question: Should the subject index on microfiche be arranged in classified or in alphabetical order? How should the hard-copy "index to the index” be arranged?

Assume a microfiche reader where the user must manually insert the fiche and find the frame.

Key idea

- Meaningful arrangement

Answer

Context: It was argued that in a subject card catalog subjects should be arranged alphabetically by the preferred term used to designate the subject. The idea was that the user could think of a term, look for that term in the alphabetically arranged catalog, and in one step find documents. Of course this would work only if the term the user thought of happened to be the subject heading used in the library catalog. Users who were not successful could have profited from a hierarchically arranged browsable list of subject headings or a descriptor-find index.

The microfiche index should be arranged by subject. This allows the person to look through all related terms under a subject without switching microfiche, possibly finding more useful terms with no extra work. There is no downside to this since the user must find first the microfiche from an overall index, so the search is always two steps

The index to the index should be alphabetical (including synonyms) so that the user can easily look up the term they have in mind. In addition, a classified arrangement of the index to the index would be useful as it would be easier for people to browse.

The best way of arranging a catalog depends on how it can be accessed. Computerized indexes have different challenges/advantages than other formats. In any system, both access by word or term (through an alphabetical index in print and search capability online) and through a meaningful classified arrangement should be provided.

In online systems, alphabetically arranged lists of terms rarely serve a purpose; the user can just type in a term and the system will do the search. On the other hand terms arranged in a meaningful hierarchy allow a user to browse even if she does not have a term in mind.

40 13. **Design and development of an online information retrieval system for courses**

You are charged with the **design and development of an online information retrieval system for courses** at the University at Buffalo. The system should serve

(1) students in course selection and

(2) curriculum committees who want to know what courses exist in a given area (such as *statistics* or *communication in organizations* before approving a new course.

Discuss your approach (describe the workings of the system you propose to the extent feasible in 40 minutes; bulleted lists for some pieces are fine)

Think about your own problems in finding appropriate courses, especially courses outside LIS that would be very interesting for **your** interests.

Design a system that would make your life easier.

What pieces of knowledge does such a system need?

Question 13. Design and development of an online information retrieval system for courses

You are charged with the design and development of an online information retrieval system for courses at the University at Buffalo. The system should serve

(1) students in course selection and

(2) curriculum committees who want to know what courses exist in a given area (such as *statistics* or *communication in organizations* before approving a new course.

Discuss your approach (describe the workings of the system you propose to the extent feasible in 40 minutes; bulleted lists for some pieces are fine)

Key ideas

- User orientation. Requirements analysis based on the problem to be solved
- Entity-relationship modeling

Answer

First write down what functions the system must support (questions, other types of processing). In real life: user study. On exam: based on your own experience.
Even in real life, the systems analyst would list functions to the extent she is able to based on common sense and looking at other systems to ask users more informed questions.

Some functions for students (and advisers)

A student needs to know about courses suitable for his interests and his program.

Search courses by

- subject (a considerable problem),
- level of the course (undergraduate, master, PhD)
- intended audience / learning objectives / relationship to later jobs
- time offered in a given semester (for several semesters in advance),
- location offered (Buffalo campus, Rochester, online)
- instructor (for each offering of the course),
- student evaluation scores from previous offerings and/or evaluation of the instructor (for example as found on the Web) (use with care)
- availability of open sections at times feasible for the student

Show all these types of information plus

- course descriptions (belong to course) and
- syllabi (belong to course offering (course section) because the syllabus depends on the instructor and is updated regularly)
- how difficult is it to get into the course.
- precise location (room number) of section registered for

Help in creating plan of study:

- Make sure that courses are taken in a prescribed sequence (if any),
- Make sure student has prerequisites for a course by the time she plans to take it,
- Make sure the plan of study meets all requirements of the student's program (e.g., MLS) and of any certifications or civil service exams the student wants to obtain.
- Based on a list of student interests and time and financial aid constraints and based on a three-year teaching plan, put together some suggested plans of study (this would be an expert system for building a plan of study.

The system should be able to select courses of interest to the student based on the criteria mentioned above and then, given the schedule constraints of the student, map out a tentative sequence of courses which would be kept on file under the student's name. This tentative schedule would be updated as new information about courses becomes available or as the student's schedule requirements change. At registration time, the student would look over the plan, finalize selections for the coming semester, and be registered automatically (no separate input of data necessary). If advisor approval is required, this could be easily integrated by giving the advisors passwords; only a user with the proper password could finalize course selections.

Some functions for a curriculum committee

A curriculum committee must evaluate whether a proposed course is needed and the quality of the course

Search:

- Find courses similar to the proposed course

Display:

- For each of these courses show demand based on past enrollment
- Is the course part of a new program that is expected to draw many students
- Mode of offering (seated, online synchronous, online asynchronous)
- Syllabus (to evaluate material presented, readings, student activities etc.)
- Instructor CV (to evaluate instructor qualifications)

After the requirements analysis, construct a conceptual data schema.

The requirements written above are written in such a way that the entity types and relationship types should jump out at you. You need to distinguish between Course and Course offering (see Lecture 1.2 and Organizing Information Chapter 3).

Just some examples

Course	<*belongsTo*>	(Program, RequirementStatus)
CourseOffering	<*hasSyllabus*>	Text
Course	<*hasStatus*>	CourseStatus [proposed, inCatalog, inactive]

12 14. **Degree of precombination and exhaustivity and specificity of indexing**

Compare a system using shelf arrangement based on an index language like LCC or DDC with a system based on postcombination (such as a computerized IR system) with respect to the exhaustivity and specificity of indexing that can be achieved. What can you say about retrieval performance in both cases?

Note: did this in an in-lecture exercise in Lecture 13.2. What did you learn from that example?

Question 14. Degree of precombination and exhaustivity and specificity of indexing

Compare [1] a system based on postcombination (such as a computerized IR system) with [2] a system using shelf arrangement based on an index language like LCC or DDC with respect to the exhaustivity and specificity of indexing that can be achieved. What can you say about retrieval performance in both cases?

Key idea

- Postcombination and precombination

Answer

The in-lecture exercise in Lecture 13.2 comparing indexing with **[1] the faceted London Education Classification (LEC)** and **[2] the Dewey Decimal Classification (DDC)** did this . The upshot of this exercise is that a post-combination system enables both higher specificity and higher exhaustivity of indexing. This supports better retrieval performance as illustrated in Lecture 13.2 based on Organizing Information Chapter 16, specifically Section 16.3

This general trend applies also to the comparison of **[1] precombination systems with a high degree of enumeration**, such as LCC, and **[2] precombination systems that rely more on building new precombined descriptors by the indexer**, such as DDC. For example, in [1] LCC, many combinations of a subject and a place (such as railways in Egypt) are enumerated, with specificity in the place facet often limited to the level of entire countries. On the other hand, [2] DDC does not enumerate any subject-place combinations. The indexer builds a class, using the subject from the schedules and place from Table 2, which gives places to quite a high level of specificity down to counties.

Entirely optional. Read only if you are interested in linguistic applications
No such question will be on the final

30 15. **Design a large lexical and classification database**

Assume you have to **design a large lexical and classification database** that has the ambitious objective of serving as a tool for both natural language processing and indexing and retrieval. What information should be included for each term or concept?

There will be nothing like this on the final exam.

If you are interested in an in-depth treatment, go to this paper
http://www.dsoergel.com/UBLIS571DS-14.1-14.2-1ReadingSoergelSemWebFull.pdf
and see p. 10 – 11 (far more comprehensive than we could have done as a class exercise in 15 minutes, never mind a response to an exam question).

Final review. Natural language processing (NLP). See Lecture Notes

Final review. Precombination vs postcombination See Lecture Notes