3. Information Organization
{and,or,vs} Search

INFO 202 - 5 September 2007

Bob Glushko
Plan for INFO 202 Lecture #3

The Information Life Cycle

"Search" != "Search Engine"

Different Models of the Search Process

Relevance, Recall, and Precision

The IO and IR tradeoff
The Information Life Cycle: Creation / Organization

- Recognize a need for it
- Plan to create or collect it
- Design a structure for it... or not
- Plan to manage it
- Author or collect it
The Information Life Cycle: Retrieval

Find it
Filter it
Transform it
Combine or assemble it
Integrate it
The Information Life Cycle: Usage

- Publish it
- Syndicate it
- Generate something from it
- Reuse, revise, repurpose, or retarget it
- Act on it
Where People Search for Information

Physical libraries

The web, using a search engine like google/yahoo/msn/etc.

Personal information collections

The people in professional or social networks

An employer's intranet or business systems

Information resources not searchable on the web, like those from Lexis/Nexis, Westlaw, etc. etc.
The Business of Search

Search techniques and technologies have enormous influence on how people find and think about information.

Search is obviously very important to Google, Yahoo!, and other "search" companies, as well as to Microsoft, SAP, Oracle and other application vendors.

Every business, on and off the web, except very small ones, need effective search capabilities to stay in business.
"Search" != "Search Engine"

- Google
- Answers.com
- Healthline.com
- Eurekster.com
- Alacrastore
What Do People Search for On Search Engines?

Queries as Zeitgeist

- Google Trends
- Lycos Top 50

Average query length is between two and three words. Why don't people exploit the power of search engines?

Self-reports don't match query logs. Why not?
Models of the Search Process

Classical model

Alternative models

- Iterative models - IR as problem-solving dialog
- "Berry picking" and foraging
The Classical Model of Search

Assumes a "go to the library and use the IR system" approach, often with an intermediary of a trained searcher actually doing the search.

Assumes the information need remains the same throughout the search process.

Goal is to maximize recall and precision simultaneously.

The value is in the retrieved document set.
Schematic View of Classical Search
Classical Search: Processing [1]

The user translates an information need or question(s) into a QUERY.

The query expresses the information need in a format or as a set of descriptive features that the system can handle.

The processable representation of these features make up the INDEX or INDICES.
Contrast the Classical Model with "Finding Out About"
Classical Search: Processing [2]

The system matches the descriptive features in the query against the features that describe the "documents" (or pointers to documents) stored by the system.

Documents are retrieved when the degree of the match exceeds some measure of similarity (which might be "exact match" for some queries or systems).

The system presents the retrieved documents according to the measure of similarity.

The user assesses the RELEVANCE of the documents (information that sometimes can be provided to the system).
Classical Search: Processing [3]
Relevance

A document in some collection is relevant if it:

- Answers a precise question precisely: *What is the capital of California?* (Sacramento)
- Partially answers a question: *Where is Emeryville?* (Near Berkeley)
- Suggests a source for more information: *What is lymphodema?* (Look in this Medical Dictionary)
- Provides background information
- Reminds about other knowledge
Recall and Precision

ALL DOCUMENTS

Relevant

Retrieved
Recall and Precision [2]

RECALL is the proportion of the relevant documents that are retrieved

PRECISION is the proportion of the retrieved documents that are relevant

Goal: High recall and precision - Get as much good stuff as possible while getting as little junk as possible
High Recall but Low Precision
Low Recall but High Precision
High Recall and High Precision
Problems with the Standard Model

- People don't just search "formal" information sources
- Users learn during the search process
- They refine their original queries and generate new ones
Iterative Models - IR as Dialog

IR is an active process of Finding Out About

It involves making hypotheses, evaluating whether process is on a useful path, proceeding forward (and backward)

Users can recognize elements of a useful answer, even when incomplete

Questions and understanding changes as the process continues
Berry Picking Model [1]

A sketch of a searcher... "moving through many actions towards a general goal of satisfactory completion of research related to an information need." (after Bates 89)
Berry Picking Model [2]

Interesting information is scattered like berries among bushes

New information may yield new ideas and new directions

The information need is not satisfied by a single, final retrieved set - it is satisfied by a series of selections and bits of information found along the way

Strategies include footnote chasing, citation searching, journal run, area scanning, subject searches, and author searches
Information Foraging -- "Adaptive Interaction with Information"
Information Foraging - Overview

A more recent and rigorous theory about the search process called INFORMATION FORAGING was proposed by Peter Pirolli and Stuart Card.

Uses concepts of foraging theory from evolutionary biology

- How do predators decide which prey to hunt? == How does an "informavore" choose which information resources to pursue?

- Animals adapt their search strategies == Informavores also adapt search tactics, and modify their own information environments and information organization to make their searches more productive.
Information Foraging - Discussion and Implications

Are you more likely to launch a web search or walk to the library to find some information?

Where are the "meaty" information resources more likely to be found?

What takes more effort?

Foraging theory holds that animals use "scent" to assess the likely value of a food source and to determine the direction to hunt. What are the "scents" for information resources?
Two Little Red Books Full of Good Quotes
The effectiveness of a system for organizing information is a direct function of the intelligence put into organizing it (Preface, ix).

While some access problems are caused by new technology, others -- those that stem from the variety of information, the many faces of its users, and the anomalies that characterize the language of retrieval -- have been around a long time (p. 2).

Whether users search library shelves or the Internet, some will retrieve too little, some too much, and some will be unable to formulate adequate search requests (p. 2).

It has never been easy to explain why colossal labor should be needed to organize information (p. 10).
Organization {and,or,vs} Search [1]

We organize to enable retrieval

The more effort we put into organizing information, the more effectively it can be retrieved

The more effort we put into retrieving information, the less it needs to be organized first

We need to think in terms of investment, allocation of costs and benefits between the organizer and retriever

The allocation differs according to the relationship between them; who does the work and who gets the benefit?
An AUTHOR anticipates the interests of an AUDIENCE and creates information that is a balance between what the author wants to say and what he or she thinks the audience wants to know.

How precisely the author knows the anticipated information needs shapes the choices made about the extent and nature of the information organization.

The author designs or selects a structure for the information... or not.

These structuring decisions made by authors in creating and organizing information impact its retrieval and use.
Organization {and,or,vs} Search [3]

Collections of unstructured information contain documents from many authors targeting many different audiences.

So people looking for information will likely have different purposes and use language differently from the authors.

So then the information retrieved has to be evaluated for how well it answers the user's questions.

With structured information, the answer retrieved is completely correct with respect to the underlying semantics of the information; if you find it, it is what you wanted.
People are familiar with many characteristics of their personal information because it is "stuff they've seen"

Any organization they impose is likely to be highly idiosyncratic or biased... or seem that way to other people

The context in which they created or encountered it is especially useful in retrieving it

Search followed by iterative contextual filtering is an effective retrieval strategy because it replaces a recall task with the easier recognition one
Enterprise information is produced by people doing their jobs and its structure and quality is often governed by policy or technology support.

Enterprise information can exist in numerous formats and applications, many of which are hard to get at via enterprise search because they are "silos".

But because these formats are known, it can be worth the effort to invest in information extraction and text mining.

Information types span the spectrum from unstructured to semi-structured to structured.

In enterprise search, a person usually has a very clear idea of what they are looking for, and there is often a single source for the "right" answer.
Enterprise Search
Information Extraction
Information Extraction Not Necessary

<Restaurant>
  <Name>El Floridita Restaurant</Name>
  <Address>
    <Number>1253</Number>
    <Street>N. Vine Street</Street>
    <City>Los Angeles</City>
  </Address>
  <Phone>(213) 871-8612</Phone>
  <Review>Named after Hemingway's favorite hangout... Try the
    <Entree>ropa vieja</Entree> for <Price>$12.95</Price>
    ...
  </Review>
  <CreditCards>
    <CreditCard>AE</CreditCard>
    <CreditCard>CB</CreditCard>
    ...
  </CreditCards>
</Restaurant>
Readings for INFO 202 Lecture #4


George Lakoff. Women, Fire, and Dangerous Things. Chapters 1 and 2
Assignment 1 - "Everything is Miscellaneous"

Explain what Weinberger means by "everything" and "miscellaneous," citing (by page number) at least three passages in the book that illustrate the scope and meaning of each concept.

Take a clear stand on whether Weinberger has convinced you or not that "everything in miscellaneous." If you agree, cite (by page number) at least two arguments in the book against "authoritative organization." If you disagree, cite at least two arguments (or make two of your own) and explain why you think Weinberger's case falls short.

Your paper should be about 500-750 words (between 2 and 3 printed pages), and no cover page is necessary. Bring three copies to class on Tuesday September 10, so that each of the TAs and the professor can read your paper.
Class of 2009 Party

TONIGHT Wednesday, September 5, 6-9pm

At my house in the Berkeley hills, 34 Stephens Way

No public transportation, so this will encourage you to talk to each other