3. Organization {and,or,vs} Retrieval

INFO 202 - 8 September 2008

Bob Glushko

Plan for INFO 202 Lecture #3

Definitions and frameworks for "Information"

The Information Life Cycle

Relevance, Recall, and Precision

The IO and IR Tradeoff

What is Information?

Many different ways to define it; depends on your discipline or perspective or problem

A common sense is "something you don't know" or "news about something"

But news to one person might be "old news" to someone who already knows it

So if something happens that we expected, we get less information than if something unexpected happens

Bates "Fundamental Forms of Information"

(Journal of the American Society for Information Science & Technology, June 2006)

Information is "the pattern of organization of matter and energy"

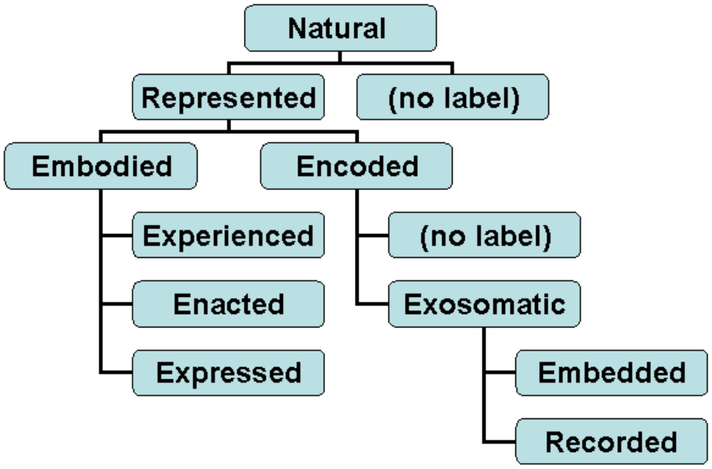
So this exists independently of any observer or recipient, but each observer can construct, store, and act upon it in different subjective ways

Environmental and evolutionary factors influence this so there are similarities in how information is experienced

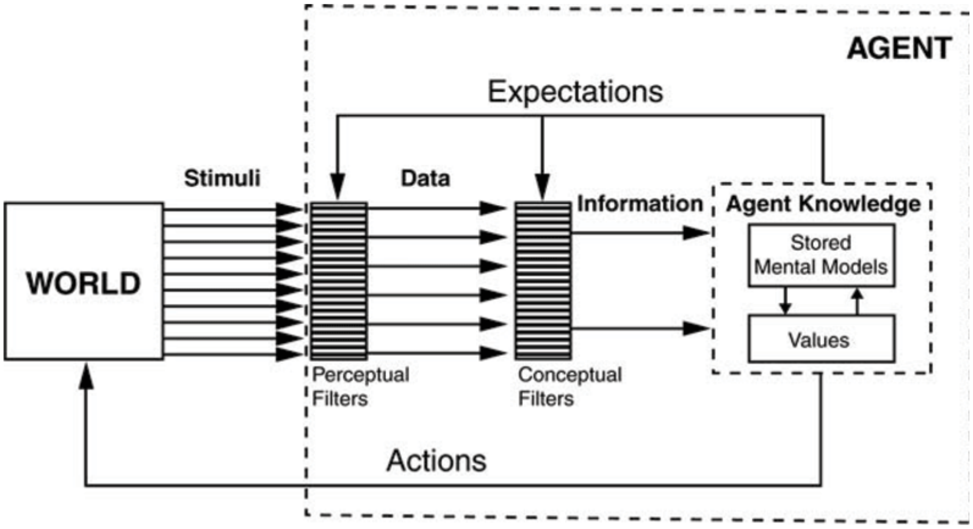
Anything human beings interact with or observe can be a source of information

Living beings can assign meaning to information, but patterns of organization of matter and energy are not inherently meaningful

Bates Taxonomy of Information Forms



Information Hierarchy Process Model (Boisot & Canals, 2004)



"Data, information and knowledge: have we got it right?" Journal of

Taylor's Information Hierarchy

Taylor presents an information hierarchy of DATA, INFORMATION, KNOWLEDGE, and WISDOM as "different levels of comprehension of symbols"

- DATA
- INFORMATION
- KNOWLEDGE
- WISDOM

Information Hierarchy Example

DATA - the value we collect from a measuring instrument is 25

INFORMATION - the instrument is a refractometer, which measures the sugar content of grapes on the Brix scale.

KNOWLEDGE - the sugar content of grapes predicts the potential alcohol content of wine fermented from them

WISDOM - If the summer weather is cooler than usual, grapes will contain more acid so growers usually pick later with a higher Brix to help balance the tart flavors from the acid

Is the Encyclopedia Britannica wisdom? Is Wikipedia? What about this encyclopedia?

Information Quality

Do you know the source of the information?

Is the source authentic?

Is the information authoritative or credible?

Information, News, Uncertainty and Truth

"I just feel fortunate to live in a world with so much disinformation at my fingertips."

Weinberger's "Three Orders of Order"

The order of things - putting physical things into places to organize them

Physical metadata - organizing things by using information recorded in physical form (printed documents, card catalogs)

Digital metadata - organizing things or information by using information recorded in digital form

Bits vs. Atoms

the distinction that Weinberger makes is built upon the contrast between "bits" and "atoms" first expressed by Nicholas Negroponte of the MIT Media Lab

Information encoded as bits can move several orders of magnitude faster than atoms can

It can be in many places at once (broadcasting, networking) unlike atoms that have to be in one place

Quotations from Weinberger

"We have organized our ideas with principles designed for use in a world limited by the laws of physics"

"The solution to the overabundance of information is more information... [but] the real world limits the amount of additional data we can supply"

"Our bookstores... work well for people trying to find what they came in for, whereas there are almost as many ways to organize for browsers as there are browsers"

"Instead of everything having its place, it's better if thing can get assigned multiple places simultaneously"

"the more ways our digital photos can be sorted, ordered, clustered, and made sense of -- the more miscellaneous they are -- the better"

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

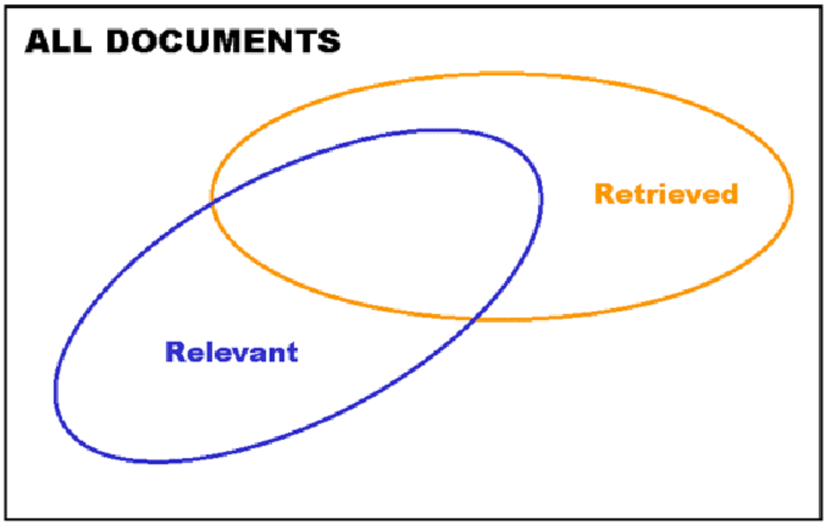
Alacrastore

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



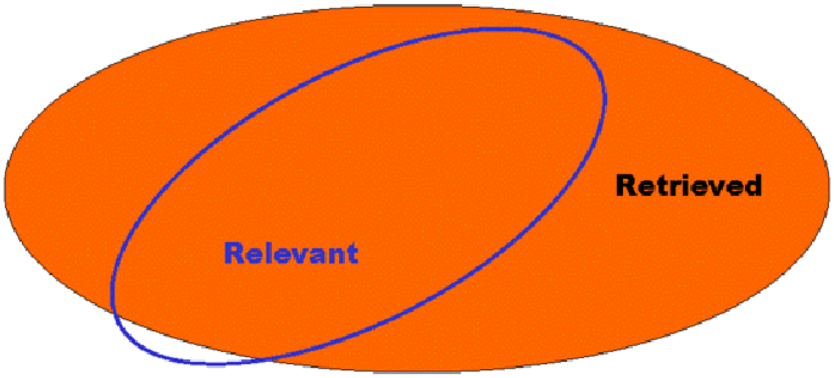
Recall and Precision [2]

RECALL is the proportion of the relevant documents that are retrieved

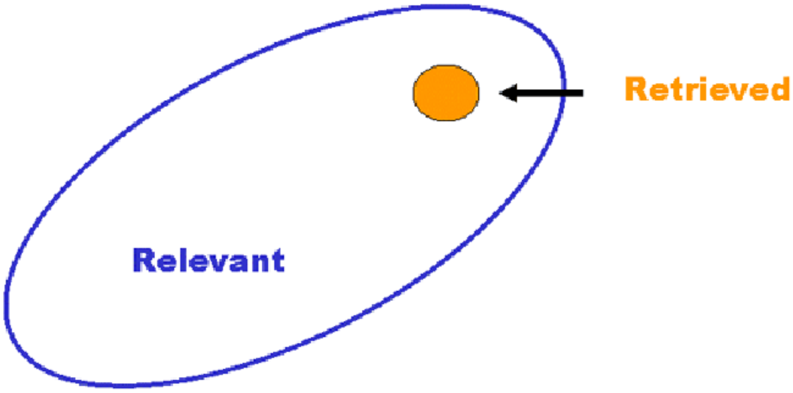
PRECISION is the proportion of the retrieved documents that are relevant

Goal: High recall and precision - Get as much good stuff as possible while getting as little junk as possible

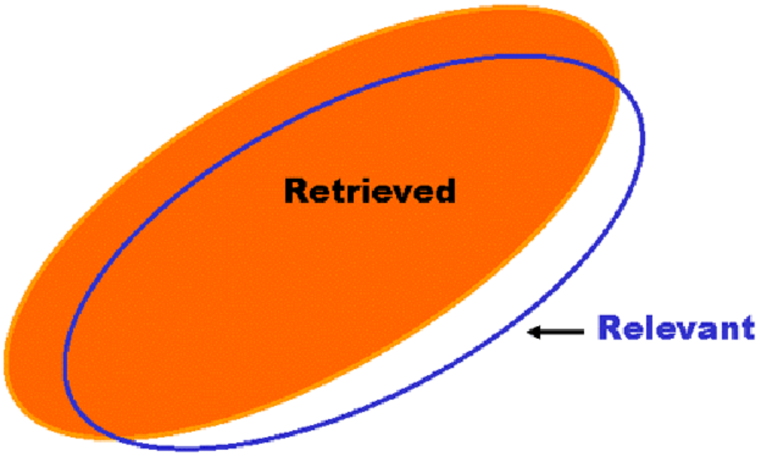
High Recall but Low Precision



Low Recall but High Precision



High Recall and High Precision



Quotations from Svenonius

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?

Organization {and,or,vs} Search [2]

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

Organization {and,or,vs} Search for Personal Information

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

Organization {and,or,vs} Search for Enterprise Information

Enterprise information is produced by people doing their jobs and its structure and quality is often governed by policy or technology support

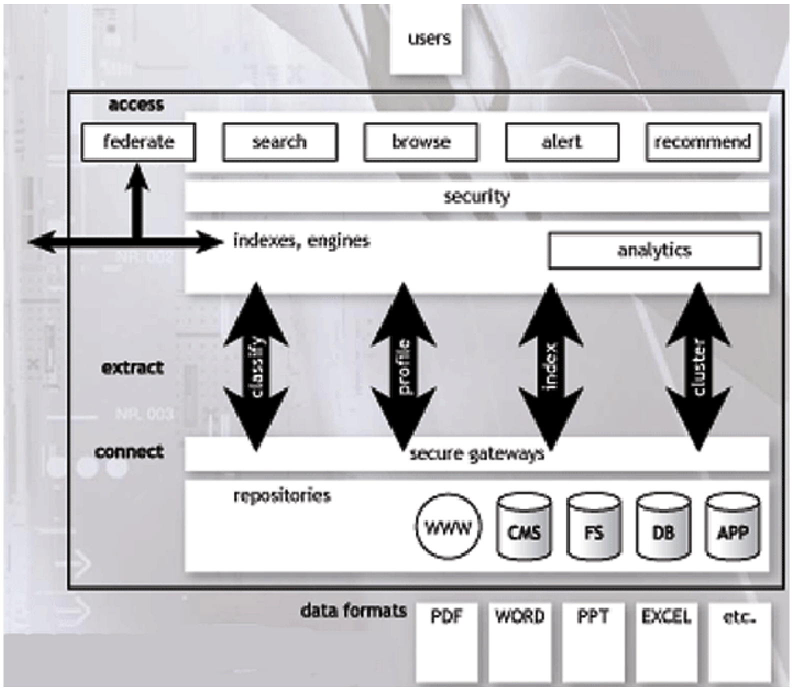
Enterprise information can exist in numerous formats and applications, many of which are hard to get at via enterprise search because they are "silos"

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

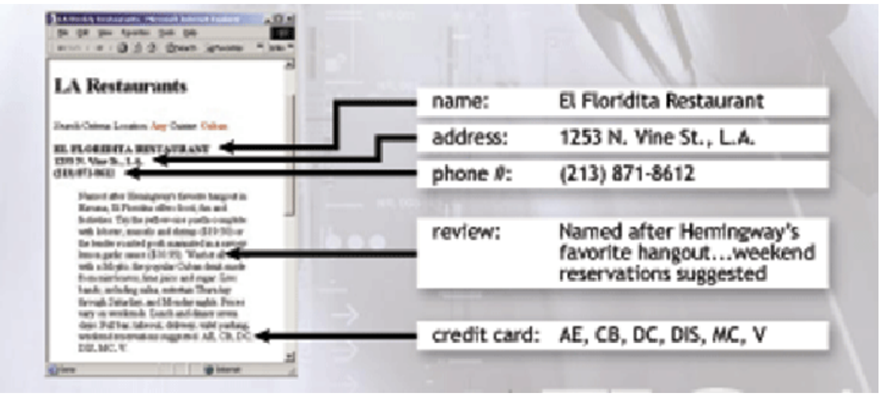
Information types span the spectrum from unstructured to semi-structured to structured

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



name: El Floridita Restaurant

address: 1253 N. Vine St., L.A.

phone #: (213) 871-8612

review: Named after Hemingway's favorite hangout…weekend reservations suggested

credit card: AE, CB, DC, DIS, MC, V

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>
```

Your First Discussion Section

M 11-12, 12-1, 4-5

Svenonius vs Weinberger - and the "tradeoff "

Party Tomorrow

6-9pm

34 Stephens Way, Berkeley

Reading for 10 September

Robert J. Glushko and Tim McGrath, Document Engineering, Chapter 2, "XML Foundations" - you can download it from:

http://people.ischool.berkeley.edu/~glushko/
DocumentEngineeringBookDraft/DEBook/ch2_FINAL.pdf