30. Course Review

INFO 202 - 10 December 2008

Bob Glushko

Today's Agenda

About the final exam on 12/15

Key concepts and themes

The most important 4% of the lecture slides

The Final Exam

Next Monday from 9am to 1pm in 202

5 of 8 short answers, 2 of 4 long answers

open book, open note, open study guide... but not open Web

IO Buzzwords

Semantics	Syntax	Semiotics
Categorization	Classification	"the Work"
Taxonomy	Folksonomy	
Vocabulary	Tag	Web 2.0
Metadata	Metamodel	Facet
Conceptual Model	Schema	Microformat
Ontology	Information Architecture	
Integration	Mash-up	
Interoperability	Compliance	Standardization

IR Buzzwords

Collection	Corpus	
Browsing	“Berrypicking”	Foraging
Recall	Precision	
Link	Anchor	Page Rank
Crawling	Data Mining	Clustering
Boolean	Bayesian	
Term Weighting	TF/IDF	
Vectors	Dimensionality Reduction	LSI

What Does it Mean to Describe Something?

Identify / scope the thing to be described

Study it to identify its important properties or features

Compare it with other things like and unlike it

Select or develop a system / vocabulary for description using "good" categories and terms

Create the descriptions, measurements, and other statistics about the object, either "by hand" or by some automated / computational process

Why Is Organizing Information Difficult?

People use different words for the same things, and the same words for different things

Organization is always (explicitly or implicitly) being done in some context

Every context produces different biases in naming and categorization

What is the Context / Scope of Information Organization?

Individual concepts and information components

Documents and databases

Personal collection

Within a group / community / enterprise

Between enterprises

"In the world"

Who Organizes Information?

Professionals (the emphasis of traditional library science)

Authors

Users ("folksonomy" and other end-user "tagging")

Machines (computer programs)

Key Points in Today's Lecture

Naming is a challenging and often contentious activity

The "uncontrolled" words that people naturally use to describe things or concepts are "embodied" in their context and experiences ... so they are often different or even "bad" with respect to the words used by others

A "controlled" vocabulary creates an artificial language to be used in place of the "natural" ones to facilitate agreement in some context and for some purposes

Instances and Classes / Categories

We can treat information as a description of, or as being attached to an INSTANCE of something, a particular realization or implementation or occurrence

Or, we can treat information as a description of a CLASS or CATEGORY of things that we are treating as equivalent

It is important to be precise about which descriptive perspective you're taking

Document Types

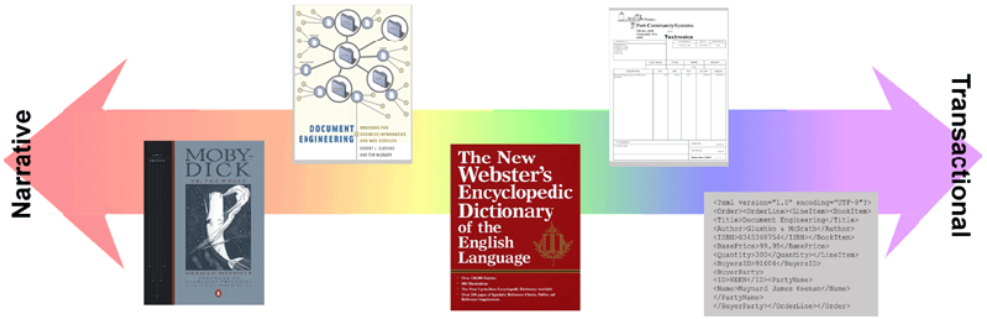
Any definition of "document" allows for a notion of different types of documents

This idea can be very intuitive and very informal, or we can be more precise and define a *model of a document type*
as the rules or constraints that distinguish one type from another

This expression of the model is *conceptual* and is independent of the syntax and technology in which document instances are ultimately implemented

But most of the time the model is ultimately implemented as a *physical* view in some specific syntax like XML

The Document Type Spectrum



Information Management and the Document Type Spectrum

XML Vocabularies

When a model of a document type is (AT BEST PARTIALLY) encoded in XML it is often called an XML "vocabulary" or XML "application"

But this conflicts with the more common usage of that word to mean "software application" so I try to always use vocabulary

The best thing about XML is the ease with which anyone can create a new vocabulary

The worst thing about XML is the same as the best thing: the ease with which anyone can create a new vocabulary

Summary: Why Study Categorization?

Categorization is central to how we organize information and the world, and categories are involved whenever we communicate, analyze, predict, or classify

- Informally with "cultural" categories
- Formally with "institutional" categories
- Whenever we design data structures, programming language class hierarchies, user or application interfaces, ...

Categorization is much messier than our computer systems and applications would like

But understanding how people (and each of us) categorize can help us design better systems and interfaces

Every Classification is "Biased"

Every classification system takes a point of view

Every classification system implicitly or explicitly distinguishes between "good" or "standard" and "bad" or "nonstandard" ways of understanding things

Categories for CAFE

Quotes 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)

Defining What Something Means

By Reference or Enumeration

Definition

Definition in a controlled vocabulary

Data types

Metadata

Metamodels

Formal assertions

Ontologies and thesauri

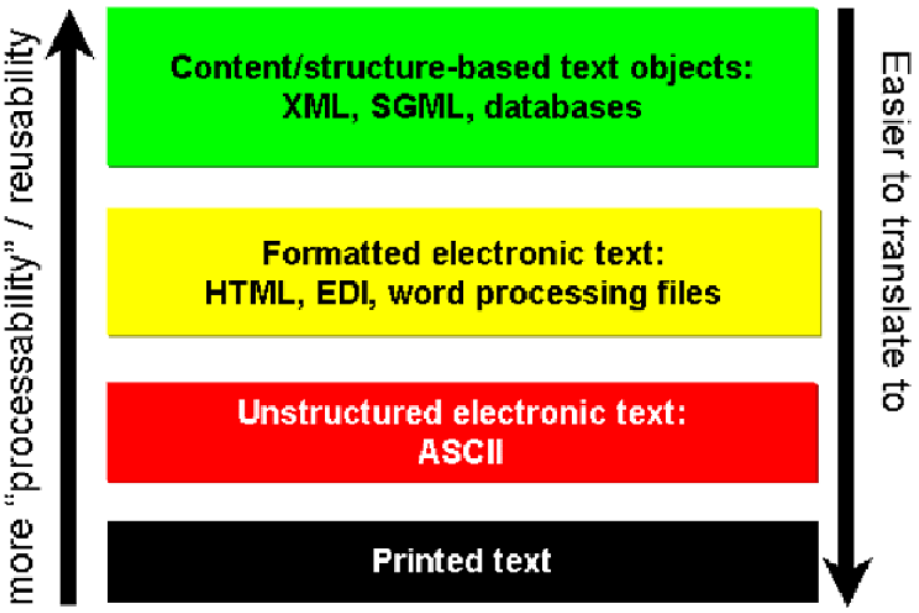
Not All Containers Are Equally Separable, Valuable or Usable

Not all information content can be separated from its container (sometimes the medium is the message)

But it is important to think of the information content abstractly if you can because that's the key to representing the same information in multiple formats, media, or technologies

Some information formats or representations are inherently more reusable or adaptable than others

"Information IQ"



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

But Svenonius Warns Us...

Because the choice objective is capable of spawning description *ad infinitum*, it is economically untenable (p. 23)

Attempts to cope with the unwanted economic consequences of open-ended objectives surface periodically as a rethinking of a "core" set of essential metadata to be used in description (p. 23)

An important question is whether the bibliographic universe can be organized both intelligently (to meet the traditional bibliographic objectives) and automatically (p. 25)

Any task that requires an organizing intelligence to engage in research is costly (p. 26)

How Much Metadata, What Kind, and by Whom?

You must consider the tradeoffs between organization and retrieval

Not all documents / resources need the same amount of metadata

The same metadata elements or attributes might need different amounts of semantic precision for different document types or contexts (*"A laxer form of vocabulary control"*-- Svenonius p. 26)

Doctorow on Metadata

People lie

People are lazy

People are stupid

People delude themselves

Metadata metrics distort it

Metadata suffers from "the vocabulary problem"

The Modeling Debate [1]

Some problems and some domains are inherently complex and a careful, rigorous modeling approach is required

- This "heavyweight" position argues that there are no modeling shortcuts

But some people argue that modeling "involves a substantial amount of work that is often political, tedious, and unpleasant" that should be avoided whenever possible

- Some domains and use cases might be simple enough ("Microformats") that less "heavyweight" modeling approaches could suffice

The Modeling Debate [2]

You should always look to see if someone has already modeled your problem domain (Cover Pages and OASIS)

If the underlying conceptual model of an existing vocabulary doesn't fit your requirements and you must develop your own, you have many choices to make about scope, abstraction, and granularity

Enterprise Information & Data Management Goals

Run the business more efficiently through greater automation and "straight through" or "end-to-end" processing of the information that it creates and receives

Consolidation of data from multiple business units to create a unified view of the {customer, supply chain, etc}

Get end-to-end visibility of business processes

Take different perspectives (from high level aggregation to resolving individual data anomalies or inconsistency)

Make better decisions more quickly

Enterprise (and Inter-enterprise) Information & Data Management Challenges

Internal to a firm, application "silos" or "stovepipes" may have been created over time and not have been designed to share information with each other

Each of these systems has a specific purpose and a data model customized for that purpose - so these models may be incomplete or incompatible with respect to each other

This causes problems when business processes can span multiple departments, business applications, or even multiple firms

These problems are greater when information or data comes from outside the firm

Semantic Integration – the Transformation Requirement

Information can't be reliably exchanged between systems to integrate business processes or support decision making unless semantics are unambiguous

Transformation of the incoming information is often required before it can make sense to the receiving application or service

Semantic integration is the process by which a common semantic "data model" or "object model" is created through transformation

What's the most powerful semantic integration processor?

Social Categorization in the Enterprise

Tagging and bookmarking are being adapted for use in business organizations and large enterprises

Some significant differences with "open web" categorization

- Every user is authenticated to a "real" identity
- Organizational norms and incentives restrict/shape the purposes and nature of the categorization

These applications can capture expertise and interests implicitly and at lower cost than traditional knowledge management applications

Models of Information Retrieval [1]

The core problems of information retrieval are finding relevant documents and ordering the found documents according to relevance

The IR model explains how these problems are solved:

- ...By specifying the representations of queries and documents in the collection being searched
- ...And the information used, and the calculations performed, that order the retrieved documents by relevance
- (And optionally, the model provides mechanisms for using relevance feedback to improve precision and results ordering)

Models of Information Retrieval [2]

BOOLEAN model -- representations are sets of index terms, set theory operations with Boolean algebra calculate relevance as binary

VECTOR models -- representations are vectors with non-binary weighted index terms, linear algebra operations yield continuous measure of relevance

Models of Information Retrieval [3]

STRUCTURE models -- combine representations of terms with information about structures within documents (i.e., hierarchical organization) and between documents (i.e. hypertext links and other explicit relationships) to determine which parts of documents and which documents are most important and relevant

PROBABILISTIC models -- documents are represented by index terms, and the key assumption is that the terms are distributed differently in relevant and non relevant documents.

Motivating Term Weighting from the Boolean Model

The Boolean model represents documents as a set of index terms that are either present or absent

This binary notion doesn't fit our intuition that terms differ in how much they suggest what the document is about

We will capture this notion by assigning weights to each term in the index

Same Idea, for Left-Brain Folks

Keywords, index terms, controlled vocabulary terms -- are not strictly properties of any single document. They reflect a relationship between an individual document and the set of documents it belongs to, from which it might be selected

The value of a potential keyword varies inversely with the number of documents in which it occurs -- the most informative words are those that occur infrequently but when they occur they occur in clusters, with most of the occurrences in a small number of documents out of the collection

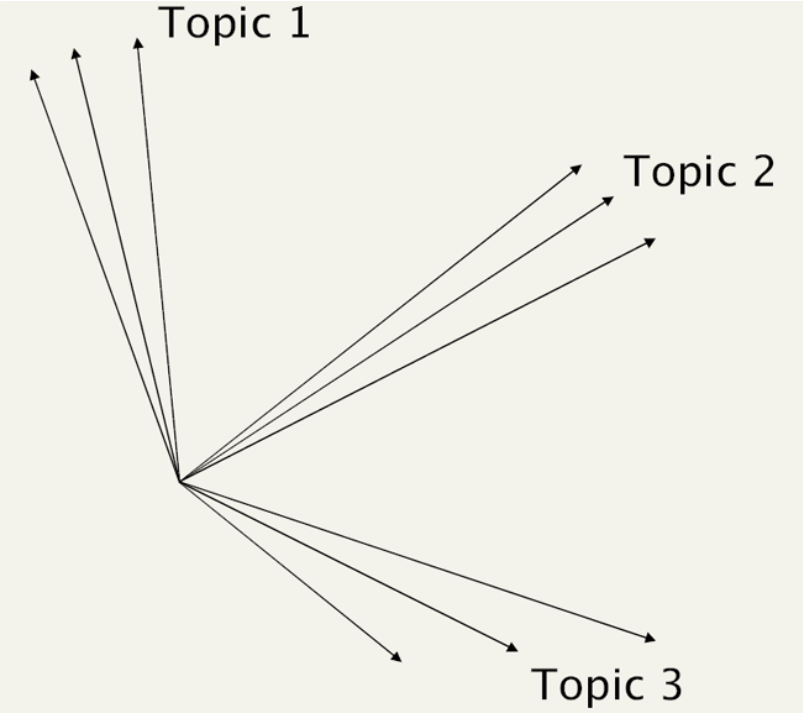
Dimensionality Reduction to Improve the Vector Model

The dimensionality of the space in the simple vector model is the number of different terms in it

But the "semantic dimensionality" of the space is number of distinct topics represented in it

The number of topics is much lower than the number of terms (in a given collection, untapped synonymy is more important than unnoticed polysymy)

"Topic Space," Not "Term Space"



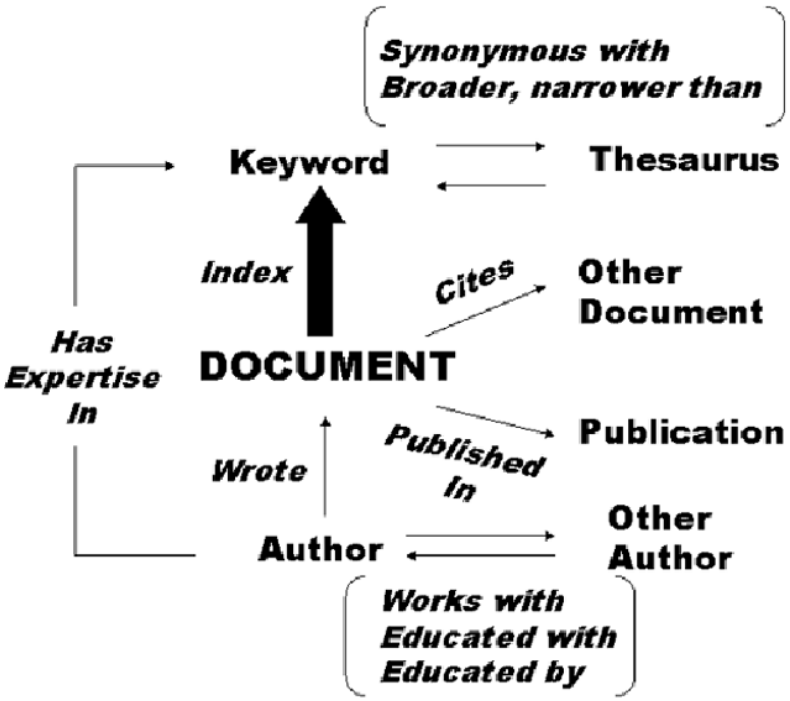
Structure-Based IR Models

Combine representations of terms with information about structures within documents (i.e., hierarchical organization) and between documents (i.e. hypertext links and other explicit relationships)

This structural information tells us what documents and parts of documents are most important and relevant

This information gives us additional justification for determining relevance and ordering a result set

The Index in Context



Moving Beyond "Reachability"

We've been treating linking in purely structural terms - is one thing connected to another - but we can refine that into two perspectives:

- *Relational*
  analysis treats links as indicators of the amount of connectedness or the direction of flow between documents, people, groups, journals, disciplines, domains, organizations, or nations
- *Evaluative*
  analysis treats links as indicators of the level of quality, importance, influence or performance of documents, people, groups. ...

Real World Applications

Machine Translation

Spelling Suggestions/Corrections

Grammar Checking

Speech Processing

Text Categorization and Clustering

Combining Linguistic and Statistical Approaches

Both the linguistic and data-driven or statistical approaches are seen as integral and complementary parts of an NLP application

Systems employ sophisticated techniques for dictionaries and grammars to identify parts of speech and do morphological analysis

But the statistics of co-occurrence / conditional probability yield many practical techniques for estimating the substitutability or semantic equivalence of words in larger text segments that make no use of their "languageness"

In particular, the web is such a huge corpus that statistical approaches can be surprisingly informative and robust

Why This Course Is Challenging / Unique / Essential

We deal with deep intellectual issues that have challenged philosophers and other deep thinkers for millennia

You must making the transition to studying information / content IN a discipline to studying information / content AS a discipline

You must learn to look past the presentation / rendition / technology reification / thinginess of information to see it more abstractly as structure and meaning

The diversity of perspectives and backgrounds in your class will be one of the challenges in this course

The One Minute Course

It is important to think ABSTRACTLY about information and the conceptual tasks of organizing and finding it

We experience information as individuals, in association with other individuals, or as part of a business or business ecosystem

We must recognize the profound impact of new technologies and their co-evolution with IO and IR techniques

But we must still appreciate the "classical" knowledge that is often ignored by those chasing new technologies

This is Not the End of 202

It IS the end of the semester: 1500 pages, 8 assignments, 29 lectures, and 10 section meetings

But how you think about information has changed immensely

And that will shape the remainder of your ISchool experience and the rest of your life