

LBSC 690: Week 11

Information Retrieval and Search



Jimmy Lin
College of Information Studies
University of Maryland

Monday, April 16, 2007

What is IR?

- “Information?”
 - How is it different from “data”?
 - Information is data in context
 - Not necessarily text!
 - How is it different from “knowledge”?
 - Knowledge is a basis for making decisions
 - Many “knowledge bases” contain decision rules
- “Retrieval?”
 - Satisfying an information need
 - “Scratching an information itch”

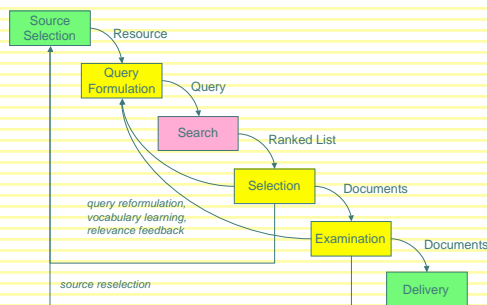
What types of information?

- Text (Documents and portions thereof)
- XML and structured documents
- Images
- Audio (sound effects, songs, etc.)
- Video
- Source code
- Applications/Web services

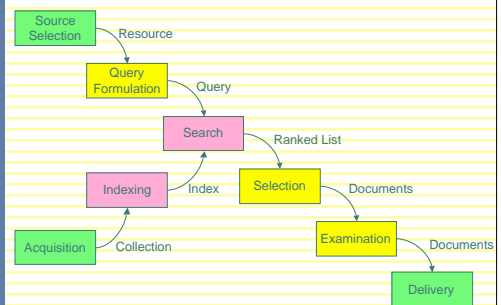
The Big Picture

- The four components of the information retrieval environment:
 - User (user needs)
 - Process
 - System
 - Data
- What computer geeks care about! What we care about!

The Information Retrieval Cycle



Supporting the Search Process



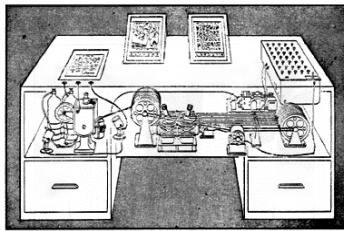
How is the Web Indexed?

- Spiders and crawlers
- Robot exclusion
- Deep vs. Surface Web

Modern History

- The “information overload” problem is much older than you may think
- Origins in period immediately after World War II
 - Tremendous scientific progress during the war
 - Rapid growth in amount of scientific publications available
- The “Memex Machine”
 - Conceived by Vannevar Bush, President Roosevelt's science advisor
 - Outlined in 1945 Atlantic Monthly article titled “As We May Think”
 - Foreshadows the development of hypertext (the Web) and information retrieval system

The Memex Machine



Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify microfilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (LIFE 1933, p. 120).

Types of Information Needs

- Retrospective
 - “Searching the past”
 - Different queries posed against a static collection
 - Time invariant
- Prospective
 - “Searching the future”
 - Static query posed against a dynamic collection
 - Time dependent

Retrospective Searches (I)

- *Ad hoc* retrieval: find documents “about this”
 - Identify positive accomplishments of the Hubble telescope since it was launched in 1991.
 - Compile a list of mammals that are considered to be endangered, identify their habitat and, if possible, specify what threatens them.
- Known item search
 - Find Jimmy Lin's homepage.
 - What's the ISBN number of “Modern Information Retrieval”?
- Directed exploration
 - Who makes the best chocolates?
 - What video conferencing systems exist for digital reference desk services?

Retrospective Searches (II)

- Question answering
 - “Factoid”** Who discovered Oxygen?
When did Hawaii become a state?
Where is Ayer's Rock located?
What team won the World Series in 1992?
 - “List”** What countries export oil?
Name U.S. cities that have a “Shubert” theater.
 - “Definition”** Who is Aaron Copland?
What is a quasar?

Prospective "Searches"

- Filtering
 - Make a binary decision about each incoming document
 - Spam or not spam?
- Routing
 - Sort incoming documents into different bins?
 - Categorize news headlines: World? Nation? Metro? Sports?

Why is IR hard?

- Why is it so hard to find the text documents you want?
- What's the problem with language?
 - Ambiguity
 - Synonymy
 - Polysemy
 - Morphological Variation
 - Paraphrase
 - Anaphora
 - Pragmatics

"Bag of Words" Representation

- Bag = a "set" that can contain duplicates
 - "The quick brown fox jumped over the lazy dog's back"
 - {back, brown, dog, fox, jump, lazy, over, quick, the, the}
- Vector = values recorded in any consistent order
 - {back, brown, dog, fox, jump, lazy, over, quick, the}
 - [1 1 1 1 1 1 1 1 2]

Bag of Words Example

Document 1	Term	Document 1	Document 2
The quick brown fox jumped over the lazy dog's back.	aid	0	1
	all	0	1
	back	1	0
	brown	1	0
	come	0	1
	dog	1	0
	fox	1	0
	good	0	1
	jump	1	0
	lazy	1	0
	men	0	1
	now	0	1
	over	1	0
	party	0	1
	quick	1	0
	their	0	1
	time	0	1
Document 2			
Now is the time for all good men to come to the aid of their party.			

Stopword List
for
is
of
the
to

Boolean "Free Text" Retrieval

- Limit the bag of words to "absent" and "present"
 - "Boolean" values, represented as 0 and 1
- Represent terms as a "bag of documents"
 - Same representation, but rows rather than columns
- Combine the rows using "Boolean operators"
 - AND, OR, NOT
- Result set: every document with a 1 remaining

Boolean Free Text Example

Term	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
aid	0	0	0	1	0	0	0	1
all	0	1	0	1	0	1	0	0
back	1	0	1	0	0	0	1	0
brown	1	0	1	0	1	0	1	0
come	0	1	0	1	0	1	0	1
dog	0	0	1	0	1	0	0	0
fox	0	0	1	0	1	0	1	0
good	0	1	0	1	0	1	0	1
jump	0	0	1	0	0	0	0	0
lazy	1	0	1	0	1	0	1	0
men	0	1	0	1	0	0	0	1
now	0	1	0	0	0	1	0	1
over	1	0	1	0	1	0	1	1
party	0	0	0	0	0	1	0	1
quick	1	0	1	0	0	0	0	0
their	1	0	0	0	1	0	1	0
time	0	1	0	1	0	1	0	0

- dog AND fox
 - Doc 3, Doc 5
- dog NOT fox
 - Empty
- fox NOT dog
 - Doc 7
- dog OR fox
 - Doc 3, Doc 5, Doc 7
- good AND party
 - Doc 6, Doc 8
- good AND party NOT over
 - Doc 6

Why Boolean Retrieval Works

- Boolean operators approximate natural language
 - Find documents about a good party that is not over
- AND can discover relationships between concepts
 - good party
- OR can discover alternate terminology
 - excellent party
- NOT can discover alternate meanings
 - Democratic party

The Perfect Query Paradox

- Every information need has a perfect set of documents
 - If not, there would be no sense doing retrieval
- Every document set has a perfect query
 - AND every word to get a query for document 1
 - Repeat for each document in the set
 - OR every document query to get the set query
- But can users realistically expect to formulate this perfect query?
 - Boolean query formulation is hard!

Why Boolean Retrieval Fails

- Natural language is way more complex
 - She saw the man on the hill with a telescope
- AND "discovers" nonexistent relationships
 - Terms in different paragraphs, chapters, ...
- Guessing terminology for OR is hard
 - good, nice, excellent, outstanding, awesome, ...
- Guessing terms to exclude is even harder!
 - Democratic party, party to a lawsuit, ...

Proximity Operators

- More precise versions of AND
 - "NEAR n" allows at most n-1 intervening terms
 - "WITH" requires terms to be adjacent and in order
- Easy to implement, but less efficient
 - Store a list of positions for each word in each doc
 - Stopwords become very important!
 - Perform normal Boolean computations
 - Treat WITH and NEAR like AND with an extra constraint

Boolean Retrieval

- Strengths
 - Accurate, if you know the right strategies
 - Efficient for the computer
- Weaknesses
 - Often results in too many documents, or none
 - Users must learn Boolean logic
 - Sometimes finds relationships that don't exist
 - Words can have many meanings
 - Choosing the right words is sometimes hard

Ranked Retrieval Paradigm

- Some documents are more relevant to a query than others
 - Not necessarily true under Boolean retrieval!
- "Best-first" ranking can be superior
 - Select n documents
 - Put them in order, with the "best" ones first
 - Display them one screen at a time
 - Users can decide when they want to stop reading

Ranked Retrieval: Challenges

- "Best first" is easy to say but hard to do!
 - The best we can hope for is to approximate it
- Will the user understand the process?
 - It is hard to use a tool that you don't understand
- Efficiency becomes a concern

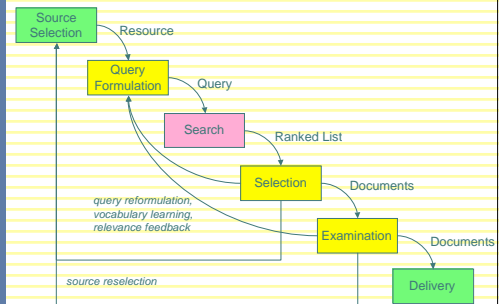
Similarity-Based Queries

- Create a query "bag of words"
- Find the similarity between the query and each document
 - For example, count the number of terms in common
- Rank order the documents by similarity
 - Display documents most similar to the query first
- Surprisingly, this works pretty well!

Counting Terms

- Terms tell us about documents
 - If "rabbit" appears a lot, it may be about rabbits
- Documents tell us about terms
 - "the" is in every document: not discriminating
- Documents are most likely described well by rare terms that occur in them frequently
 - Higher "term frequency" is stronger evidence
 - Low "collection frequency" makes it stronger still

The Information Retrieval Cycle



Search Output

- What now?
 - User identifies relevant documents for "delivery"
 - User issues new query based on content of result set
- What can the system do?
 - Assist the user to identify relevant documents
 - Assist the user to identify potentially useful query terms

Selection Interfaces

- One dimensional lists
 - What to display? title, source, date, summary, ratings, ...
 - What order to display? retrieval status value, date, alphabetic, ...
 - How much to display? number of hits
 - Other aids? related terms, suggested queries, ...
- Two+ dimensional displays
 - Clustering, projection, contour maps, VR
 - Navigation: jump, pan, zoom

Query Enrichment

- Relevance feedback
 - User designates "more like this" documents
 - System adds terms from those documents to the query
- Manual reformulation
 - Initial result set leads to better understanding of the problem domain
 - New query better approximates information need
- Automatic query suggestion

Example Interfaces

- Google: keyword in context
- Microsoft Live: query refinement suggestions
- Exalead: faceted refinement
- Vivisimo: clustered results
- Kartoo: cluster visualization
- WebBrain: structure visualization
- Grokker: "map view"

Evaluating IR Systems

- User-centered strategy
 - Given several users, and at least 2 retrieval systems
 - Have each user try the same task on both systems
 - Measure which system works the "best"
- System-centered strategy
 - Given documents, queries, and relevance judgments
 - Try several variations on the retrieval system
 - Measure which ranks more good docs near the top

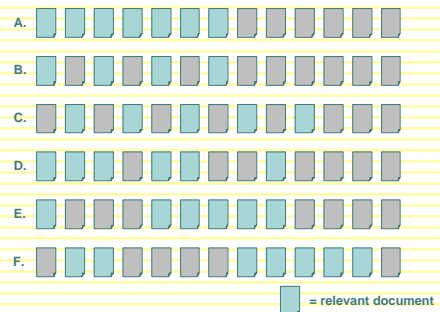
Good Effectiveness Measures

- Capture some aspect of what the user wants
- Have predictive value for other situations
 - Different queries, different document collection
- Easily replicated by other researchers
- Easily compared
 - Optimally, expressed as a single number

Defining "Relevance"

- Hard to pin down: a central problem in information science
- Relevance relates a topic and a document
 - Not static
 - Influenced by other documents
- Two general types
 - Topical relevance: is this document about the correct subject?
 - Situational relevance: is this information useful?

Which is the Best Rank Order?



Set-Based Measures

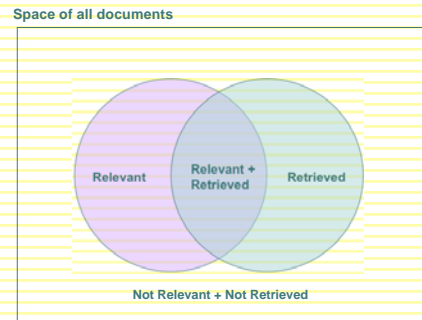
	Relevant	Not relevant
Retrieved	A	B
Not retrieved	C	D

Collection size = $A+B+C+D$
 Relevant = $A+C$
 Retrieved = $A+B$

- **Precision** = $A \div (A+B)$
- **Recall** = $A \div (A+C)$
- **Miss** = $C \div (A+C)$
- **False alarm (fallout)** = $B \div (B+D)$

When is precision important?
 When is recall important?

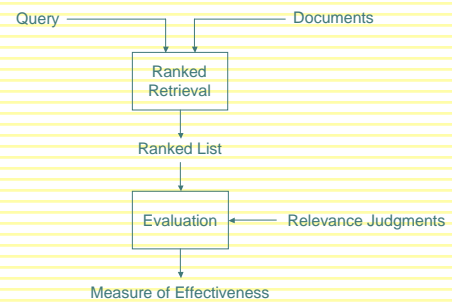
Another View



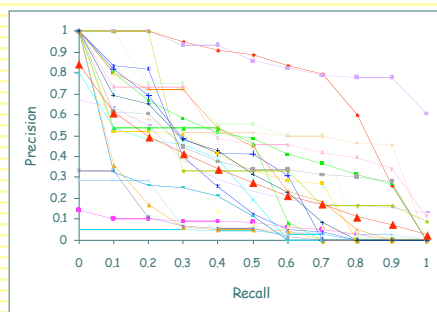
Precision and Recall

- **Precision**
 - How much of what was found is relevant?
 - Often of interest, particularly for interactive searching
- **Recall**
 - How much of what is relevant was found?
 - Particularly important for law, patents, and medicine

Abstract Evaluation Model



ROC Curves



User Studies

- Goal is to account for interface issues
 - By studying the interface component
 - By studying the complete system
- **Formative evaluation**
 - Provide a basis for system development
- **Summative evaluation**
 - Designed to assess performance

Quantitative User Studies

- Select independent variable(s)
 - e.g., what info to display in selection interface
- Select dependent variable(s)
 - e.g., time to find a known relevant document
- Run subjects in different orders
 - Average out learning and fatigue effects
- Compute statistical significance
 - Null hypothesis: independent variable has no effect
 - Rejected if $p < 0.05$

Qualitative User Studies

- Observe user behavior
 - Instrumented software, eye trackers, etc.
 - Face and keyboard cameras
 - Think-aloud protocols
 - Interviews and focus groups
- Organize the data
 - For example, group it into overlapping categories
- Look for patterns and themes
- Develop a "grounded theory"

Questionnaires

- Demographic data
 - For example, computer experience
 - Basis for interpreting results
- Subjective self-assessment
 - Which did they think was more effective?
 - Often at variance with objective results!
- Preference
 - Which interface did they prefer? Why?

By now you should know...

- Why information retrieval is hard
- Why information retrieval is more than just querying a search engine
- The difference between Boolean and ranked retrieval (and their advantages/disadvantages)
- Basics of evaluating information retrieval systems