

Computational Linguistics 1

CMSC/LING 723, LBSC 744



Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland

Lecture 25: 6 December 2011

Agenda

- HW6 due today!
- Questions, comments, concerns?
- Information Retrieval (IR)
- Question Answering (QA)

Information Retrieval

- What is the task?
 - Given a query and a document collection, return a (ranked) set of documents
 - Documents should be relevant to the query
 - Variations on task
 - Passages instead of documents
- Approach generally assumes
 - Indexing words and documents where they appear
 - Evaluation of 'distance' from query to document
 - Some kind of 'term weighting' to derive distance

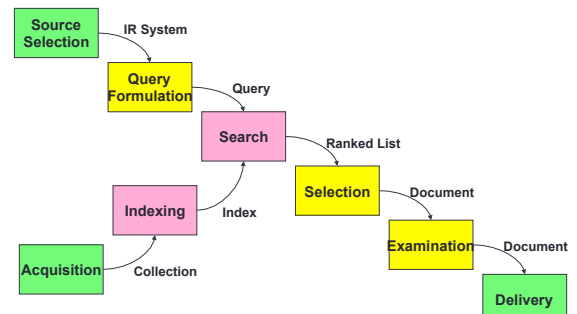
Uses of IR

- Find something that you want
 - The information need may or may not be **explicit**
- Known item search
 - Find the class home page
- Answer seeking
 - Is Lexington or Louisville the capital of Kentucky?
- Directed exploration
 - Who makes videoconferencing systems?

The Big Picture

- The four components of the information retrieval environment:
 - User (user needs)
 - Process
 - System
 - Data

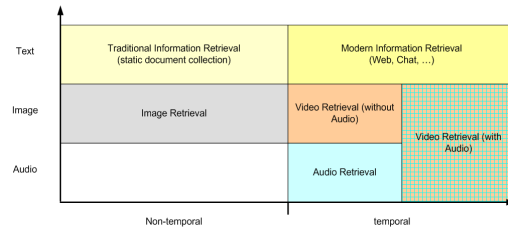
Supporting the Search Process



Data Collection/Ways of Finding Text

- Searching metadata
 - Using controlled or uncontrolled vocabularies
- Searching content
 - Characterize documents by the words they contain
- Searching behavior
 - User-Item: Find similar users
 - Item-Item: Find items that cause similar reactions

Information Retrieval Types



Information Retrieval

- Indexing
 - Indexing words and documents where they appear
- Distance from query to document
 - Evaluation of 'distance' from query to document
- Term weighting
 - Some kind of 'term weighting' to derive distance

Indexing the Data Collection

- For large data collection, don't want to re-process with every query
- Off-line (prior to query), build an *inverted index*
 - For each word, collect the documents within which word occurs (the "postings")
 - Store for easy access when given query words
- Simplest use: for a set of query terms, return the intersection of their postings
 - For example: "calcium regulation"
 - Find postings of calcium C and regulation R
 - Return $C \cap R$

Index Quality

- Crawl quality
 - Comprehensiveness, dead links, duplicate detection
- Document analysis
 - Frames, metadata, imperfect HTML, ...
- Document extension
 - Anchor text, source authority, category, language, ...
- Document restriction (ephemeral text suppression)
 - Banner ads, keyword spam, ...

"Exact Match" Retrieval

- Simplest use of an index:
 - For a set of query terms, return the intersection of their postings
 - For example: "calcium regulation"
 - Find postings of calcium C and regulation R
 - Return $C \cap R$
- Find all documents with some characteristic
 - Indexed as "Presidents -- United States"
 - Containing the words "Clinton" and "Peso"
 - Read by my boss
- A set of documents is returned
 - Hopefully, not too many or too few
 - Usually listed in date or alphabetical order

Ranked Retrieval

- Put most useful documents near top of a list
 - Possibly useful documents go lower in the list
- Users can read down as far as they like
 - Based on what they read, time available, ...
- Provides useful results from weak queries
 - Untrained users find exact match harder to use

Limitations of Simple Retrieval

- Too many documents may be returned
 - e.g., for "calcium regulation" $|C| = x, |R| = y, |C \cap R| = z$
- Or not enough
 - e.g., if $C \cap R = \emptyset$
- Even if a manageable number of documents, need to rank them
- Both problems are handled if we have a way of scoring the **relevance** of documents in $C \cap R$ to the query
 - Assume "most relevant" doc = most similar to query
 - Just return the top k in ranked order
 - Vector space model to measure document/query distance

Information Retrieval

- Indexing
 - Indexing words and documents where they appear
- Distance from query to document
 - Evaluation of 'distance' from query to document
- Term weighting
 - Some kind of 'term weighting' to derive distance

Vector Space Model

- Each document has a vector
 - Dimensions represent words
 - Weights in each dimension related to frequency
 - Vector typically length normalized
- For retrieval, find vectors close to query vector
 - cosine similarity
 - Assuming normalized n -dimensional vectors, for query q and document d

$$\cos(q, d) = \sum_{i=1}^n q[i]d[i]$$

Simple Example: Counting Words

Query: recall and fallout measures for information retrieval

Documents:

- 1: Nuclear fallout contaminated Texas.
- 2: Information retrieval is interesting.
- 3: Information retrieval is complicated.

| | 1 | 2 | 3 | Query |
|--------------|---|---|---|-------|
| complicated | | | 1 | |
| contaminated | 1 | | | |
| fallout | 1 | | | 1 |
| information | | 1 | 1 | 1 |
| interesting | | 1 | | |
| nuclear | 1 | | | |
| retrieval | | 1 | 1 | 1 |
| Texas | 1 | | | |

Vector-Space for Example Query

- To find best fit:
 - Create vector space and produce document/query vectors
 - Measure cosine similarity between document and query vectors
 - Rank documents by cosine from query

More complicated example...

Cloning of Cardiac, Kidney, and Brain Promoters of the Feline *ncx1* Gene* (Received for publication, January 2, 1997, and in revised form, February 19, 1997) Kimberly V. Barnes, Guangmao Cheng, Myra M. Dawson and Donald R. Menick From the Cardiology Division, Department of Medicine and the Gazes Cardiac Research Institute, Medical University of South Carolina, Charleston, South Carolina 29425-2221

ABSTRACT

The Na⁺-Ca²⁺ exchanger (NCX1) plays a major role in calcium efflux and therefore in the control and regulation of intracellular calcium in the heart. The exchanger has been shown to be regulated at several levels including transcription. NCX1 mRNA levels are up-regulated in both cardiac hypertrophy and failure. In this work, the 5-end of the *ncx1* gene has been cloned to study the mechanisms that mediate hypertrophic stimulation and cardiac expression. The feline *ncx1* gene has three exons that encode 5-untranslated sequences that are under the control of three tissue-specific promoters. The cardiac promoter drives expression in cardiocytes, but not in mouse 1 cells. Although it contains at least one enhancer (2000 to 1250 base pairs (bp)) and one or more negative elements (1250 to 250 bp), a minimum promoter (250 to +200 bp) is sufficient for cardiac expression and -adrenergic stimulation.

Challenges in more complicated example

- Words occur more than once – what's in the vector?
 - The more frequent, the more important, correct?
 - What are the most frequent words in this document?
- Related words (regulation, regulated) are counted separately
- Orthographic and notational variants of named entities, e.g., "Na⁺-Ca²⁺", "NCX1", "ncx1"
- Punctuation attached to words; sentence initial caps
- Noise and garbage in text (see word counts)
- Boils down to two issues
 - **term weighting** and text normalization

Information Retrieval

- Indexing
 - Indexing words and documents where they appear
- Distance from query to document
 - Evaluation of 'distance' from query to document
- Term weighting
 - Some kind of 'term weighting' to derive distance

Term-Weighting

- Assume "most relevant" doc = most similar to query
 - But not all terms should be weighted equally
- **Term-weighting** based on two criteria:
 - Repeated words are good cues to meaning
 - Rarely used words make searches more selective
- Compare weights with query
 - Add up the weights for each query term
 - Put the documents with the highest total first

Which Terms to Emphasize?

- Major factors
 - Uncommon terms are more selective
 - Repeated terms provide evidence of meaning
- Adjustments
 - Give more weight to terms in certain positions
 - Title, first paragraph, etc.
 - Give less weight each term in longer documents
 - Ignore documents that try to "spam" the index
 - Invisible text, excessive use of the "meta" field, ...

TF*IDF

- Stands for "term frequency times inverse document frequency"
 - Term frequency: number of times word appears in document(s)
 - Document frequency: number of documents word appears in at least once
- One of the best known techniques in all of NLP
 - Karen Sparck-Jones (1972) "A statistical interpretation of term specificity and its application in retrieval"
- A *family* of approaches that
 - Rewards the frequency of a term in the document(s) of interest
 - Penalizes the document frequency of a term in the total (static) collection
 - Exactly how they are combined or scaled can differ widely

TF*IDF Example

- Let $c_{ij} = c_i(w_j)$ be the count of word w_j in document d_i
- Let $f_j = |\{d_i : c_{ij} > 0\}|$, the number of documents with w_j
- Let D be the total number of documents
- Then the suggested weight (in Manning and Schuetze) is

$$\text{weight}(i, j) = \begin{cases} (1 + \log c_{ij}) \log \frac{D}{f_j} & \text{if } c_{ij} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

- If you try a few examples, it does have the desired effect
- Other options are log-likelihood ratio and log-odds


Information Retrieval


- Indexing
 - Indexing words and documents where they appear
- Distance from query to document
 - Evaluation of 'distance' from query to document
- Term weighting
 - Some kind of 'term weighting' to derive distance
- Evaluation


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


Which is the Best Rank Order?


A. 


B. 

C. 

D. 

E. 

F. 

 = relevant document

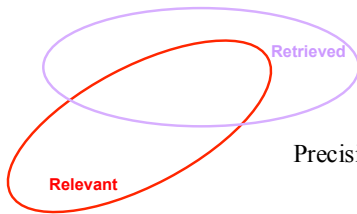
Evaluation

- What can be measured that reflects the searcher's ability to use a system? (Cleverdon, 1966)
 - Coverage of information
 - Form of presentation
 - Effort required/Ease of use
 - Time and space efficiency
- Effectiveness
- Recall
 - Precision

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

Measures of Effectiveness

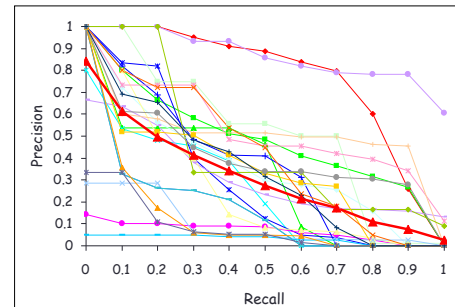


$$\text{Precision} = \frac{|\text{Ret} \cap \text{Rel}|}{|\text{Ret}|}$$

$$\text{Recall} = \frac{|\text{Ret} \cap \text{Rel}|}{|\text{Rel}|}$$

slide from Bonnie Door

Precision-Recall Curves



Source: Ellen Voorhees, NIST

IR: Summary

- Search is a process engaged in by people
- Content and behavior offer useful evidence
- Relation between query and documents is complex
 - Every word is not equal
- Evaluation must consider many factors

Computational Linguistics 1

41

Agenda

- HW6 due today!
- Questions, comments, concerns?
- Information Retrieval (IR)
- Question Answering (QA)

Computational Linguistics 1

42

Question Answering

- Canonical task: factoid question answering
 - e.g., Who was the first pick of the Portland Trailblazers in the 2007 NBA draft?
- Many different possible answer types to such questions
 - Proper name: *Greg Oden* (can be person, organization, etc.)
 - Dates: June 5, 1977
 - Quantities or measurements: 7 feet tall
- Granularity of answer is relatively small
 - Requires information extraction techniques
- Ultimately a somewhat limited class of questions

Computational Linguistics 1

slide from Brian Roark

43

Factoid Question Answering

- Extract relevant information from query
 - What is the likely answer type
 - Collocations from query for finding answer
- Retrieve relevant passages
- Extract possible answers (e.g., entities) from passages
- Classify possible answers
- Rank possible answers
- Perhaps techniques for combining multiple answers into one

Computational Linguistics 1

44

More-Complex Questions

- Questions don't typically conform to simple factoid structure
 - <160>What is the role of PrnP in mad cow disease?
 - <174>How does BRCA1 ubiquitinating activity contribute to cancer?
 - <175>How does L2 interact with L1 to form HPV11 viral capsids?
- Factoid Q&A won't suffice
 - Answers won't be entities or any small phrase
 - Answer "type" will likely be sentences or paragraphs
- Standard ad hoc IR document retrieval also insufficient
 - Want information easier to peruse

Query-Driven Summarization

- Can reach same task through "query-driven" summarization
 - Given a query and a document set, produce a summary
- Also reach a very similar task in IR, when "passages" rather than documents are returned
- Outside of factoid question answering, QA can be placed on the IR/summarization continuum

Agenda: Summary

- HW6 due today!
- Questions, comments, concerns?
- Information Retrieval (IR)
- Question Answering (QA)

Next time:

- Summarization
- Information Extraction (IE)