Improving Multimodal Image Retrieval Using Relevance Feedback and Query Expansion

Jayashree Kalpathy-Cramer,
Steven Bedrick, Bill Hersh

Dept. Medical Informatics & Clinical Epidemiology,
Oregon Health and Science University, Portland, OR

OREGON HEALTH & SCIENCE UNIVERSITY

Agenda

- Historical Perspective on Relevance Feedback
  - Information retrieval
    - Query Expansion
    - Moving query towards “ideal query”
  - Image retrieval
    - Classification/machine learning
    - Statistical model of information need
- Challenges in Evaluation
- Demo
  - Query expansion using thesaurus
  - Query expansion using pseudo relevance feedback
  - RSVP + relevance feedback with Bayesian classifier (text)
  - RSVP + relevance feedback with images using LSA (?)

OREGON HEALTH & SCIENCE UNIVERSITY

Relevance feedback

- (One of) First Interactive techniques in IR
  - originally developed in the information retrieval domain for text documents.
  - allows the user to provide input on the relevance of the initial document set
    - understand users need
  - used by the system to retrieve documents that are “similar” to relevant documents
    - Formulate “ideal” query
    - Recognize when see it

OREGON HEALTH & SCIENCE UNIVERSITY

Use of Relevance Feedback in Text retrieval

- Explicit feedback
  - User marks relevance
  - Traditional implementation
    - modified, expanded query
- Pseudo feedback
  - Assume top x relevant, use terms from these to build better query
- Implicit feedback
  - Based on user behavior

Explicit Feedback - How does it work?

- User starts with a (short, simple) query
- System returns ordered list of documents
- The user marks returned documents as relevant or non-relevant.
- The system computes a better representation of the information need (query) based on feedback.
- Can go through one or more iterations.

OREGON HEALTH & SCIENCE UNIVERSITY

Explicit Feedback

- Feedback explicitly provided by users
  - Document level or term level
- How best to use feedback?
  - Are terms selected manually or automatically?
  - Lots of parameters
  - Which terms
  - How to weight terms
  - Document level - how to use
  - How to rerank?
  - Do you show already shown documents?
  - Depends on corpus
- Generally accepted that it works
  - Difficult to evaluate

Pseudo Relevance Feedback

- Automatic local analysis - depends on specific query
- Pseudo relevance feedback attempts to automate the manual part of relevance feedback.
  - Retrieve an initial set of relevant documents.
  - *Assume* that top *m* ranked documents are relevant.
  - Do relevance feedback
- Similar questions
  - How many documents/terms for pseudo relevance
  - How many iterations?
- Danger of query drift

OREGON
HEALTH
&SCIENCE
UNIVERSITY

Implicit RF

- User behavior as feedback
  - No explicit marking of relevance
- Example:
  - Select : click=> relevance
  - View: longer time => more relevant
  - Save, Bookmark
  - Link, Cite

Does Relevance Feedback Work?

- System centered research- relevance feedback works
  - – More terms, longer query
  - – Need adequate judgments
  - – Need medium-large set of relevant documents
  - – Empirically, one round of relevance feedback is often very useful
- User centered research - mixed

Feedback issues

- Users lazy/reluctant to provide feedback
- Context of new terms important
- Poor terms can cause users to lose trust in sytem's recommendation
- Cognitively demanding- requires more from the user
- Control - show user what is going on vs. doing it magically
  - Makes it hard to understand why a particular document was retrieved
- Users aren't able to pick best term
  - People and systems don't agree on what are good terms
- Negative Feedback is difficult, not predictable
- Query Drift

Query Expansion

- In relevance feedback, users give additional input (relevant/non-relevant) on documents, which is used to re-weight terms in the documents

- In query expansion, users give additional input (good/bad search term) on words or phrases.

Types of Query Expansion

- Global Analysis: (static; of all documents in collection)
  - Controlled vocabulary
    - Maintained by editors (e.g., medline)
  - Manual thesaurus
    - E.g. MedLine: physician, syn: doc, doctor, MD, medico
  - Automatically derived thesaurus
    - (co-occurrence statistics)
  - Refinements based on query log mining
    - Common on the web
- Local Analysis: (dynamic)
  - Analysis of documents in result set

Thesaurus-based Query Expansion

- Does not require user input
- Each term is expanded with synonyms from thesaurus
  - May weight added terms less than original query terms.
- Can increase recall.
- May significantly decrease precision, particularly with ambiguous terms
  - Terms can map to many different synonyms depending on context
- There is a high cost of manually producing a thesaurus
  - NLM Metathesaurus, MeSH, Snomed etc

Relevance Feedback in Image Retrieval

- Way to bridge the semantic gap
  - attempt to discern information need
  - a human and a system interact to refine queries
  - image features and respective weights can be non-intuitive for humans
    - the weights are dynamically updated

OREGON HEALTH & SCIENCE UNIVERSITY

Relevance Feedback in Image Retrieval

- Feature Selection
  - Identify most discrimant features
- Re-weight features
- Classification
- Statistical model of user's needs
- Distance learning
  - DistBoost
- Affinity Propagation
- Online learning
  - Update after each selection

Evaluation

- Hard to evaluate relevance feedback
  - How to compare algorithms
  - How to isolate impact of adding new terms
  - Don’t want to give people same document again
    - Measure of diversity
  - Information need can change as person sees more documents
- How to deal with judged documents?
  - Evaluate only unrated documents to avoid testing on training data
    - What if only few relevant documents in collection
- Effectiveness depends on initial query, especially for pseudo relevance
  - Query drift

OREGON HEALTH & SCIENCE UNIVERSITY

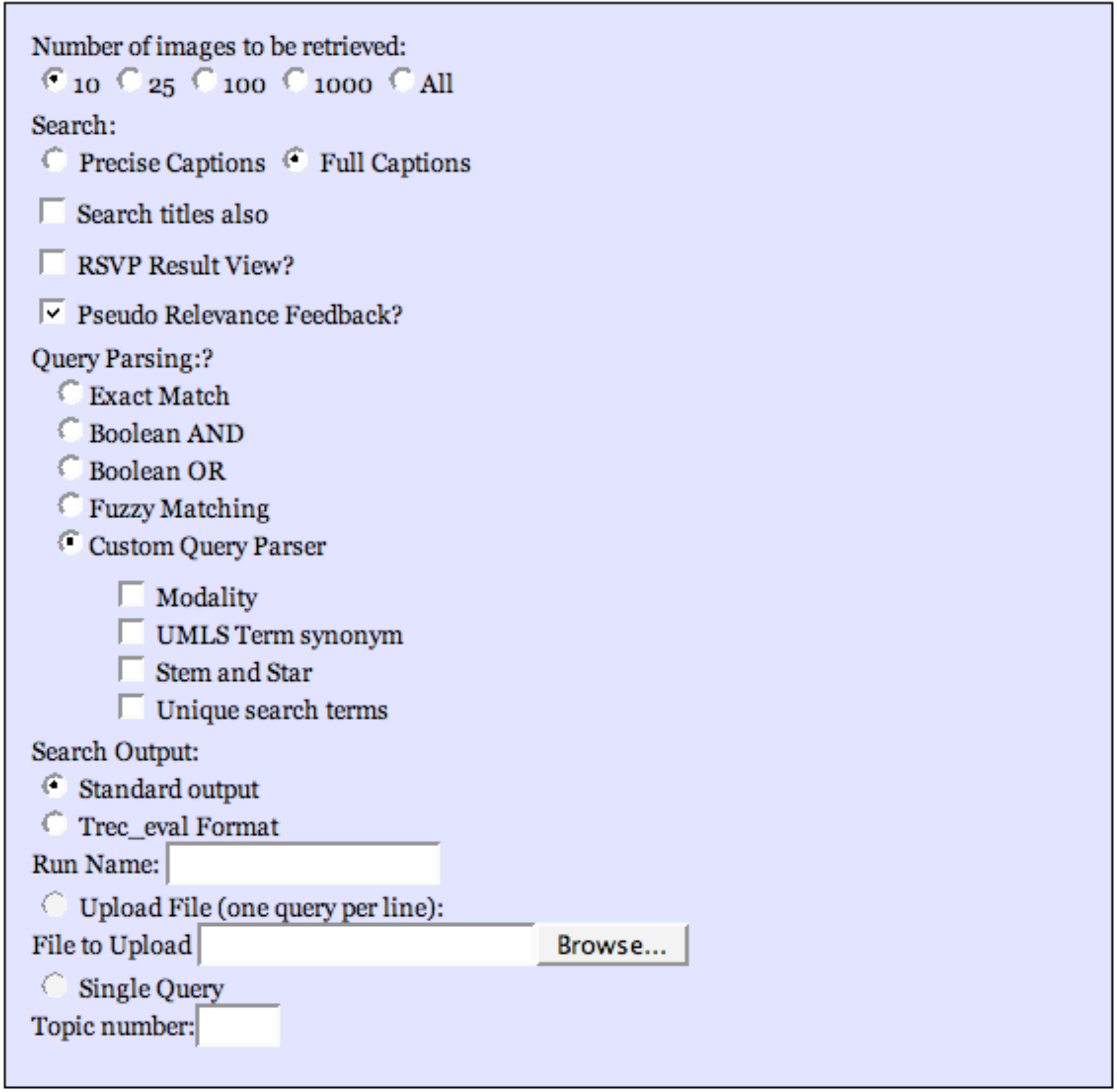
Relevance Feedback in ImageCLEF

- Barriers
  - Not all participants have resources for interactive searches
    - No system
    - No personnel
    - No time
  - How to compare fairly?
    - Gather information about judged
    - Separate initial query from feedback system

OREGON HEALTH & SCIENCE UNIVERSITY

OHSU's interactive runs at ImageCLEF med

- Only (best :-) group to submit interactive run
- Options selected by user
  - Query parsing
    - Modality Detection -filter out images of not desired modality
  - Pseudo relevance feedback
  - Relevance feedback
  - UMLS (thesaurus) based expansion



Number of images to be retrieved:
● 10 ○ 25 ○ 100 ○ 1000 ○ All

Search:
○ Precise Captions ● Full Captions

□ Search titles also

□ RSVP Result View?

☑ Pseudo Relevance Feedback?

Query Parsing:?
- ○ Exact Match
- ○ Boolean AND
- ○ Boolean OR
- ○ Fuzzy Matching
- ● Custom Query Parser
  - □ Modality
  - □ UMLS Term synonym
  - □ Stem and Star
  - □ Unique search terms

Search Output:
○ Standard output ●

Search Output:
● Standard output
○ Trec_eval Format

Run Name:

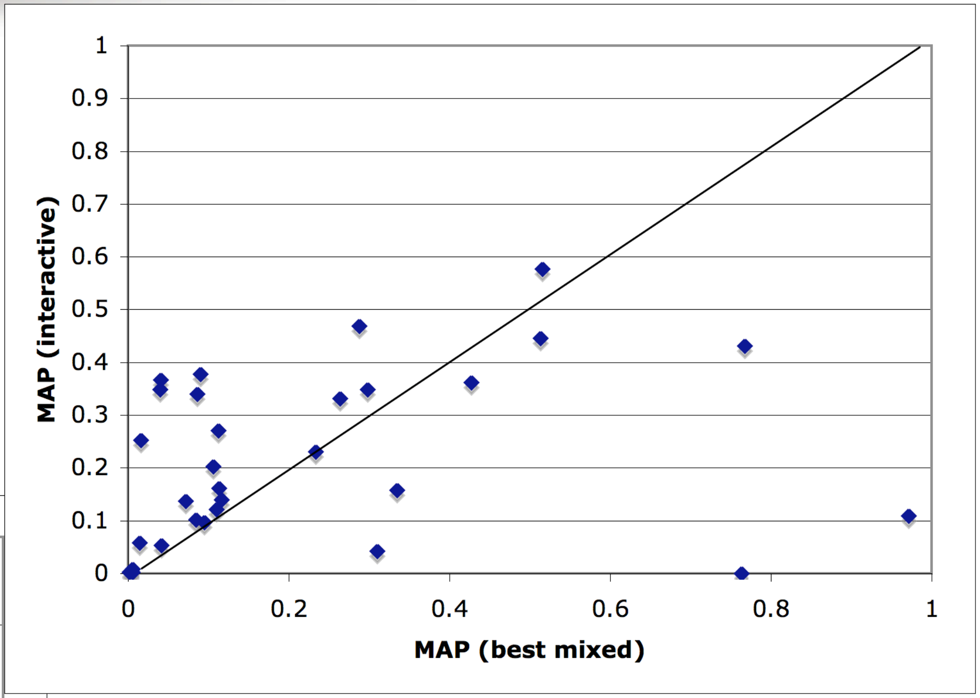
○ Upload File (one query per line):

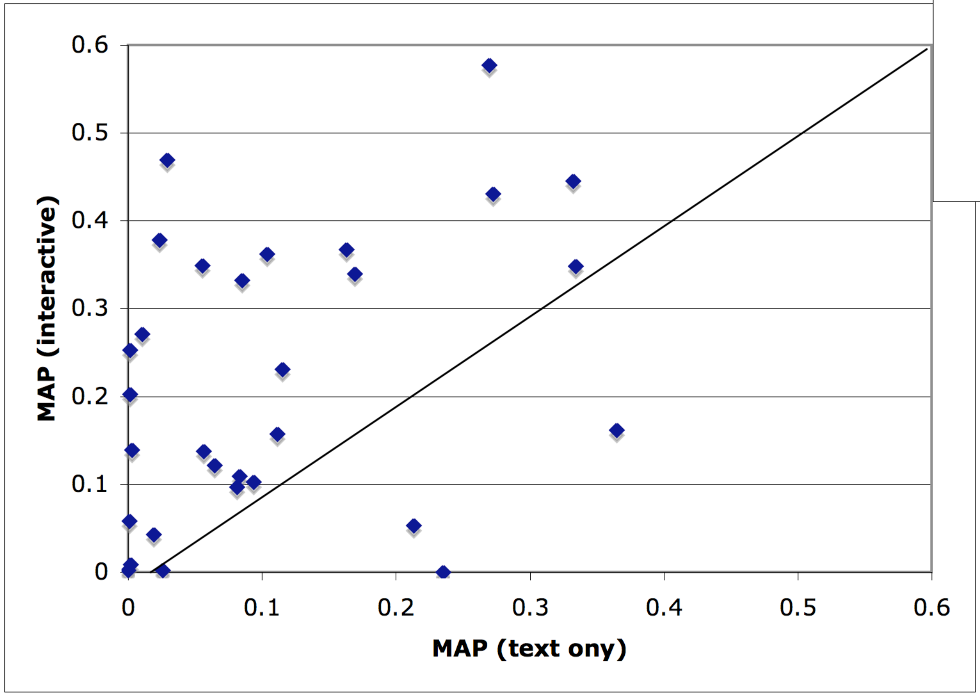
File to Upload Browse...

○ Single Query

Topic number:

MAP interactive vs. baseline, best automatic





OREGON HEALTH & SCIENCE UNIVERSITY

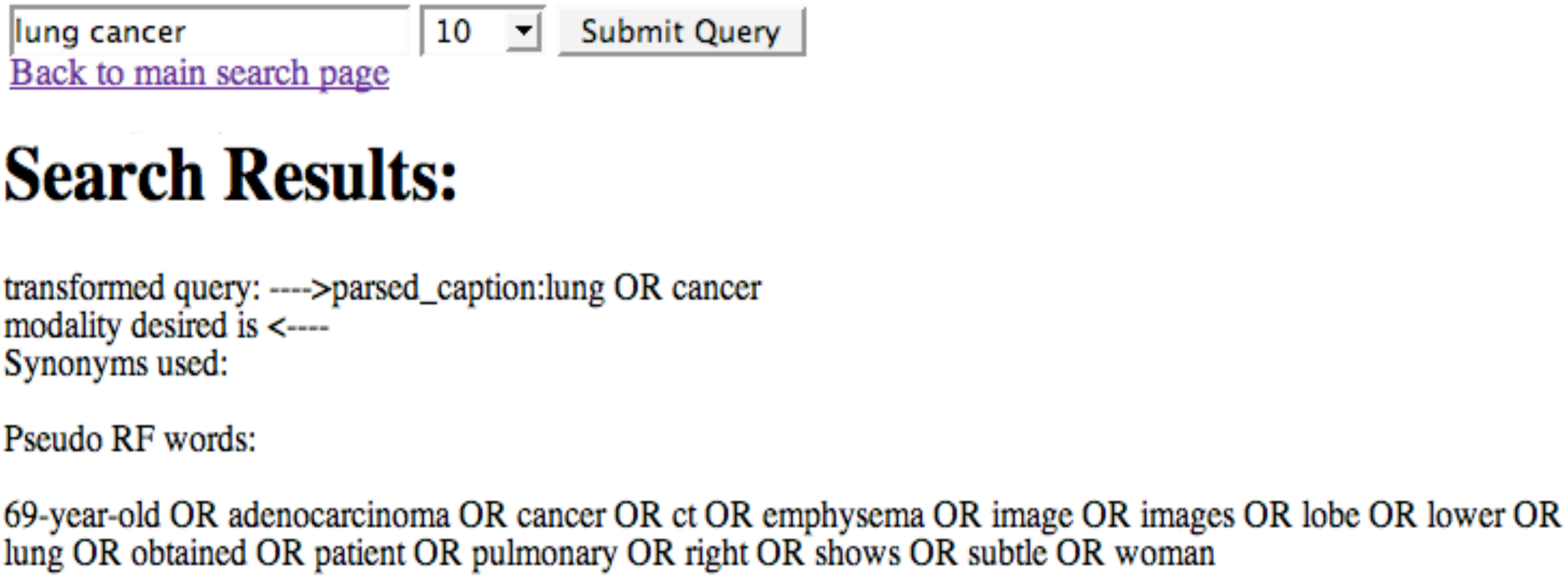
Our experience with interactive searches

- Time consuming
- Easier to make mistakes while submitting run
- How many images should we judge?
  - Is it fair
- Difference of opinion between judges
- Can improve the search experience

Demonstration

- Query Expansion using pseudo relevance feedback
- Query Expansion using thesaurus
- RSVP + feedback

Query Expansion using Pseudo RF



lung cancer

10

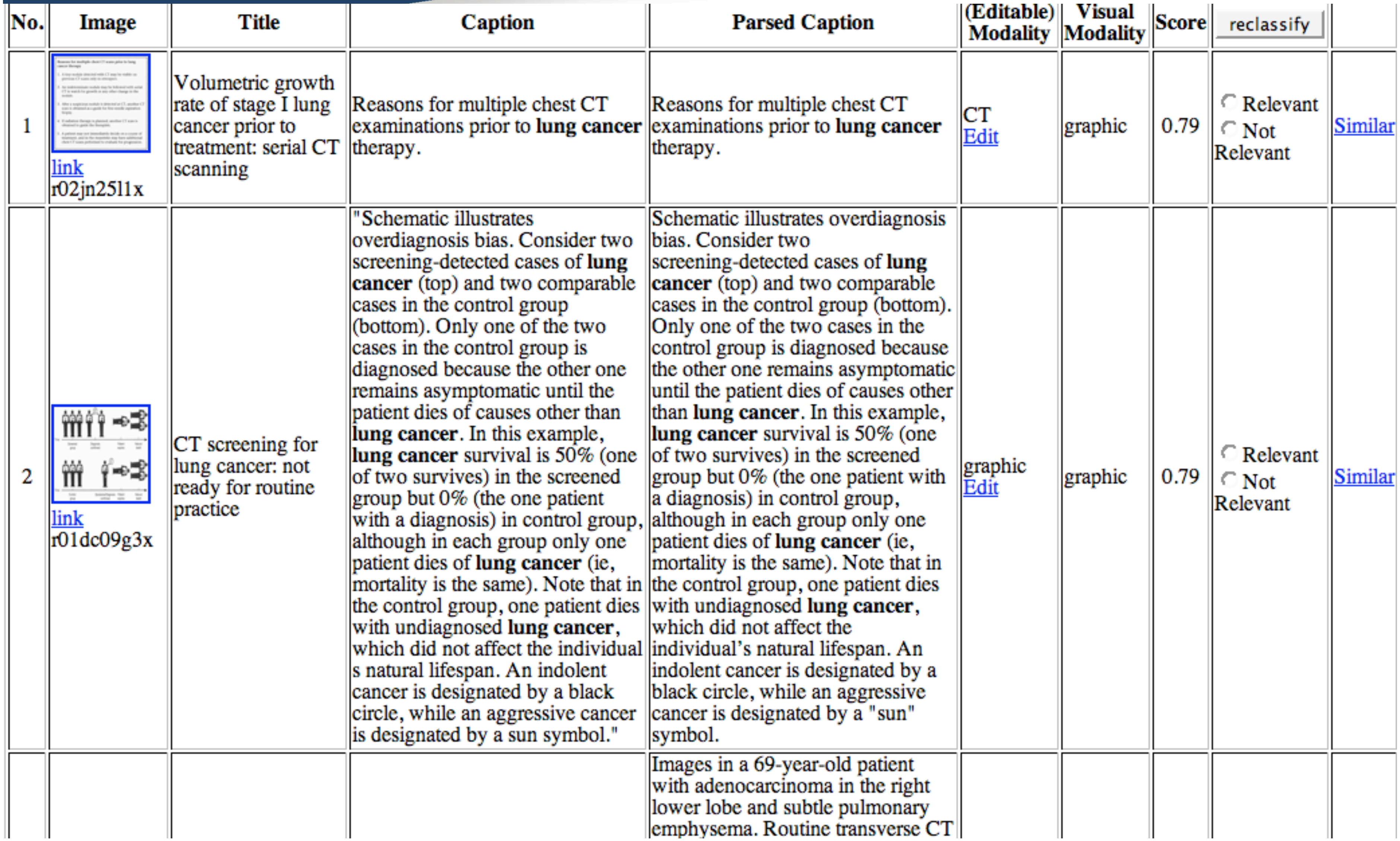
Search Results:

transformed query: ---->parsed_caption:lung OR cancer
modality desired is <----
Synonyms used:

Pseudo RF words:

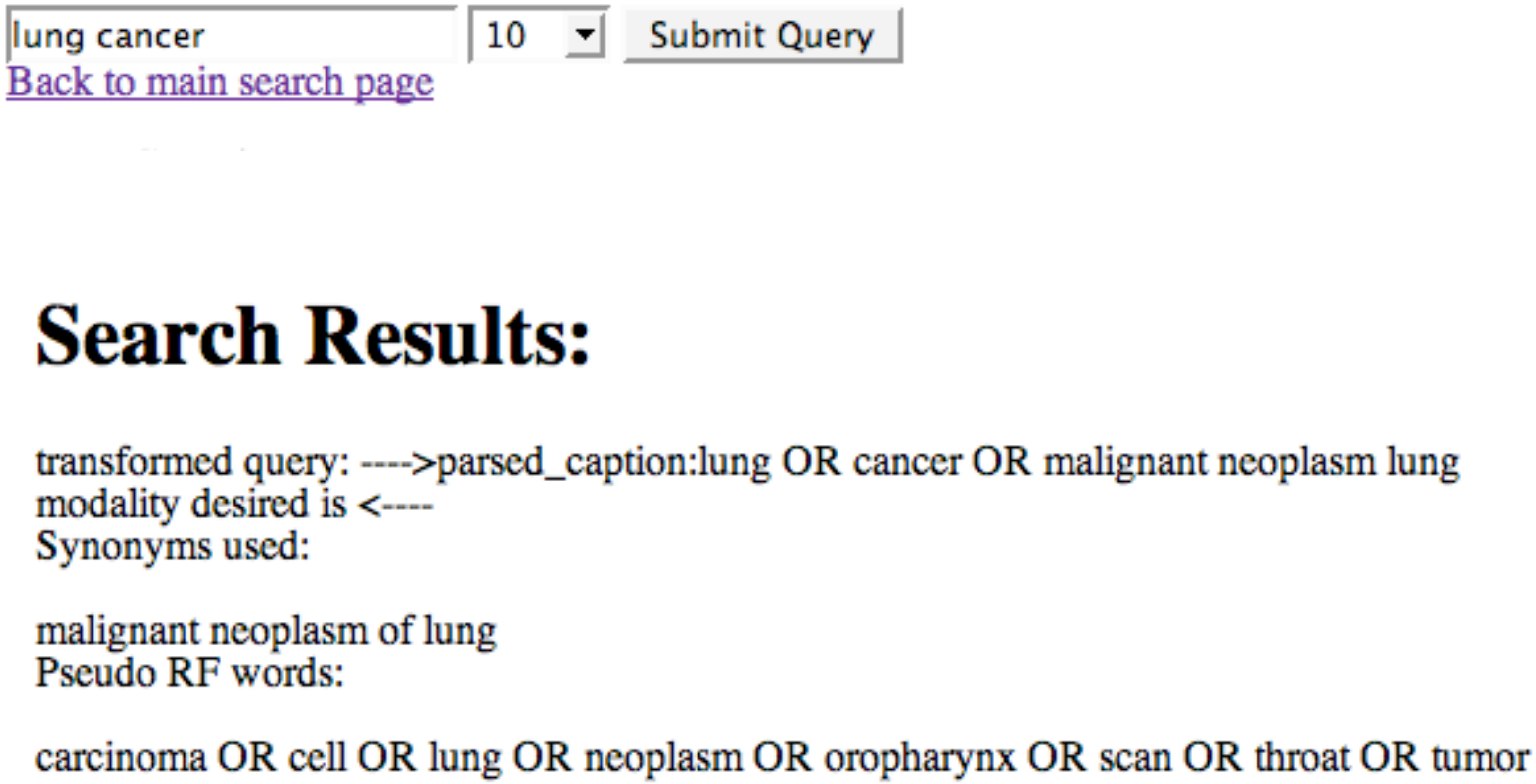
69-year-old OR adenocarcinoma OR cancer OR ct OR emphysema OR image OR images OR lobe OR lower OR lung OR obtained OR patient OR pulmonary OR right OR shows OR subtle OR woman

Pseudo RF + thesaurus



No.	Image	Title	Caption	Parsed Caption	(Editable) Modality	Visual Modality	Score	reclassify	
1	link r02jn25l1x	Volumetric growth rate of stage I lung cancer prior to treatment: serial CT scanning	Reasons for multiple chest CT examinations prior to **lung cancer** therapy.	Reasons for multiple chest CT examinations prior to **lung cancer** therapy.	CT Edit	graphic	0.79	○ Relevant ○ Not Relevant	Similar
2	link r01dc09g3x	CT screening for lung cancer: not ready for routine practice	"Schematic illustrates overdiagnosis bias. Consider two screening-detected cases of **lung cancer** (top) and two comparable cases in the control group (bottom). Only one of the two cases in the control group is diagnosed because the other one remains asymptomatic until the patient dies of causes other than **lung cancer**. In this example, **lung cancer** survival is 50% (one of two survives) in the screened group but 0% (the one patient with a diagnosis) in control group, although in each group only one patient dies of **lung cancer** (ie, mortality is the same). Note that in the control group, one patient dies with undiagnosed **lung cancer**, which did not affect the individual s natural lifespan. An indolent cancer is designated by a black circle, while an aggressive cancer is designated by a sun symbol."	Schematic illustrates overdiagnosis bias. Consider two screening-detected cases of **lung cancer** (top) and two comparable cases in the control group (bottom). Only one of the two cases in the control group is diagnosed because the other one remains asymptomatic until the patient dies of causes other than **lung cancer**. In this example, **lung cancer** survival is 50% (one of two survives) in the screened group but 0% (the one patient with a diagnosis) in control group, although in each group only one patient dies of **lung cancer** (ie, mortality is the same). Note that in the control group, one patient dies with undiagnosed **lung cancer**, which did not affect the individual's natural lifespan. An indolent cancer is designated by a black circle, while an aggressive cancer is designated by a "sun" symbol.	graphic Edit	graphic	0.79	○ Relevant ○ Not Relevant	Similar
				Images in a 69-year-old patient with adenocarcinoma in the right lower lobe and subtle pulmonary emphysema. Routine transverse CT					

Query Expansion using Pseudo RF + thesaurus



lung cancer

10

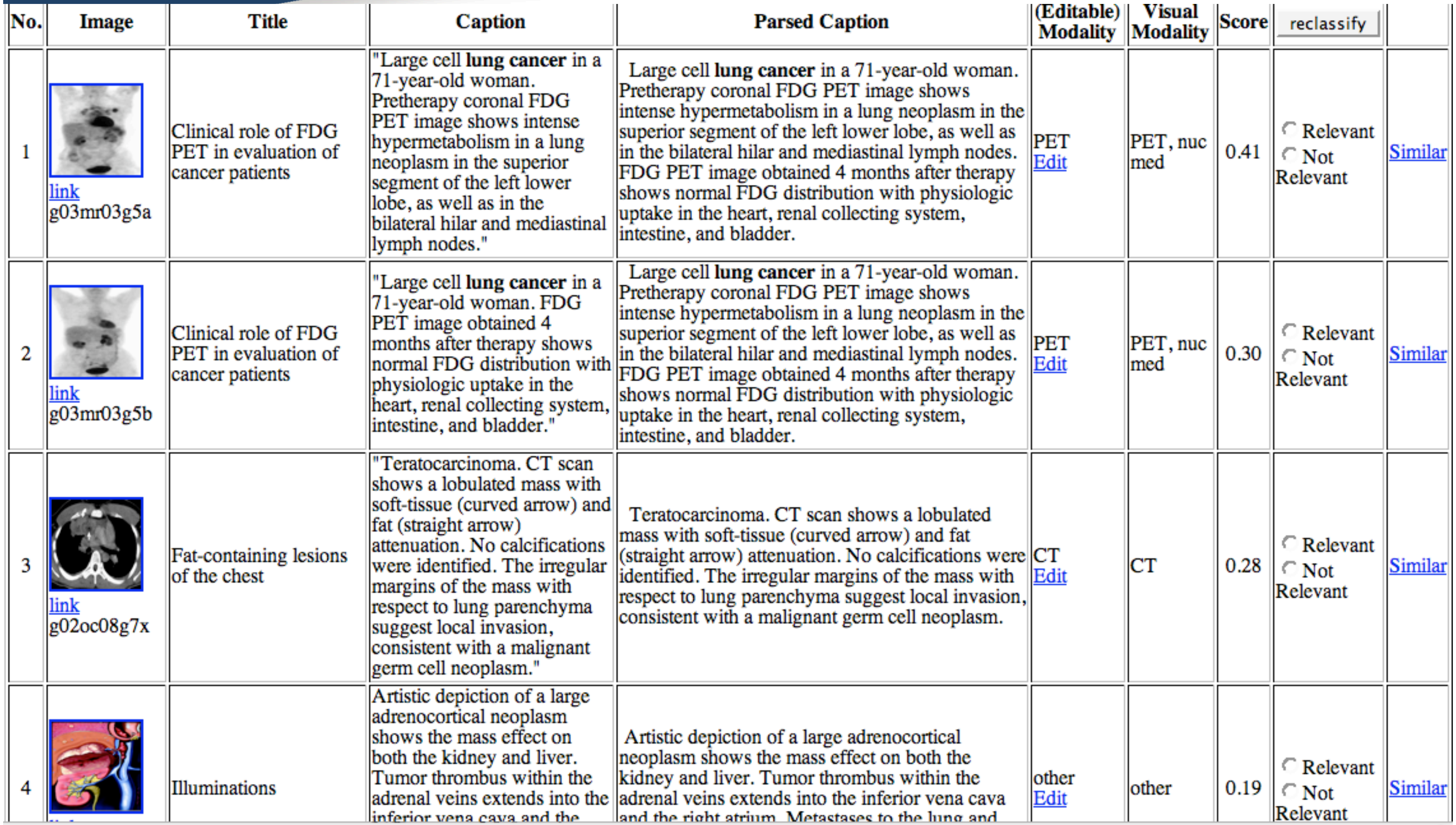
Search Results:

transformed query: ---->parsed_caption:lung OR cancer OR malignant neoplasm lung
modality desired is <----
Synonyms used:

malignant neoplasm of lung
Pseudo RF words:

carcinoma OR cell OR lung OR neoplasm OR oropharynx OR scan OR throat OR tumor

Pseudo RF + thesaurus



No.	Image	Title	Caption	Parsed Caption	(Editable) Modality	Visual Modality	Score	reclassify	
1	link g03mr03g5a	Clinical role of FDG PET in evaluation of cancer patients	"Large cell **lung cancer** in a 71-year-old woman. Pretherapy coronal FDG PET image shows intense hypermetabolism in a lung neoplasm in the superior segment of the left lower lobe, as well as in the bilateral hilar and mediastinal lymph nodes."	Large cell **lung cancer** in a 71-year-old woman. Pretherapy coronal FDG PET image shows intense hypermetabolism in a lung neoplasm in the superior segment of the left lower lobe, as well as in the bilateral hilar and mediastinal lymph nodes. FDG PET image obtained 4 months after therapy shows normal FDG distribution with physiologic uptake in the heart, renal collecting system, intestine, and bladder.	PET Edit	PET, nuc med	0.41	○ Relevant ○ Not Relevant	Similar
2	link g03mr03g5b	Clinical role of FDG PET in evaluation of cancer patients	"Large cell **lung cancer** in a 71-year-old woman. FDG PET image obtained 4 months after therapy shows normal FDG distribution with physiologic uptake in the heart, renal collecting system, intestine, and bladder."	Large cell **lung cancer** in a 71-year-old woman. Pretherapy coronal FDG PET image shows intense hypermetabolism in a lung neoplasm in the superior segment of the left lower lobe, as well as in the bilateral hilar and mediastinal lymph nodes. FDG PET image obtained 4 months after therapy shows normal FDG distribution with physiologic uptake in the heart, renal collecting system, intestine, and bladder.	PET Edit	PET, nuc med	0.30	○ Relevant ○ Not Relevant	Similar
3	link g02oc08g7x	Fat-containing lesions of the chest	"Teratocarcinoma. CT scan shows a lobulated mass with soft-tissue (curved arrow) and fat (straight arrow) attenuation. No calcifications were identified. The irregular margins of the mass with respect to lung parenchyma suggest local invasion, consistent with a malignant germ cell neoplasm."	Teratocarcinoma. CT scan shows a lobulated mass with soft-tissue (curved arrow) and fat (straight arrow) attenuation. No calcifications were identified. The irregular margins of the mass with respect to lung parenchyma suggest local invasion, consistent with a malignant germ cell neoplasm.	CT Edit	CT	0.28	○ Relevant ○ Not Relevant	Similar
4		Illuminations	Artistic depiction of a large adrenocortical neoplasm shows the mass effect on both the kidney and liver. Tumor thrombus within the adrenal veins extends into the inferior vena cava and the	Artistic depiction of a large adrenocortical neoplasm shows the mass effect on both the kidney and liver. Tumor thrombus within the adrenal veins extends into the inferior vena cava and the right atrium. Metastases to the lung and	other Edit	other	0.19	○ Relevant ○ Not Relevant	Similar

Future Plans

- User studies (Google grant)
- Eye tracking (?)
- User interface design
  - List / Thumbview /RSVP

Conclusions

- Relevance feedback and query expansion can be quite useful
  - Painful for the user / Time consuming
  - CBIR systems have benefited from it
  - Machine learning techniques can help
    - On-line learning with dynamic updates
- Hard to evaluate efficacy

THANKS

- Questions?

OREGON HEALTH & SCIENCE UNIVERSITY

Sources

- Adapted from Lectures by

  Prabhakar Raghavan (Yahoo, Stanford)

  Christopher Manning (Stanford)

  Hinrich Schuetze (Stuttgart)

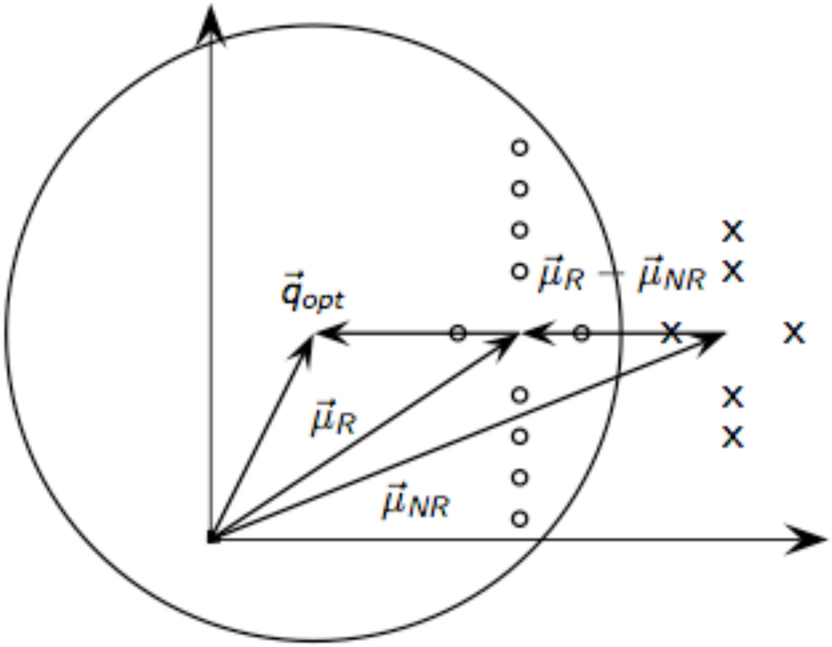
  http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html

Rocchio Algorithm

- The Rocchio algorithm incorporates relevance feedback information into the vector space model.
- Want to maximize *sim* (*Q*, C_r) - *sim* (*Q*, C_{nr})
- The optimal query vector for separating relevant and non-relevant documents (with cosine sim.):

$$\vec{Q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\vec{d}_j \notin C_r} \vec{d}_j$$

- Q_{opt} = optimal query; C_r = set of rel. doc vectors; *N* = collection size



$\vec{q}_{opt}$ separates relevant/nonrelevant perfectly.

Local and Global Methods

- Local methods
  - Relevance feedback
  - Pseudo relevance feedback
- Global methods
  - Query expansion/reformulation
    - Thesauri (or WordNet)
    - Automatic thesaurus generation
  - Global indirect relevance feedback

Use of Relevance Feedback

- Move query point to be closer to relevant objects
- Change weights of features to give more importance to discriminant features

Automatic Thesaurus Generation

- Attempt to generate a thesaurus automatically by analyzing the collection of documents
- Two main approaches
  - Co-occurrence based (co-occurring words are more likely to be similar)
  - Shallow analysis of grammatical relations
- Co-occurrence based is more robust, grammatical relations are more accurate.

One option for evaluation

- Remove from the corpus any documents for which feedback was provided.
- Measure recall/precision performance on the remaining *residual collection*.
- Compared to complete corpus, specific recall/precision numbers may decrease since relevant documents were removed.
- However, **relative** performance on the residual collection provides fair data on the effectiveness of relevance feedback.