

CSCI 5417
Information Retrieval Systems

Jim Martin

Lecture 8
9/15/2011

Today 9/15

- Finish evaluation discussion
- Query improvement
 - Relevance feedback
 - Pseudo-relevance feedback
- Query expansion

Evaluation

- Summary measures
 - Precision at fixed retrieval level
 - Perhaps most appropriate for web search: all people want are good matches on the first one or two results pages
 - But has an arbitrary parameter of k
 - 11-point interpolated average precision
 - The standard measure in the TREC competitions: you take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents, using interpolation (the value for 0 is always interpolated!), and average them
 - Evaluates performance at all recall levels

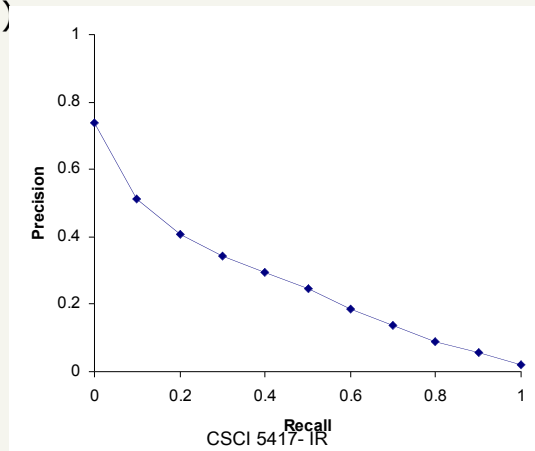
9/19/11

CSCI 5417- IR

3

Typical (good) 11 point precisions

- SabIR/Cornell 8A1 11pt precision from TREC 8 (1999)



9/19/11

CSCI 5417- IR

4

Yet more evaluation measures...

- Mean average precision (MAP)
 - Average of the precision value obtained for the top k documents, each time a relevant doc is retrieved
 - Avoids interpolation, use of fixed recall levels
 - MAP for query collection is arithmetic avg.
 - Macro-averaging: each query counts equally

9/19/11

CSCI 5417- IR

5

Recall/Precision

		■ R	P	MAP
■ 1	R	■ 10%	100%	100
■ 2	N	■ 10	50	
■ 3	N	■ 10	33	
■ 4	R	■ 20	50	50
■ 5	R	■ 30	60	60
■ 6	N	■ 30	50	
■ 7	R	■ 40	57	57
■ 8	N	■ 40	50	
■ 9	N	■ 40	44	
■ 10	N	■ 40	40	
		■		.6675

9/19/11

CSCI 5417

6

Variance

- For a test collection, it is usual that a system does poorly on some information needs (e.g., MAP = 0.1) and excellently on others (e.g., MAP = 0.7)
- Indeed, it is usually the case that the variance in performance of the same system across queries is much greater than the variance of different systems on the same query.
- That is, there are easy information needs and hard ones!

9/19/11

CSCI 5417

7

Finally

- All of these measures are used for distinct *comparison* purposes
 - System A vs System B
 - System A (1.1) vs System A (1.2)
 - Approach A vs. Approach B
 - Vector space approach vs. Probabilistic approaches
 - Systems on different collections?
 - System A on med vs. trec vs web text?
- They don't represent absolute measures

9/19/11

CSCI 5417

8

From corpora to test collections

- Still need
 - Test queries
 - Relevance assessments
- Test queries
 - Must be germane to docs available
 - Best designed by domain experts
 - Random query terms generally not a good idea
- Relevance assessments
 - Human judges, time-consuming
 - Human panels are not perfect

9/19/11

CSCI 5417

9

Pooling

- With large datasets it's impossible to really assess recall.
 - You would have to look at every document.
- So TREC uses a technique called pooling.
 - Run a query on a representative set of state of the art retrieval systems.
 - Take the union of the top N results from these systems.
 - Have the analysts judge the relevant docs in this set.

9/19/11

CSCI 5417

10

TREC

- TREC Ad Hoc task from first 8 TRECs is standard IR task
 - 50 detailed information needs a year
 - Human evaluation of pooled results returned
 - More recently other related things: Web track, HARD, Bio, Q/A
- A TREC query (TREC 5)
 - <top>
 - <num> Number: 225
 - <desc> Description:
What is the main function of the Federal Emergency Management Agency (FEMA) and the funding level provided to meet emergencies? Also, what resources are available to FEMA such as people, equipment, facilities?
 - </top>

9/19/11

CSCI 5417

11

Critique of Pure Relevance

- Relevance vs [Marginal Relevance](#)
 - A document can be redundant even if it is highly relevant
 - Duplicates
 - The same information from different sources
 - Marginal relevance is a better measure of utility for the user.
- Using facts/entities as evaluation units more directly measures true relevance.
- But harder to create evaluation set

9/19/11

CSCI 5417

12

Search Engines...

- How does any of this apply to the big search engines?

9/19/11

CSCI 5417

13

Evaluation at large search engines

- Recall is difficult to measure for the web
- Search engines often use precision at top k , e.g., $k = 10$
- Or measures that reward you more for getting rank 1 right than for getting rank 10 right.
 - NDCG (Normalized Cumulative Discounted Gain)
- Search engines also use non-relevance-based measures
 - Clickthrough on first result
 - Not very reliable if you look at a single clickthrough ... but pretty reliable in the aggregate.
 - Studies of user behavior in the lab
 - A/B testing
 - Focus groups
 - Diary studies

9/19/11

CSCI 5417

14

A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a system up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an "automatic" measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most

9/19/11

CSCI 5417

15

Query to think about

- E.g., Information need: *I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
- Query: **wine red white heart attack effective**

9/19/11

CSCI 5417- IR

16

Sources of Errors (unranked)

	Relevant	Not Relevant
Retrieved	a	b
Not Retrieved	c	d

- What's happening in boxes c and b?

9/19/11

CSCI 5417- IR

17

Retrieved/Not Relevant (b)

- Documents are retrieved but are found to be not relevant...
 - Term overlap between query and doc but not relevant overlap...
 - About other topics entirely
 - Terms in isolation are on target
 - Terms are homonymous (off target)
 - About the topic but peripheral to information need

9/19/11

CSCI 5417- IR

18

Not Retrieved/Relevant (c)

- No overlap in terms between the query and docs (zero hits)
 - Documents and users using different vocabulary
 - Synonymy
 - Automobile vs. car
 - HIV vs. AIDS
- Overlap but not enough
 - Problem with weighting schemes?
 - Tf-IDF
 - Problem with similarity metric?
 - Cosine?

9/19/11

CSCI 5417- IR

19

Ranked Results

- Contingency tables are somewhat limited as tools because they're cast in terms of retrieved/not retrieved.
- That's rarely the case in ranked retrieval
 - Problems b and c are duals of the same problem
 - Why was this irrelevant document ranked higher than this relevant document.
 - Why was this irrelevant doc ranked so high?
 - Why was this relevant doc ranked so low?

9/19/11

CSCI 5417- IR

20

Discussion Examples

■ Query

<top>

<num> Number: OHSU42

<title> 43 y o pt with delirium, hypertension,
tachycardia

<desc> Description: thyrotoxicosis, diagnosis and
management

</top>

9/19/11

CSCI 5417- IR

21

Examples: Doc 1

.W A 57-year-old woman presented with palpitations, muscle weakness, bilateral proptosis, goiter, and tremor. The thyroxine (T4) level and the free T4 index were increased while the total triiodothyronine (T3) level was normal. Iodine 123 uptake was increased, and a scan revealed an enlarged gland with homogeneous uptake. Repeated studies again revealed an increased T4 level and free T4 index and normal total and free T3 levels. A protirelin test showed a blunted thyrotropin response. Treatment with propylthiouracil was associated with disappearance of symptoms and normal T4 levels, but after 20 months of therapy, hyperthyroidism recurred and the patient was treated with iodine 131. This was an unusual case of T4 toxicosis because the patient was not elderly and was not exposed to iodine-containing compounds or drugs that impair T4-to-T3 conversion. There was no evidence of abnormal thyroid hormone transport or antibodies.

9/19/11

CSCI 5417- IR

22

Examples: Doc 2

.W A 25-year-old man presented with diffuse metastatic pure choriocarcinoma, **thyrotoxicosis**, and cardiac tamponade. No discernable testicular primary tumor was found. The patient's peripheral blood karyotype was 47, XXY and phenotypic features of Klinefelter's syndrome were present. The patient was treated with aggressive combination chemotherapy followed by salvage surgery and remains in complete remission 3 years after **diagnosis**. Pure choriocarcinoma, although rare as a primary testicular neoplasm, accounts for 15% of extragonadal germ cell tumors in general and 30% of germ cell tumors in patients with Klinefelter's syndrome. Historically, the diagnosis of pure choriocarcinoma has been thought to convey a very poor prognosis. The occurrence of hyperthyroidism is unique to tumors containing choriocarcinomatous elements and the **management** of this disorder is discussed. Treatment of extragonadal germ cell tumors is also discussed with special reference to the roles of combination chemotherapy and salvage surgery.

9/19/11

CSCI 5417- IR

23

So...

- We've got 2 errors here.
 - Doc 1 relevant but not returned
 - What could we do to make it relevant?
 - Doc 2 returned (because of term overlap) but not relevant
 - Why isn't it relevant if it contains the terms?

9/19/11

CSCI 5417- IR

24

Examples: Doc 1

- .T A case of thyroxine **thyrotoxicosis**.
- .W A 57-year-old woman presented with palpitations, muscle weakness, bilateral proptosis, goiter, and tremor. The thyroxine (T4) level and the free T4 index were increased while the total triiodothyronine (T3) level was normal. Iodine 123 uptake was increased, and a scan revealed an enlarged gland with homogeneous uptake. Repeated studies again revealed an increased T4 level and free T4 index and normal total and free T3 levels. A protirelin test showed a blunted thyrotropin response. Treatment with propylthiouracil was associated with disappearance of symptoms and normal T4 levels, but after 20 months of therapy, hyperthyroidism recurred and the patient was treated with iodine 131. This was an unusual case of T4 toxicosis because the patient was not elderly and was not exposed to iodine-containing compounds or drugs that impair T4-to-T3 conversion. There was no evidence of abnormal thyroid hormone transport or antibodies.

9/19/11

CSCI 5417- IR

25

Break

- Quiz is Tuesday 27th
 - Here in class
 - Closed book
 - 1 page cheat sheet ok

9/19/11

CSCI 5417- IR

26

Questions?

- Office hours (ECOT 726)
 - Mondays 10-11:30
 - Thursday 2-3:30
 - And when my door is open

9/19/11

CSCI 5417- IR

27

Readings

- Chapter 1
- Chapter 2: Skip 2.3, 2.4.3
- Chapter 3: skip 3.4
- Chapter 4
- Chapter 6: skip 6.1, 6.4.4
- Chapter 7
- Chapter 8
- Chapter 9:
- Chapter 12: skip 12.4

9/19/11

CSCI 5417- IR

28

Improving Things

- Relevance feedback
- Pseudo-relevance feedback
- Query expansion

- All are focused on creating better queries
- Other directions
 - Weighting scheme (alter the vector space)
 - Similarity scheme (something other than cosine).

9/19/11

CSCI 5417- IR

29

Relevance Feedback

- Relevance feedback: Gather user feedback on relevance of docs in initial set of results
 - User issues a (short, simple) query
 - The **user** marks returned documents as relevant or non-relevant.
 - The **system** computes a better representation of the information need based on feedback.
 - Relevance feedback can go through one or more **iterations**.
- Idea
 - it may be difficult to formulate a good query when you don't know the collection well,
 - But users can tell what they like when they see it

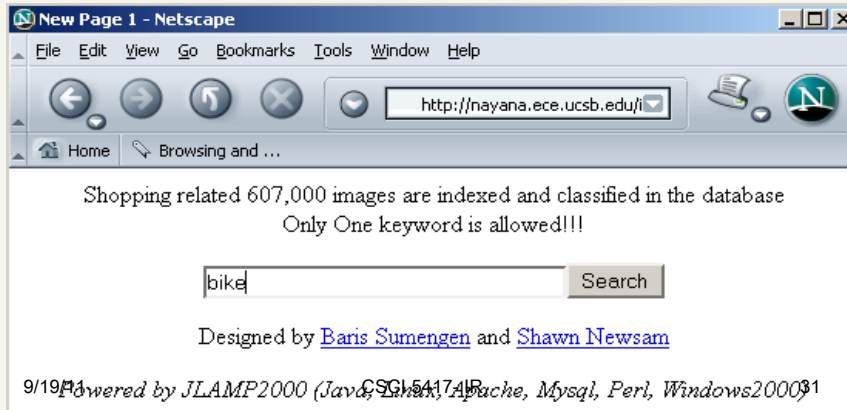
9/19/11

CSCI 5417- IR

30








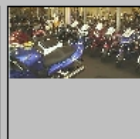




Relevance Feedback: Example

- Image search engine <http://nayana.ece.ucsb.edu/imsearch/imsearch.html>



Results for Initial Query













Browse Search Prev Next Random

					
(144473, 16458) 0.0 0.0 0.0	(144457, 252140) 0.0 0.0 0.0	(144456, 262857) 0.0 0.0 0.0	(144456, 262863) 0.0 0.0 0.0	(144457, 252134) 0.0 0.0 0.0	(144483, 265154) 0.0 0.0 0.0
					
(144483, 264644) 0.0 0.0 0.0	(144483, 265153) 0.0 0.0 0.0	(144518, 257752) 0.0 0.0 0.0	(144538, 525937) 0.0 0.0 0.0	(144456, 249611) 0.0 0.0 0.0	(144456, 250064) 0.0 0.0 0.0

9/19/11 CSCI 5417- IR 32









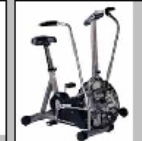
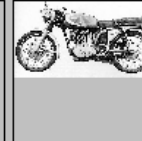


Relevance Feedback

Browse Search Prev Next Random

					
(144473, 16458) 0.0 0.0 0.0	(144457, 252140) 0.0 0.0 0.0	(144456, 262857) 0.0 0.0 0.0	(144456, 262863) 0.0 0.0 0.0	(144457, 252134) 0.0 0.0 0.0	(144483, 265154) 0.0 0.0 0.0
					
(144483, 264644) 0.0 0.0 0.0	(144483, 265153) 0.0 0.0 0.0	(144518, 257752) 0.0 0.0 0.0	(144538, 525937) 0.0 0.0 0.0	(144456, 249611) 0.0 0.0 0.0	(144456, 250064) 0.0 0.0 0.0
9/19/11		CSCI 5417- IR			33

Results after Relevance Feedback

Browse Search Prev Next Random

					
(144538, 523493) 0.54182 0.231944 0.309876	(144538, 523835) 0.56319296 0.267304 0.295889	(144538, 523529) 0.584279 0.280881 0.303398	(144456, 253569) 0.64501 0.351395 0.293615	(144456, 253568) 0.650275 0.411745 0.23853	(144538, 523799) 0.66709197 0.358033 0.309059
					
(144473, 16249) 0.6721 0.393928 0.278178	(144456, 249634) 0.675018 0.4639 0.211118	(144456, 253693) 0.676901 0.47645 0.200451	(144473, 16328) 0.700339 0.391337	(144483, 265264) 0.70170796 0.36176 0.339948	(144478, 512410) 0.70297 0.4691134 0.233859
		CSCI 5417- IR			

Theoretical Optimal Query

- 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
- Unrealistic: we don't know relevant documents.

9/19/11

CSCI 5417- IR

35

Relevance Feedback in vector spaces

- We can modify the query based on relevance feedback and apply standard vector space model.
- Use only the docs that were marked.
- Relevance feedback *can improve* recall and precision
- But *it is most useful* for increasing *recall* in situations where recall is important
 - Users can be expected to review results and to take time to iterate

9/19/11

CSCI 5417- IR

36

Rocchio 1971 Algorithm (SMART)

- Used in practice:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

- q_m = modified query vector; q_0 = original query vector; α, β, γ : weights (hand-chosen or set empirically); D_r = set of known relevant doc vectors; D_{nr} = set of known irrelevant doc vectors
- New query moves toward relevant documents and away from irrelevant documents
- Tradeoff α vs. β/γ : If we have a lot of judged documents, we want a higher β/γ .
- Term weight can go negative
 - Negative term weights are ignored (set to 0)

9/19/11

CSCI 5417- IR

37

Positive vs. Negative Feedback

- Positive feedback is more valuable than negative feedback (so, set $\gamma < \beta$; e.g. $\gamma = 0.25$, $\beta = 0.75$).
- Many systems only allow positive feedback ($\gamma=0$).
- Or a single negative document
 - Ide-dec-hi

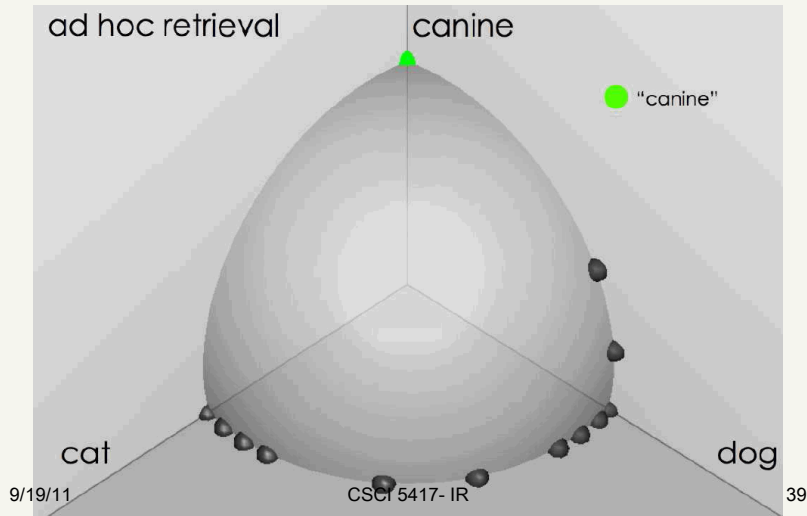
9/19/11

CSCI 5417- IR

38

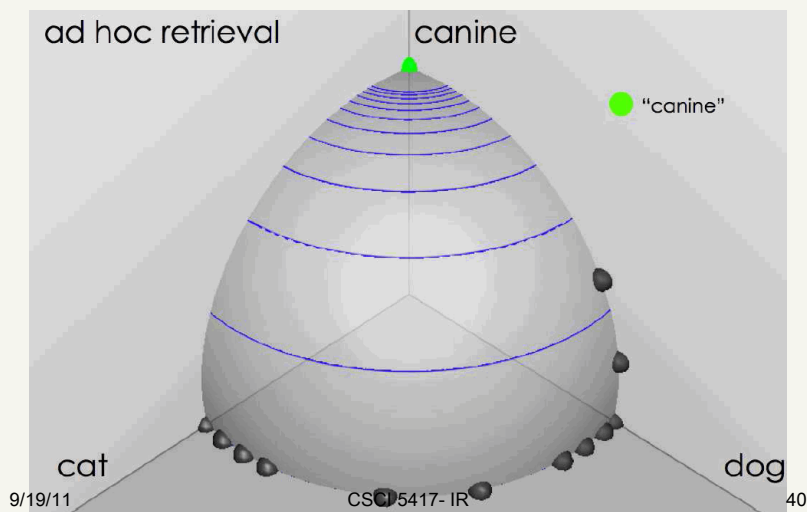
Ad hoc results for query *canine*

source: Fernando Diaz



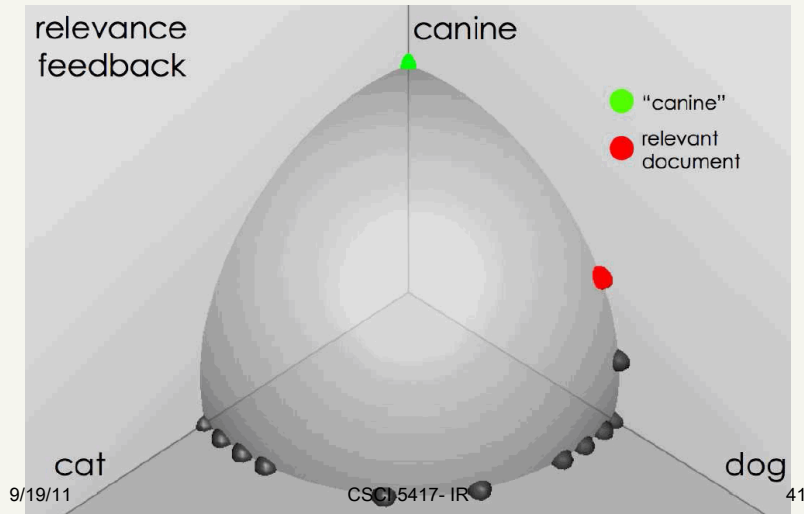
Ad hoc results for query *canine*

source: Fernando Diaz



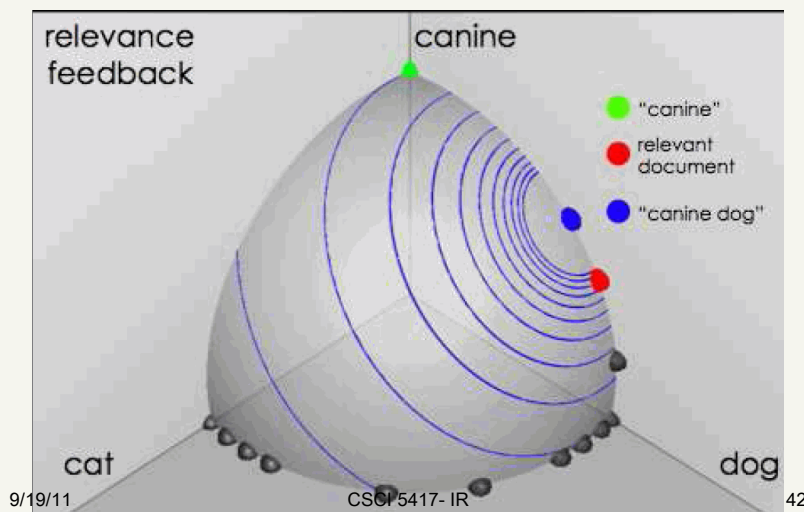
User feedback: Select what is relevant

source: Fernando Diaz



Results after relevance feedback

source: Fernando Diaz



Relevance Feedback: Assumptions

- A1: User has sufficient knowledge for initial query.
- A2: Relevance prototypes are “well-behaved”.
 - Term distribution in relevant documents will be similar
 - Term distribution in non-relevant documents will be different from those in relevant documents
 - Either: All relevant documents are tightly clustered around a single prototype.
 - Or: There are different prototypes, but they have significant vocabulary overlap.

9/19/11

CSCI 5417- IR

43

Violation of Assumptions

- User does not have sufficient initial knowledge to form a reasonable starting query
 - Misspellings (Brittany Speers).
 - Cross-language information retrieval
 - Mismatch of searcher’s vocabulary vs. collection vocabulary
 - Cosmonaut/astronaut

9/19/11

CSCI 5417- IR

44

Relevance Feedback: Practical Problems

- Why do most search engines not use relevance feedback?

9/19/11

CSCI 5417- IR

45

Relevance Feedback: Problems

- Long queries are inefficient for typical IR engines
 - Long response times for user.
 - High cost for retrieval system.
 - Partial solution:
 - Only reweight certain prominent terms
 - Perhaps top 20 by term frequency
- **Users are often reluctant to provide explicit feedback**
- It's often harder to understand why a particular document was retrieved after applying relevance feedback

9/19/11

CSCI 5417- IR

46

Relevance Feedback Summary

- Relevance feedback has been shown to be very effective at improving relevance of results.
 - Requires enough judged documents, otherwise it's unstable (≥ 5 recommended)
 - Requires queries for which the set of relevant documents is medium to large
- Full relevance feedback is painful for the user.
- Full relevance feedback is not very efficient in most IR systems.
- Other types of interactive retrieval may improve relevance by as much with less work.

9/19/11

CSCI 5417- IR

47

Pseudo Relevance Feedback

- 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
- Mostly works
- Found to improve performance in TREC ad-hoc task
 - Danger of query drift

9/19/11

CSCI 5417- IR

48

Query Expansion

- In relevance feedback, users give additional input (relevant/non-relevant) on **documents**, which is used to reweight terms in the documents
- In query expansion, users give additional input (good/bad search term) on **words or phrases**.

9/19/11

CSCI 5417- IR

49

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**

9/19/11

CSCI 5417- IR

50

Controlled Vocabulary

The screenshot shows the PubMed search interface. At the top, there are logos for NCBI, PubMed, and the National Library of Medicine (NLM). Below the logos, there are navigation tabs for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, and Taxonomy. A search bar contains the text "cancer" and a "Go" button. Below the search bar, there are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". On the left side, there is a sidebar with "About Entrez" and "Text Version" links. Below that, there are sections for "Entrez PubMed" (Overview, Help | FAQ, Tutorial, New/Noteworthy, E-Utilities) and "PubMed Services" (Journals Database, MeSH Browser, Single Citation). The main content area displays the "PubMed Query:" as `("neoplasms"[MeSH Terms] OR cancer[Text Word])`. At the bottom of the page, there is a "Search" button, a "URL" field containing "CSCI 5417- IR", and the page number "51".

Thesaurus-based Query Expansion

- This doesn't require user input
- For each term, t , in a query, expand the query with synonyms and related words of t from the thesaurus
 - feline → feline cat
- May weight added terms less than original query terms.
- Generally increases recall.
- Widely used in many science/engineering fields
- May significantly decrease precision, particularly with ambiguous terms.
 - "interest rate" → "interest rate fascinate evaluate"
- There is a high cost of manually producing a thesaurus
 - And for updating it for scientific changes

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
 - Entities that are grown, cooked, eaten, and digested are more likely to be food items.
- Co-occurrence based is more robust, grammatical relations are more accurate.

9/19/11

CSCI 5417- IR

53

Automatic Thesaurus Generation Discussion

- Quality of associations is usually a problem.
 - Term ambiguity may introduce irrelevant statistically correlated terms.
 - "Apple computer" → "Apple red fruit computer"
- Problems:
 - False positives: Words deemed similar that are not
 - False negatives: Words deemed dissimilar that are similar
- Since terms are highly correlated anyway, expansion may not retrieve many additional documents.

9/19/11

CSCI 5417- IR

54

Query Expansion: Summary

- Query expansion is often effective in increasing recall.
 - Fairly successful for subject-specific collections
 - Not always with general thesauri
- In most cases, precision is decreased, often significantly.
- Overall, not as useful as relevance feedback; *may* be as good as pseudo-relevance feedback

9/19/11

CSCI 5417- IR

55

So...

- For HW part 2...
 - Stemming? Stoplists?
 - Better query formulation?
 - Selection?
 - Expansion
 - Automatic?
 - Thesaurus?
 - Better/different weighting scheme
 - Pseudo relevance feedback?
 - Boosting?

9/19/11

CSCI 5417- IR

56