CS276A

Information Retrieval

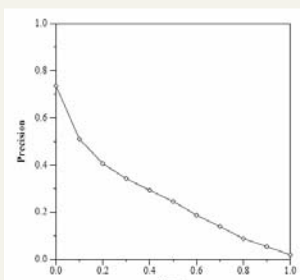
Lecture 9

Recap of the last lecture

- Results summaries
- Evaluating a search engine
  - Benchmarks
  - Precision and recall

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

- Recall Level Ave. Precision
- 0.00 0.7360
- 0.10 0.5107
- 0.20 0.4059
- 0.30 0.3424
- 0.40 0.2931
- 0.50 0.2457
- 0.60 0.1873
- 0.70 0.1391
- 0.80 0.0881
- 0.90 0.0545
- 1.00 0.0197
- Average precision: 0.2553



This lecture

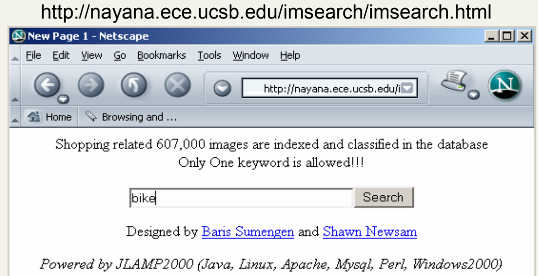
- Improving results
  - For high recall. E.g., searching for *aircraft* didn't match with *plane;* nor *thermodynamic* with *heat*
- Options for improving results…
  - Relevance feedback
  - The complete landscape
    - Global methods
      - Query expansion
        - Thesauri
        - Automatic thesaurus generation
    - Local methods
      - Relevance feedback
      - Pseudo relevance feedback

Relevance Feedback

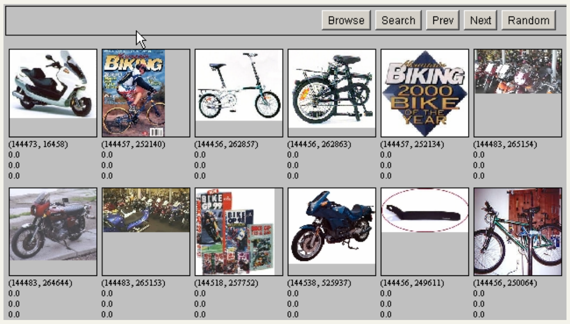
- Relevance feedback: user feedback on relevance of docs in initial set of results
  - User issues a (short, simple) query
  - The user marks returned documents as relevant or non-relevant.
  - The system computes a better representation of the information need based on feedback.
  - Relevance feedback 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, so iterate

Relevance Feedback: Example

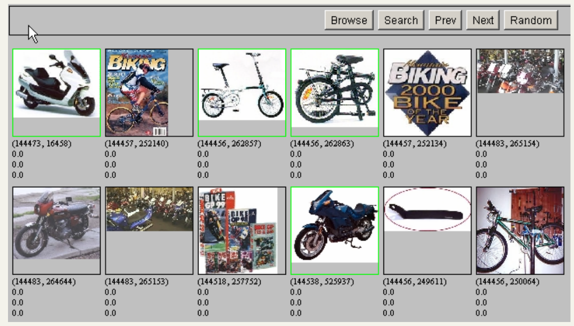
- Image search engine http://nayana.ece.ucsb.edu/imsearch/imsearch.html



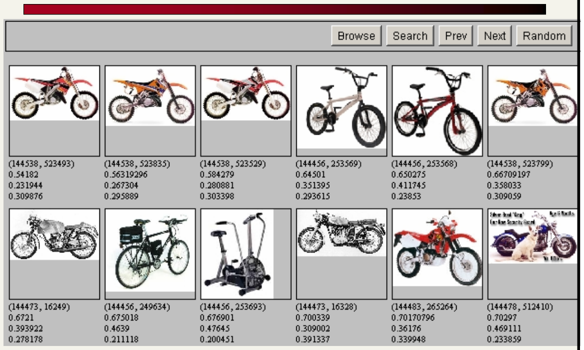
Results for Initial Query



Relevance Feedback



Results after Relevance Feedback



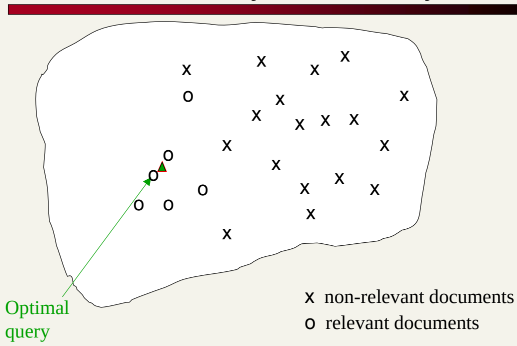
Rocchio Algorithm

- 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:

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

- Q_{opt} = optimal query; C_r = set of rel. doc vectors; N = collection size
- Unrealistic: we don't know relevant documents.

The Theoretically Best Query



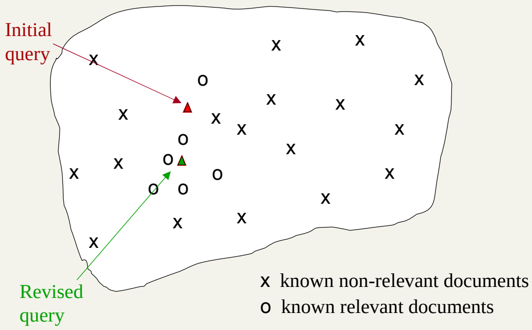
Rocchio 1971 Algorithm (SMART)

- 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.
- Negative term weights are ignored

Relevance feedback on initial query

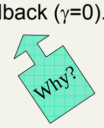


Relevance Feedback in vector spaces

- 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
- Relevance feedback 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

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 (γ=0).



Probabilistic relevance feedback

- Rather than reweighting in a vector space…
- If user has told us some relevant and irrelevant documents, then we can proceed to build a classifier, such as a Naive Bayes model:
  - $P(t_k|R) = |\mathbf{D}_{rk}| / |\mathbf{D}_r|$
  - $P(t_k|NR) = (N_k - |\mathbf{D}_{rk}|) / (N - |\mathbf{D}_r|)$
    - t_k = term in document; $\mathbf{D}_{rk}$ = known relevant doc containing t_k; N_k = total number of docs containing t_k
- More in upcoming lectures
  - This is effectively another way of changing the query term weights
  - Preserves no memory of the original weights

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.
    - Similarities between relevant and irrelevant documents are small

Violation of A1

- User does not have sufficient initial knowledge.
- Examples:
  - Misspellings (Brittany Speers).
  - Cross-language information retrieval (hígado).
  - Mismatch of searcher’s vocabulary vs collection vocabulary
    - Cosmonaut/astronaut

Violation of A2

- There are several relevance prototypes.
- Examples:
  - Burma/Myanmar
  - Contradictory government policies
  - Pop stars that worked at Burger King
- Often: instances of a general concept
- Good editorial content can address problem
  - Report on contradictory government policies

Relevance Feedback: Cost

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

Why?

Relevance Feedback Example: Initial Query and Top 8 Results

- Query: New space satellite applications

Note: want high recall

- + 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
- + 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
- 3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
- 4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
- 5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
- 6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
- 7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
- + 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

Relevance Feedback Example: Expanded Query

2.074 new	15.106 space
30.816 satellite	5.660 application
5.991 nasa	5.196 eos
4.196 launch	3.972 aster
3.516 instrument	3.446 arianespace
3.004 bundespost	2.806 ss
2.790 rocket	2.053 scientist
2.003 broadcast	1.172 earth
0.836 oil	0.646 measure

Top 8 Results After Relevance Feedback

- + 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
- + 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
- 3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
- 4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit
- + 5. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use
- 6. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
- 7. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million
- + 8. 0.488, 12/02/87, Telecommunications Tale of Two Companies

Evaluation of relevance feedback strategies

- Use q_0 and compute precision and recall graph
- Use q_m and compute precision recall graph
  - Use all documents in the collection
    - Spectacular improvements, but … it's cheating!
    - Partly due to known relevant documents ranked higher
    - Must evaluate with respect to documents not seen by user
  - Use documents in residual collection (set of documents minus those assessed relevant)
    - Measures usually lower than for original query
    - More realistic evaluation
    - Relative performance can be validly compared
- Empirically, one round of relevance feedback is often very useful. Two rounds is sometimes marginally useful.

Relevance Feedback on the Web

- Some search engines offer a similar/related pages feature (trivial form of relevance feedback)
  - Google (link-based)
  - Altavista
  - Stanford web
- But some don't because it's hard to explain to average user:
  - Alltheweb
  - msn
  - Yahoo
- Excite initially had true relevance feedback, but abandoned it due to lack of use.

α/β/γ ??

Other Uses of Relevance Feedback

- Following a changing information need
- Maintaining an information filter (e.g., for a news feed)
- Active learning

  [Deciding which examples it is most useful to know the class of to reduce annotation costs]

Relevance Feedback Summary

- Relevance feedback has been shown to be effective at improving relevance of results.
  - Requires enough judged documents, otherwise it's unstable (≥ 5 recommended)
  - For queries in 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.

The complete landscape

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

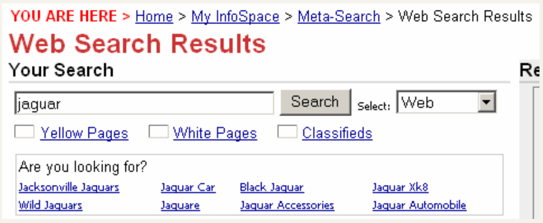
Query Reformulation: Vocabulary Tools

- Feedback
  - Information about stop lists, stemming, etc.
  - Numbers of hits on each term or phrase
- Suggestions
  - Thesaurus
  - Controlled vocabulary
  - Browse lists of terms in the inverted index

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.

Query Expansion: Example

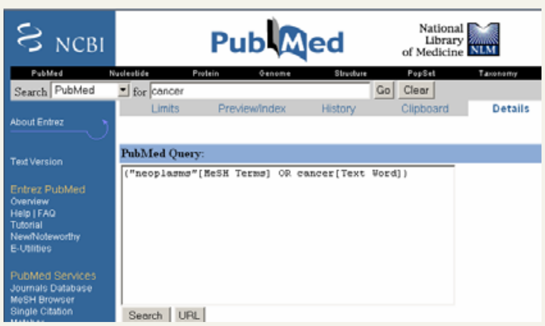


Also: see altavista, teoma

Types of Query Expansion

- Global Analysis: Thesaurus-based
  - 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:
  - Analysis of documents in result set

Controlled Vocabulary



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

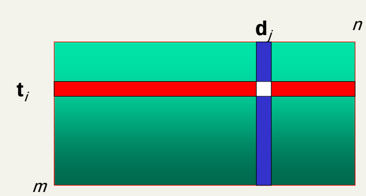
Automatic Thesaurus Generation

- 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.

Why?

Co-occurrence Thesaurus

- Simplest way to compute one is based on term-term similarities in $C = AA^T$ where A is term-document matrix.
- $w_{i,j}$ = (normalized) weighted count (t_i, $\mathbf{d}_j$)



With integer counts – what do you get for a boolean Cooccurrence matrix?

Automatic Thesaurus Generation Example

word	ten nearest neighbors
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed slig
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs gazec
Makeup	repellent lotion glossy sunscreen Skin gel p
mediating	reconciliation negotiate cease conciliation p
keeping	hoping bring wiping could some would othe
lithographs	drawings Picasso Dali sculptures Gauguin I
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate awl

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.

Query Expansion: Summary

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

Pseudo Relevance Feedback

- Automatic local analysis
- 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 (perhaps better than global analysis!)
  - Found to improve performance in TREC ad-hoc task
  - Danger of query drift

Pseudo relevance feedback: Cornell SMART at TREC 4

- Results show number of relevant documents out of top 100 for 50 queries (so out of 5000)
- Results contrast two length normalization schemes (L vs. l), and pseudo relevance feedback (adding 20 terms)

  - lnc.ltc 3210
  - lnc.ltc-PsRF 3634
  - Lnu.ltu 3709
  - Lnu.ltu-PsRF 4350

Indirect relevance feedback

[Forward pointer to CS 276B]

- DirectHit introduced a form of indirect relevance feedback.
- DirectHit ranked documents higher that users look at more often.
- Global: Not user or query specific.

Resources

MG Ch. 4.7

MIR Ch. 5.2 – 5.4

Yonggang Qiu , Hans-Peter Frei, Concept based query expansion. *SIGIR 16*: 161–169, 1993.

Schuetze: Automatic Word Sense Discrimination, Computational Linguistics, 1998.

Singhal, Mitra, Buckley: Learning routing queries in a query zone, ACM SIGIR, 1997.

Buckley, Singhal, Mitra, Salton, New retrieval approaches using SMART: TREC4, NIST, 1996.

Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4):288-297, 1990.

Harman, D. (1992): Relevance feedback revisited. *SIGIR 15*: 1-10

Xu, J., Croft, W.B. (1996): Query Expansion Using Local and Global Document Analysis, in *SIGIR 19*: 4-11