Information Retrieval (CSD510)

Relevance Feedback and Query Expansion

February 20, 2024

How can we improve recall in search?

- Two ways of improving recall: **relevance feedback** and **query expansion**
- Query(q): *car*
- A relevant document d contains *automobile* but not *car*...will not be retrieved by simple IR system.
- Even if d is the most relevant document for q!
- **Objective:** Adapt system to return relevant documents that do not contain queries q.

Improving recall

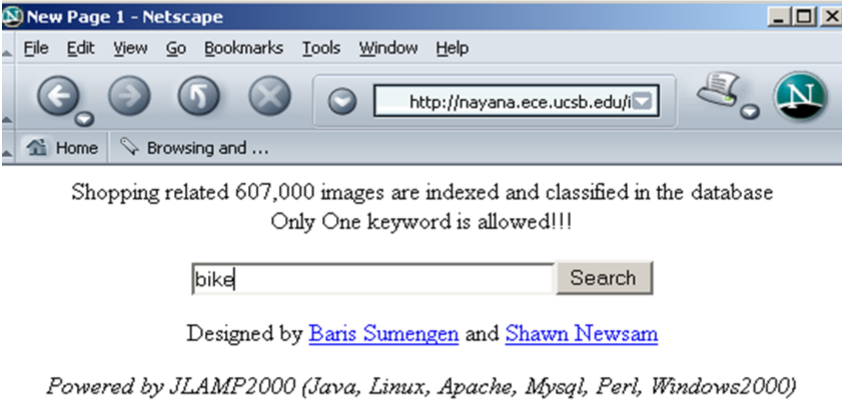
- Global method - Query expansion
  - Use thesauri
  - Automatically generate thesaurus
- Local method
  - Relevance feedback
  - Pseudo relevance feedback

Relevance feedback

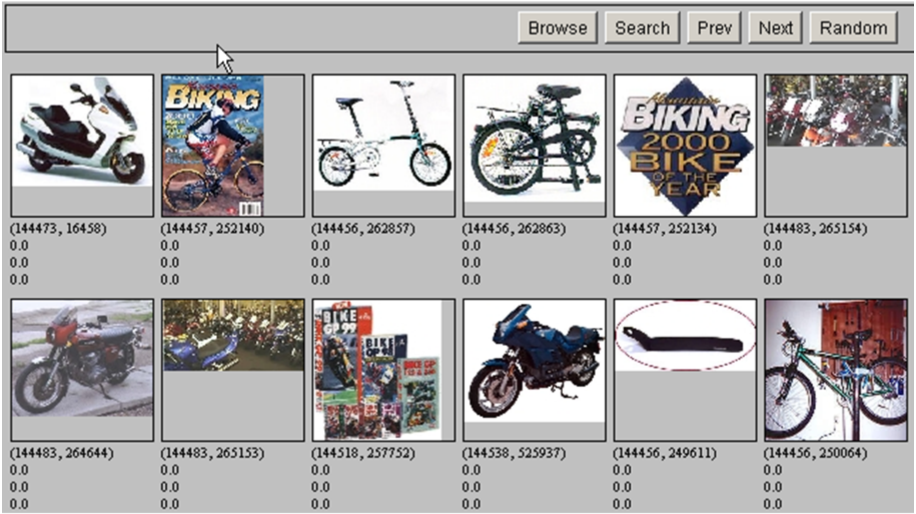
- **Relevance feedback:** user feedback on relevance of docs in initial set of results
  - User issues a (short, simple) query
  - The user marks some results as relevant or non-relevant.
  - The system computes a better representation of the information need based on feedback.
  - Relevance feedback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

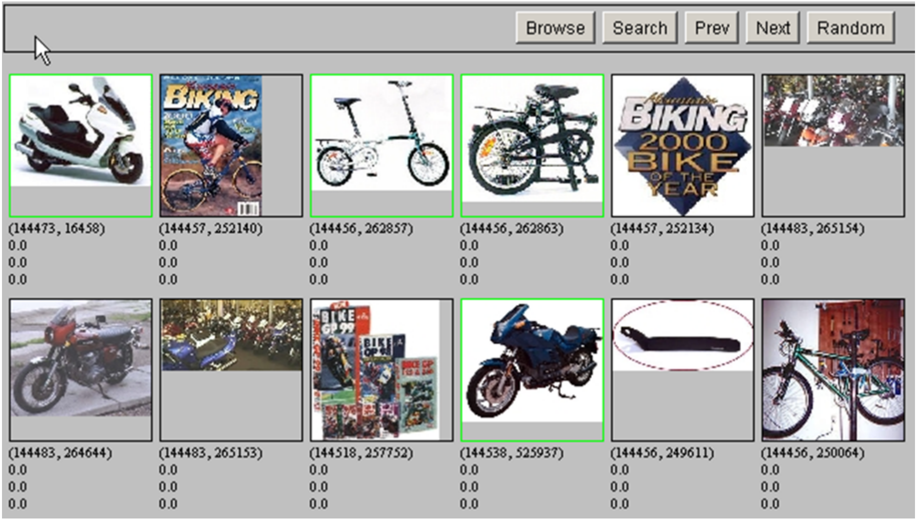
- Image search engine
- http://nayana.ece.ucsb.edu/imsearch/imsearch.html



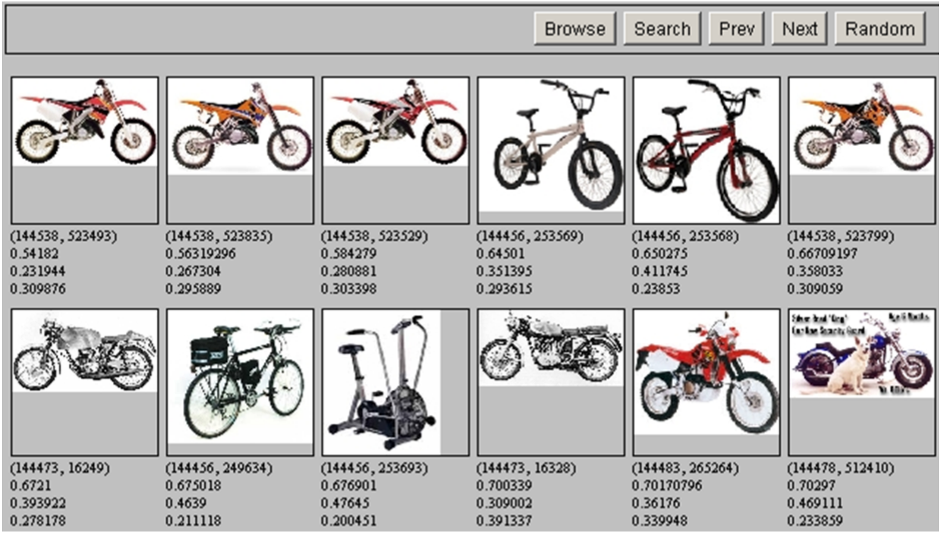
Relevance feedback: Example



Relevance feedback: Example



Relevance feedback: Example



Initial query/results

- Initial query: *New space satellite applications*
  - + 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
- User then marks relevant documents with "+".

Expanded query after relevance feedback

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

Results for expanded query

2 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan

1 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

8 5. 0.492, 12/02/87, Telecommunications Tale of Two Companies

6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use

7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers

8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Key concept: Centroid

- The centroid is the center of mass of a set of points
- Documents as points in a high-dimensional space
- Definition: Centroid

$$\vec{\mu}(C) = \frac{1}{|C|} \sum_{d \in C} \vec{d}$$

where C is the set of documents.

Rocchio Algorithm

- The Rocchio algorithm uses the vector space model to pick a relevance feedback query
- Rocchio seeks the query $\vec{q}_{opt}$ that maximizes

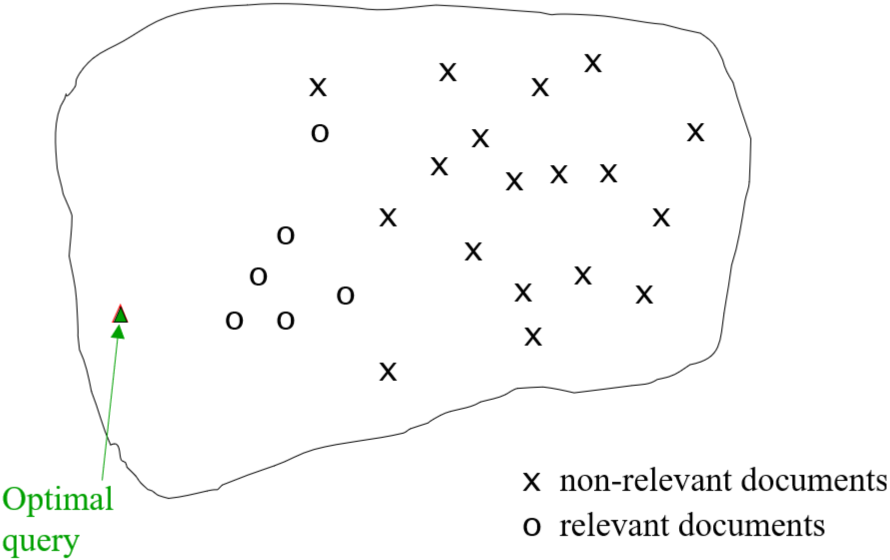
$$\vec{q}_{opt} = argmax_{\vec{q}}[cos(\vec{q}, \vec{\mu}(C_r)) - cos(\vec{q}, \vec{\mu}(C_{nr}))]$$

- Tries to separate docs marked relevant and non-relevant.

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d}_{j \in C_r}} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\vec{d}_{j \notin C_r}} \vec{d}_j$$

- Problem: we don't know the truly relevant docs

The Theoretically Best Query



Rocchio 1971 Algorithm (SMART)

- Used in practice:

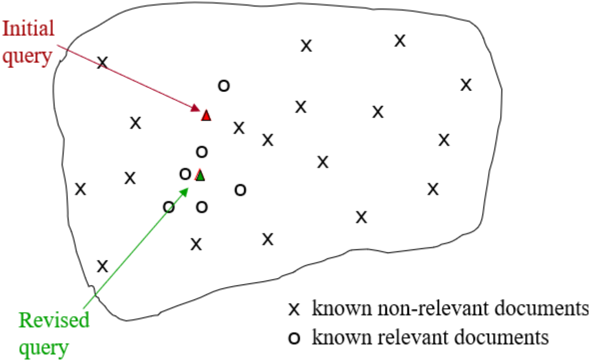
$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

- D_r = set of known relevant doc vectors
- D_{nr} = set of known irrelevant doc vectors
  - Different from C_r and C_{nr}
- q_m = modified query vector; q_0 = original query vector; α, β, γ: weights (hand-chosen or set empirically)
- New query moves toward relevant documents and away from irrelevant documents.

Subtleties to note

- Tradeoff α vs. β/γ : If we have a lot of judged documents, we want a higher β/γ.
- Some weights in query vector can go negative.
    - Negative term weights are ignored (set to 0).

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

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

Evaluation of relevance feedback strategies

- Use q_0 and compute precision and recall graph
- Use q_m and compute precision recall graph
  1. Assess on all documents in the collection...cheating!!
  2. Use documents in residual collection (set of documents minus those assessed relevant)
     - Measures usually then lower than for original query
     - But a 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.

Evaluation of relevance feedback strategies

- Use q_0 and compute precision and recall graph
- Use q_m and compute precision recall graph
  1. Assess on all documents in the collection...cheating!!
     - Partly due to known relevant documents ranked higher
     - Must evaluate with respect to documents not seen by user
  2. Use documents in residual collection (set of documents minus those assessed relevant)
     - Measures usually then lower than for original query
     - But a 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.

Evaluation of relevance feedback

- Second method – assess only the docs not rated by the user in the first round
  - Could make relevance feedback look worse than it really is
  - Can still assess relative performance of algorithms
- Most satisfactory – use two collections each with their own relevance assessments
  - q_0 and user feedback from first collection
  - q_m run on second collection and measured

Relevance Feedback on the Web

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

Pseudo relevance feedback

- Pseudo-relevance feedback automates the "manual" part of true relevance feedback
- Pseudo-relevance algorithm:
  - Retrieve a ranked list of hits for the user's query
  - Assume that the top k documents are relevant.
  - Do relevance feedback (e.g., Rocchio)
- Works very well on average
- But can go horribly wrong for some queries.
- Several iterations can cause query drift.

Indirect relevance feedback

- On the web, DirectHit introduced a form of indirect relevance feedback.
- DirectHit ranked documents higher that users look at more often.
  - Clicked on links are assumed likely to be relevant
- Globally: Not necessarily user or query specific.
  - This is the general area of **clickstream mining**
- Today – handled as part of machine-learned ranking

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

How do we augment the user query?

- Manual thesaurus
  - E.g. MedLine: physician, syn: doc, doctor, MD, medico
  - Can be query rather than just synonyms
- Global Analysis: (static; of all documents in collection)
  - Automatically derived thesaurus
  - Refinements based on query log mining
- Local Analysis: (dynamic)
  - Analysis of documents in result set

Thesaurus-based query expansion

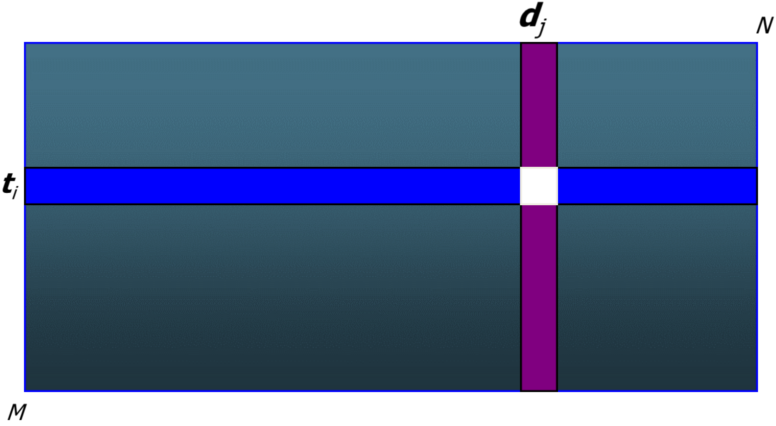
- For each term, *t*, in a query, expand the query with synonyms and related words of *t* from the thesaurus
  - feline → feline cat
- May weight added terms less than original query terms.
- Generally increases recall
- Widely used in many science/engineering fields
- May significantly decrease precision, particularly with ambiguous terms.
  - "interest rat" → "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
- Fundamental notion: similarity between two words
- Definition 1: Two words are similar if they co-occur with similar words.
- Definition 2: Two words are similar if they occur in a given grammatical relation with the same words.
- **Example:** You can harvest, peel, eat, prepare, etc. apples and pears, so apples and pears must be similar.

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) weight for (t_i, d_j)
- For each t_i, pick terms with high values in C



Co-occurence-based thesaurus: Example

Word	Nearest neighbors
absolutely	absurd, whatsoever, totally, exactly, nothing
bottomed	dip, copper, drops, topped, slide, trimmed
captivating	shimmer, stunningly, superbly, plucky, witty
doghouse	dog, porch, crawling, beside, downstairs
makeup	repellent, lotion, glossy, sunscreen, skin, gel
mediating	reconciliation, negotiate, case, conciliation
keeping	hoping, bring, wiping, could, some, would
lithographs	drawings, Picasso, Dali, sculptures, Gauguin
pathogens	toxins, bacteria, organisms, bacterial, parasite
senses	grasp, psyche, truly, clumsy, naive, innate