Recap of the last lecture

- Evaluating a search engine
  - Benchmarks
  - Precision and recall
- Results summaries
Recap: Unranked retrieval evaluation: Precision and Recall

- **Precision**: fraction of retrieved docs that are relevant = $P(\text{relevant}|\text{retrieved})$

- **Recall**: fraction of relevant docs that are retrieved = $P(\text{retrieved}|\text{relevant})$

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Nonrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>Not Retrieved</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

- Precision $P = \frac{tp}{tp + fp}$
- Recall $R = \frac{tp}{tp + fn}$
Recap: A combined measure: $F$

- Combined measure that assesses precision/recall tradeoff is \textit{F measure} (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- People usually use balanced $F_1$ measure
  - i.e., with $\beta = 1$ or $\alpha = \frac{1}{2}$

- Harmonic mean is a conservative average
  - See CJ van Rijsbergen, \textit{Information Retrieval}
This lecture

- Improving results
  - For high recall.
  - E.g., searching for *aircraft* doesn’t match with *plane*; nor *thermodynamic* with *heat*

- Options for improving results...
  - 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 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

- We will use *ad hoc retrieval* to refer to regular retrieval without relevance feedback.
- We now look at four examples of relevance feedback that highlight different aspects.
Similar pages

Google search for "sarah brightman"

Sarah Brightman Official Website - Home Page
Official site of world's best-selling soprano. Join FAN AREA free to access exclusive perks, photo diaries, a global forum community and more...
www.sarah-brightman.com/ - 4k - Cached - Similar pages
Relevance Feedback: Example

- Image search engine http://nayana.ece.ucsb.edu/imsearch/imssearch.html
Results for Initial Query
Relevance Feedback

![Image of search results]

Information Retrieval
Results after Relevance Feedback
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
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
- 30.816 satellite
- 5.991 nasa
- 4.196 launch
- 3.516 instrument
- 3.004 bundespost
- 2.790 rocket
- 2.003 broadcast
- 0.836 oil

- 15.106 space
- 5.660 application
- 5.196 eos
- 3.972 aster
- 3.446 arianespace
- 2.806 ss
- 2.053 scientist
- 1.172 earth
- 0.646 measure
Results for expanded query

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. 0.492, 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.
- Recall that we represent documents as points in a high-dimensional space.
- Definition: Centroid

\[
\bar{\mu}(C) = \frac{1}{|C|} \sum_{d \in C} d
\]

where \( C \) is a set of documents.
Rocchio Algorithm

- The Rocchio algorithm uses the vector space model to pick a relevance feedback query.
- Rocchio seeks the query $q_{opt}$ that maximizes

$$\tilde{q}_{opt} = \arg \max_{\tilde{q}} [\cos(\tilde{q}, \tilde{\mu}(C_r)) - \cos(\tilde{q}, \tilde{\mu}(C_{nr}))]$$

- Tries to separate docs marked relevant and non-relevant

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

- Problem: we don’t know the truly relevant docs
The Theoretically Best Query

Optimal query

X non-relevant documents
O relevant documents
Rocchio 1971 Algorithm (SMART)

- Used in practice:

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

- \( D_r \) = 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; \( \alpha, \beta, \gamma \): weights (hand-chosen or set empirically)
- New query moves toward relevant documents and away from irrelevant documents
Subtleties to note

- Tradeoff $\alpha$ vs. $\beta/\gamma$: If we have a lot of judged documents, we want a higher $\beta/\gamma$.

- Some weights in query vector can go negative
  - Negative term weights are ignored (set to 0)
Relevance feedback on initial query

Initial query

X known non-relevant documents
O known relevant documents

Revised 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 ($\gamma=0$).
Aside: Vector Space can be Counterintuitive.

Doc
“J. Snow & Cholera”

Query
“cholera”

q1 query “cholera”
○ www.ph.ucla.edu/epi/snow.html
x other documents
High-dimensional Vector Spaces

- The queries “cholera” and “john snow” are far from each other in vector space.
- How can the document “John Snow and Cholera” be close to both of them?
- Our intuitions for 2- and 3-dimensional space don't work in >10,000 dimensions.
- 3 dimensions: If a document is close to many queries, then some of these queries must be close to each other.
- Doesn't hold for a high-dimensional space.
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: Problems

- 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 are often reluctant to provide explicit feedback
- It’s often harder to understand why a particular document was retrieved after applying relevance feedback
Evaluation of relevance feedback strategies

- Use $q_0$ and compute precision and recall graph
- Use $q_m$ and compute precision recall graph
  - Assess on 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 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
Evaluation: Caveat

- True evaluation of usefulness must compare to other methods taking the same amount of time.
- Alternative to relevance feedback: User revises and resubmits query.
- Users may prefer revision/resubmission to having to judge relevance of documents.
- There is no clear evidence that relevance feedback is the “best use” of the user’s time.
Relevance Feedback on the Web

- Some search engines offer a similar/related pages feature (this is a trivial form of relevance feedback)
  - Google (link-based)
  - 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.
Excite Relevance Feedback

Spink et al. 2000

- Only about 4% of query sessions from a user used relevance feedback option
  - Expressed as “More like this” link next to each result
- But about 70% of users only looked at first page of results and didn’t pursue things further
  - So 4% is about 1/8 of people extending search
- Relevance feedback improved results about 2/3 of the time
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.

- Why?
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.
Would you expect such a feature to increase the query volume at a search engine?
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
    - (co-occurrence statistics)
  - Refinements based on query log mining
    - Common on the web

- **Local Analysis**: (dynamic)
  - Analysis of documents in result set
Example of manual thesaurus

PubMed Query:

("neoplasms"[MeSH Terms] OR cancer[Text Word])
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* $\rightarrow$ *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” $\rightarrow$ “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.
- You can harvest, peel, eat, prepare, etc. apples and pears, so apples and pears must be similar.
- Co-occurrence based is more robust, grammatical relations are more accurate.
Co-occurrence Thesaurus

- Simplest way to compute one is based on term-term similarities in $C = AA^T$ where $A$ is term-document matrix.
- $w_{i,j} = \text{(normalized) weight for } (t_i, d_j)$

For each $t_i$, pick terms with high values in $C$

What does $C$ contain if $A$ is a term-doc incidence (0/1) matrix?
## Automatic Thesaurus Generation Example

<table>
<thead>
<tr>
<th>word</th>
<th>ten nearest neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>absolutely</td>
<td>absurd whatsoever totally exactly nothing</td>
</tr>
<tr>
<td>bottomed</td>
<td>dip copper drops topped slide trimmed slight</td>
</tr>
<tr>
<td>captivating</td>
<td>shimmer stunningly superbly plucky witty</td>
</tr>
<tr>
<td>doghouse</td>
<td>dog porch crawling beside downstairs gazed</td>
</tr>
<tr>
<td>Makeup</td>
<td>repellent lotion glossy sunscreen Skin gel lotion</td>
</tr>
<tr>
<td>mediating</td>
<td>reconciliation negotiate cease conciliation p</td>
</tr>
<tr>
<td>keeping</td>
<td>hoping bring wiping could some would other</td>
</tr>
<tr>
<td>lithographs</td>
<td>drawings Picasso Dali sculptures Gauguin</td>
</tr>
<tr>
<td>pathogens</td>
<td>toxins bacteria organisms bacterial parasite</td>
</tr>
<tr>
<td>senses</td>
<td>grasp psyche truly clumsy naive innate awkward</td>
</tr>
</tbody>
</table>
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.
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
    - Assuming the displayed summaries are good, etc.
- Globally: Not necessarily user or query specific.
  - This is the general area of **clickstream mining**
- Today – handled as part of machine-learned ranking
Resources

IIR Ch 9
MG Ch. 4.7
MIR Ch. 5.2 – 5.4