Query reformulation
CE-324: Modern Information Retrieval
Sharif University of Technology

M. Soleymani
Fall 2014

Most slides have been adapted from: Profs. Manning, Nayak & Raghavan (CS-276, Stanford)
Outline

- Improving results to obtain high recall

- Options for improving results...
  - Local methods
    - Relevance feedback (user feedback)
    - Indirect relevance feedback (e.g., user clicks)
    - Pseudo relevance feedback
  - Global methods
    - Query expansion
      - Thesauri
      - Automatic thesaurus generation
    - Techniques like spelling correction

No participation of the user
Relevance Feedback

- User feedback on relevance of docs in 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
Results for Initial Query
Relevance Feedback
Results after Relevance Feedback

<table>
<thead>
<tr>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
<th>Image 5</th>
<th>Image 6</th>
<th>Image 7</th>
<th>Image 8</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="image4.png" alt="Image 4" /></td>
<td><img src="image5.png" alt="Image 5" /></td>
<td><img src="image6.png" alt="Image 6" /></td>
<td><img src="image7.png" alt="Image 7" /></td>
<td><img src="image8.png" alt="Image 8" /></td>
</tr>
<tr>
<td>(144538, 523493)</td>
<td>(144538, 523835)</td>
<td>(144538, 523529)</td>
<td>(144456, 253569)</td>
<td>(144456, 253568)</td>
<td>(144538, 523799)</td>
<td>(144473, 16249)</td>
<td>(144456, 249634)</td>
</tr>
<tr>
<td>0.54182</td>
<td>0.56319296</td>
<td>0.584279</td>
<td>0.64501</td>
<td>0.650275</td>
<td>0.66709197</td>
<td>0.6721</td>
<td>0.675018</td>
</tr>
<tr>
<td>0.231944</td>
<td>0.267304</td>
<td>0.280881</td>
<td>0.351395</td>
<td>0.411745</td>
<td>0.358033</td>
<td>0.393922</td>
<td>0.4639</td>
</tr>
<tr>
<td>0.309876</td>
<td>0.295889</td>
<td>0.303398</td>
<td>0.293615</td>
<td>0.23853</td>
<td>0.309059</td>
<td>0.278181</td>
<td>0.211118</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image 9</th>
<th>Image 10</th>
<th>Image 11</th>
<th>Image 12</th>
<th>Image 13</th>
<th>Image 14</th>
<th>Image 15</th>
<th>Image 16</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image9.png" alt="Image 9" /></td>
<td><img src="image10.png" alt="Image 10" /></td>
<td><img src="image11.png" alt="Image 11" /></td>
<td><img src="image12.png" alt="Image 12" /></td>
<td><img src="image13.png" alt="Image 13" /></td>
<td><img src="image14.png" alt="Image 14" /></td>
<td><img src="image15.png" alt="Image 15" /></td>
<td><img src="image16.png" alt="Image 16" /></td>
</tr>
<tr>
<td>(1444483, 265264)</td>
<td>(144478, 512410)</td>
<td>(144483, 265264)</td>
<td>(144473, 16328)</td>
<td>(144473, 16249)</td>
<td>(144456, 253693)</td>
<td>(144456, 253693)</td>
<td>(144456, 253693)</td>
</tr>
<tr>
<td>0.700339</td>
<td>0.70297</td>
<td>0.70170796</td>
<td>0.70170796</td>
<td>0.70170796</td>
<td>0.70170796</td>
<td>0.70170796</td>
<td>0.70170796</td>
</tr>
<tr>
<td>0.36176</td>
<td>0.469111</td>
<td>0.36176</td>
<td>0.36176</td>
<td>0.36176</td>
<td>0.36176</td>
<td>0.36176</td>
<td>0.36176</td>
</tr>
<tr>
<td>0.339948</td>
<td>0.233859</td>
<td>0.339948</td>
<td>0.339948</td>
<td>0.339948</td>
<td>0.339948</td>
<td>0.339948</td>
<td>0.339948</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coordinates</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(144538, 523493)</td>
<td>0.54182</td>
<td>0.56319296</td>
<td>0.584279</td>
<td>0.64501</td>
<td>0.650275</td>
<td>0.66709197</td>
<td>0.6721</td>
<td>0.675018</td>
</tr>
<tr>
<td>0.231944</td>
<td>0.267304</td>
<td>0.280881</td>
<td>0.351395</td>
<td>0.411745</td>
<td>0.358033</td>
<td>0.393922</td>
<td>0.4639</td>
<td></td>
</tr>
<tr>
<td>0.309876</td>
<td>0.295889</td>
<td>0.303398</td>
<td>0.293615</td>
<td>0.23853</td>
<td>0.309059</td>
<td>0.278181</td>
<td>0.211118</td>
<td></td>
</tr>
<tr>
<td>(1444483, 265264)</td>
<td>0.700339</td>
<td>0.70297</td>
<td>0.70170796</td>
<td>0.70170796</td>
<td>0.70170796</td>
<td>0.70170796</td>
<td>0.70170796</td>
<td>0.70170796</td>
</tr>
<tr>
<td>0.36176</td>
<td>0.469111</td>
<td>0.36176</td>
<td>0.36176</td>
<td>0.36176</td>
<td>0.36176</td>
<td>0.36176</td>
<td>0.36176</td>
<td></td>
</tr>
<tr>
<td>0.339948</td>
<td>0.233859</td>
<td>0.339948</td>
<td>0.339948</td>
<td>0.339948</td>
<td>0.339948</td>
<td>0.339948</td>
<td>0.339948</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>new</td>
<td>space</td>
<td>15.106</td>
</tr>
<tr>
<td>satellite</td>
<td>application</td>
<td>5.660</td>
</tr>
<tr>
<td>nasa</td>
<td>eos</td>
<td>5.196</td>
</tr>
<tr>
<td>launch</td>
<td>aster</td>
<td>3.972</td>
</tr>
<tr>
<td>instrument</td>
<td>arianespace</td>
<td>3.446</td>
</tr>
<tr>
<td>bundespost</td>
<td>ss</td>
<td>2.806</td>
</tr>
<tr>
<td>rocket</td>
<td>scientist</td>
<td>2.053</td>
</tr>
<tr>
<td>broadcast</td>
<td>earth</td>
<td>1.172</td>
</tr>
<tr>
<td>oil</td>
<td>measure</td>
<td>0.646</td>
</tr>
</tbody>
</table>
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

- **Centroid**: center of mass of a set of points
  - Docs as points in a high-dimensional space

- **Definition: Centroid**

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

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

- Vector space model to pick a relevance feedback query
- **Rocchio** seeks the query $\tilde{q}_{opt}$
  \[
  \tilde{q}_{opt} = \arg \max_{\tilde{q}} [\cos(\tilde{q}, \bar{\mu}(C_r)) - \cos(\tilde{q}, \bar{\mu}(C_{nr}))]
  \]
- Tries to separate docs marked relevant and non-relevant
  \[
  \tilde{q}_{opt} = \frac{1}{|C_r|} \sum_{d_j \in C_r} \tilde{d}_j - \frac{1}{|C_{nr}|} \sum_{d_j \notin C_r} \tilde{d}_j
  \]
- **Problem**: we don’t know the truly relevant docs
The theoretically best query

- x non-relevant documents
- o relevant documents
Rocchio 1971 algorithm (SMART)

- Used in practice:
  - New query moves toward relevant docs and away from irrelevant docs

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

- \(D_r\) = 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)
Rocchio 1971 algorithm (SMART)

- The original query has been added
  - It contains important information for determining relevant docs

- Tradeoff $\alpha$ vs. $\beta$ (or $\gamma$)
  - When a lot of judged docs, we want a higher $\beta$ (or $\gamma$).

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

- Known non-relevant documents: X
- Known relevant documents: o

Initial query

Revised query
Relevance Feedback in vector spaces

- We can modify the query based on RF
  - Apply standard vector space model
  - Use only the docs that were marked

- RF can improve recall and precision

- RF is most useful for increasing *recall*
  - 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 ($\gamma = 0$).
Relevance feedback: Assumptions

- A1: User has sufficient knowledge for initial query.

- A2: Relevance prototypes are “well-behaved”
  - Term distribution in relevant docs will be similar
    - either: All relevant docs are tightly clustered around a prototype.
    - or: Different prototypes, but significant vocabulary overlap.
  - While similarities between relevant and irrelevant docs are small
    - Term distribution in non-relevant docs will be different from those in relevant docs

Multi-modal distribution of relevant docs?
Violation of A1

- User does not have sufficient initial knowledge.

Examples:
- Misspellings
- 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
- Often harder to understand why a particular doc was retrieved after applying RF
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 docs in the collection
    - Spectacular improvements, but … it’s cheating!
    - Partly due to known relevant docs ranked higher
    - Must evaluate with respect to docs not seen by user
  - Use docs in residual collection (set of docs minus assessed relevant)
    - Measures usually then lower than for original query
    - But a more realistic evaluation
    - Relative performance of variant RF methods can be validly compared
    - Difficult to comparing performance with and without RF

- Empirically, one round of RF is often very useful.
Evaluation of relevance feedback

- Assess only the docs *not* rated by the user in the first round
  - Could make RF look worse than it really is
  - Can still assess relative performance of algorithms

- Most satisfactory – use two collections each with their own relevance assessments
  - Results of $q_0$ and user feedback from the first collection
  - $q_m$ run on the second collection and measured
Evaluation: Caveat

- True evaluation of usefulness
  - compare to other methods taking the same amount of time.

- Alternative to RF: User revises and resubmits query.
  - Users may prefer revision/resubmission to having to judge relevance of docs.

- No clear evidence that RF is the “best use” of the user’s time.
Relevance feedback on the Web

- RF has been little used in web search.

- Some search engines offer a similar/related pages feature (this is a trivial form of RF)
  - Google (link-based), Altavista, Stanford WebBase

- But some don’t:
  - Alltheweb, bing, Yahoo

- Because:
  - Most people would like to complete their search in a single interaction
  - It’s hard to explain to average user
  - And also web search users are rarely concerned with getting sufficient recall
Excite relevance feedback

Excite initially had true RF, but abandoned it due to lack of use.

[Spink et al. 2000]

- Only about 4% of query sessions used RF option
  - Usually 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
- RF improved results about 2/3 of the time
Click as a metric of user preferences

- RF over the years have been relaxed to allow the use of information that is expected to be related to the query

- When user looks at the snippet of a result and decides to skip it to click on a lower result (in the ranking):
  - We can say user prefers the clicked on result to the upper results
Indirect relevance feedback

- Clicks on links were assumed to indicate that the page was likely relevant to the query
  - Assuming the displayed summaries are good, etc.

- On the web, DirectHit introduced an indirect RF.
  - DirectHit ranked docs higher that users click at more often.
  - Data about the click rates on pages was gathered globally, not necessarily user or query specific.
    - Use *clickstream mining*
Pseudo relevance feedback

- Automates the “manual” part of true RF:
  - Retrieve a ranked list of hits for the user’s query
  - Assume that the top k docs 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 through global analysis

- Local analysis methods
  - Extract information from a local set of docs retrieved to expand the query

- Global analysis methods
  - No feedback from user on the retrieved docs is required
Global methods

- Aiding the user in query formulation
- Manual thesaurus
- Building a thesaurus automatically
  - Using information from the whole set of docs to expand the query
Thesaurus

- Thesaurus: containing synonyms and related words of terms
  - Unlike a dictionary, it does not give the definition of words.

physician

syn: ||doc, doctor, MD, medical, mediciner, medico,
    ||sawbones
rel: medic, general practitioner, surgeon,
Thesaurus-based query expansion

- Expand the query with synonyms (or related words) from the thesaurus
  - Example: feline → feline cat
  - Query expansion can be interactive or automatic

- May weight added terms less than original query terms.

- Generally increases recall

- May significantly decrease precision, particularly with ambiguous terms.
  - Apple computer → Apple red fruit computer
How do we augment the user query?

- **Manual thesaurus or controlled vocabulary**
  - Thesaurus: preferred terms, synonyms, broader, or narrower terms

- **Automatic thesaurus**
  - Automatically derived thesaurus

- **Query log mining**
  - Query reformulation based on query log mining
    - Manual query reformulation of other users
Example of manual thesaurus

PubMed Query:

("neoplasms"[MeSH Terms] OR cancer[Text Word])
Manual thesaurus

- Widely used in many science/engineering fields
  - Domain-particular vocabularies

- There is a high cost of manually producing a thesaurus
  - And for updating it for scientific and terminological developments
Automatic thesaurus generation

- Attempt to generate a thesaurus automatically
  - by analyzing the collection of docs

Fundamental notion: *similarity between two words*

- **Definition 1:** Two words are similar if they *co-occur* with similar words.
  - Or co-occur in a doc or paragraph

- **Definition 2:** Two words are similar if they occur in a given grammatical relation with the same words.
Co-occurrence thesaurus

- $A_{ij} = \text{(normalized) weight for } (t_i, d_j)$

- Term-term similarities: $C = AA^T$

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

A is term-doc matrix:
Co-occurrence thesaurus

\[ a_{ij} = \frac{\text{TF}_{i,j} \times \text{ITF}_j}{\sqrt{\sum_{i=1}^{M} (\text{TF}_{i,j} \times \text{ITF}_j)^2}} \]

- \[ C_{i,i'} = \sum_{j=1}^{N} a_{ij} \times a_{i'j} \]
Automatic thesaurus generation: Example

For each $t_i$, pick some terms $t_{i'}$ with the highest values in $C_{i,i'}$

<table>
<thead>
<tr>
<th>Word</th>
<th>Nearest neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>absolutely</td>
<td>absurd, whatsoever, totally, exactly,</td>
</tr>
<tr>
<td>bottomed</td>
<td>nothing</td>
</tr>
<tr>
<td>captivating</td>
<td>dip, copper, drops, topped, slide,</td>
</tr>
<tr>
<td>doghouse</td>
<td>trimmed</td>
</tr>
<tr>
<td>makeup</td>
<td>shimmer, stunningly, superbly, plucky,</td>
</tr>
<tr>
<td>mediating</td>
<td>witty</td>
</tr>
<tr>
<td>keeping</td>
<td>dog, porch, crawling, beside, downstairs</td>
</tr>
<tr>
<td>lithographs</td>
<td>repellent, lotion, glossy, sunscreen,</td>
</tr>
<tr>
<td>pathogens</td>
<td>skin, gel</td>
</tr>
<tr>
<td>senses</td>
<td>reconciliation, negotiate, case,</td>
</tr>
<tr>
<td></td>
<td>conciliation</td>
</tr>
<tr>
<td></td>
<td>hoping, bringing, wiping, could, some,</td>
</tr>
<tr>
<td></td>
<td>would</td>
</tr>
<tr>
<td></td>
<td>drawings, Picasso, Dali, sculptures,</td>
</tr>
<tr>
<td></td>
<td>Gauguin</td>
</tr>
<tr>
<td></td>
<td>toxins, bacteria, organisms, bacterial,</td>
</tr>
<tr>
<td></td>
<td>parasite</td>
</tr>
<tr>
<td></td>
<td>grasp, psyche, truly, clumsy, naive,</td>
</tr>
<tr>
<td></td>
<td>innate</td>
</tr>
</tbody>
</table>
Automatic thesaurus generation: grammatical approach

- Grammatical relation: Example
  - harvest, peel, eat (apples)
  - harvest, peel, eat (pears)
  - $\Rightarrow$ apples and pears must be similar.

- Co-occurrence based is more robust.

- Grammatical relations are more accurate.
Query expansion by thesaurus

- Represent the query in the same vector space used for representing the index terms

- Based on the thesaurus find the words correlated with query terms

- Expand the query with the terms having highest $sim(q, v)$ (similarity of the term and whole query)
Automatic thesaurus generation: Discussion

- Quality of associations is usually a problem.

- Term ambiguity ⇒ irrelevant statistically correlated terms.
  - Example: “Apple computer” → “Apple red fruit computer”

- **Problems:**
  - False positives: Words deemed similar that are not
  - False negatives: Words deemed dissimilar that are similar

- Expansion may not retrieve many additional docs.
  - Since terms are highly correlated anyway
Query expansion vs. relevance feedback

- Less successful

- May significantly decrease the precision
  - Especially when the query contains ambiguous terms

- More understandable to the system user
  - Not requiring user input
Query assist

- How to generate alternative or expanded queries?
  - Query log mining

Would you expect such a feature to increase the query volume at a search engine?
Query assist

- User gives feedback on words or phrases (in query expansion)
  - As opposed to user feedback on docs in relevance feedback
Resources

IIR Ch 9
MG Ch. 4.7
MIR Ch. 5