Computer! Translate into Russian: "We need a courier who we can trust with sensitive documents."

Don't you mean "whom"?
Relevance Feedback
Query Expansion

David Kauchak

cs160
Fall 2009

adapted from:
http://www.stanford.edu/class/cs276/handouts/lecture9-queryexpansion.ppt
Administrative

- Assignment 3 out
Google’s page search
Anomalous State of Knowledge

- Basic paradox:
  - Information needs arise because the user doesn’t know something: “an anomaly in his state of knowledge with respect to the problem faced”
  - Search systems are designed to satisfy these needs, but the user needs to know what he is looking for
  - However, if the user knows what he’s looking for, there may not be a need to search in the first place
What is actually returned...
What does “similar pages” do?

Does this solve our problem?
Relevance feedback

- User provides feedback on relevance of documents in the initial set of results
  - User issues a query
  - The user marks some results as relevant or non-relevant
  - The system computes a better results based on the feedback
  - May iterate
An example

Image search engine:
http://nayana.ece.ucsb.edu/imsearch/imsearch.html
Results for initial query
## Results after Relevance Feedback

<table>
<thead>
<tr>
<th>Image 1</th>
<th>Score 1</th>
<th>Score 2</th>
<th>Score 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td>0.54182</td>
<td>0.267304</td>
<td>0.295889</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image 2" /></td>
<td>0.56319296</td>
<td>0.280881</td>
<td>0.303398</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image 3" /></td>
<td>0.584279</td>
<td>0.351395</td>
<td>0.293615</td>
</tr>
<tr>
<td><img src="image4.png" alt="Image 4" /></td>
<td>0.64501</td>
<td>0.50275</td>
<td>0.23853</td>
</tr>
<tr>
<td><img src="image5.png" alt="Image 5" /></td>
<td>0.66709197</td>
<td>0.338033</td>
<td>0.309059</td>
</tr>
<tr>
<td><img src="image6.png" alt="Image 6" /></td>
<td>0.6721</td>
<td>0.4639</td>
<td>0.211118</td>
</tr>
<tr>
<td><img src="image7.png" alt="Image 7" /></td>
<td>0.675018</td>
<td>0.47645</td>
<td>0.200451</td>
</tr>
<tr>
<td><img src="image8.png" alt="Image 8" /></td>
<td>0.676901</td>
<td>0.390902</td>
<td>0.391337</td>
</tr>
<tr>
<td><img src="image9.png" alt="Image 9" /></td>
<td>0.700339</td>
<td>0.36176</td>
<td>0.339948</td>
</tr>
<tr>
<td><img src="image10.png" alt="Image 10" /></td>
<td>0.70170796</td>
<td>0.469111</td>
<td>0.233859</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image 11</th>
<th>Score 1</th>
<th>Score 2</th>
<th>Score 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image11.png" alt="Image 11" /></td>
<td>0.675018</td>
<td>0.47645</td>
<td>0.200451</td>
</tr>
<tr>
<td><img src="image12.png" alt="Image 12" /></td>
<td>0.676901</td>
<td>0.390902</td>
<td>0.391337</td>
</tr>
<tr>
<td><img src="image13.png" alt="Image 13" /></td>
<td>0.700339</td>
<td>0.36176</td>
<td>0.339948</td>
</tr>
<tr>
<td><img src="image14.png" alt="Image 14" /></td>
<td>0.70170796</td>
<td>0.469111</td>
<td>0.233859</td>
</tr>
</tbody>
</table>
Ideas?

- For ranking models we represent our query as a vector of weights, which we view as a point in a high dimensional space:

  0 4 0 8 0 0

- We want to bias the query towards documents that the user selected (the “relevant documents”)
- We want to bias the query away from documents that the user did not select (the “non-relevant documents”)

Relevance feedback

- X known non-relevant documents
- O known relevant documents

Initial query
Relevance feedback

x known non-relevant documents
o known relevant documents
Relevance feedback on initial query

Initial query

Revised query

How can we "move" the query?

x known non-relevant documents
o known relevant documents
The Rocchio algorithm uses the vector space model to pick a better query.

Rocchio seeks the query $q_{opt}$ that maximizes the difference between the query similarity with the relevant set of documents ($C_r$) vs. the non-relevant set of documents ($C_{nr}$)

$$
\tilde{q}_{opt} = \arg \max_{\tilde{q}} [\text{sim}(\tilde{q}, C_r) - \text{sim}(\tilde{q}, C_{nr})]
$$
The centroid is the center of mass of a set of points.

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

- Find the new query by moving it towards the centroid of the relevant queries and away from the centroid of the non-relevant queries

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

query vector = original query vector + relevant vector − non-relevant vector

Original query

Relevant centroid

Non-relevant centroid

New query
Rocchio in action

query vector = original query vector
+ relevant vector
− non-relevant vector

Original query

Relevant centroid

Non-relevant centroid

New query
Rocchio in action

source: Fernando Diaz
Rocchio in action

source: Fernando Diaz
User feedback: Select what is relevant

source: Fernando Diaz
Results after relevance feedback

source: Fernando Diaz
Any problems with this?

\[ \tilde{q}_{opt} = \frac{1}{|C_r|} \sum_{d_j \in C_r} \tilde{d}_j - \frac{1}{|C_{nr}|} \sum_{d_j \in C_{nr}} \tilde{d}_j \]

Ignores the original query!

C_r and C_{nr} are *all* the relevant and non-relevant documents

In practice, we don’t know all of these
Rocchio 1971 Algorithm (SMART)

- Used in practice:

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

- \( D_r \) = 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
Relevance Feedback in vector spaces

- 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 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 ($\gamma = 0$)
Another example

- **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</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>new</td>
<td>2.074</td>
</tr>
<tr>
<td>satellite</td>
<td>30.816</td>
</tr>
<tr>
<td>nasa</td>
<td>5.991</td>
</tr>
<tr>
<td>launch</td>
<td>4.196</td>
</tr>
<tr>
<td>instrument</td>
<td>3.516</td>
</tr>
<tr>
<td>bundespost</td>
<td>3.004</td>
</tr>
<tr>
<td>rocket</td>
<td>2.790</td>
</tr>
<tr>
<td>broadcast</td>
<td>2.003</td>
</tr>
<tr>
<td>oil</td>
<td>0.836</td>
</tr>
<tr>
<td>space</td>
<td>15.106</td>
</tr>
<tr>
<td>application</td>
<td>5.660</td>
</tr>
<tr>
<td>eos</td>
<td>5.196</td>
</tr>
<tr>
<td>aster</td>
<td>3.972</td>
</tr>
<tr>
<td>arianespace</td>
<td>3.446</td>
</tr>
<tr>
<td>ss</td>
<td>2.806</td>
</tr>
<tr>
<td>scientist</td>
<td>2.053</td>
</tr>
<tr>
<td>earth</td>
<td>1.172</td>
</tr>
<tr>
<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
### 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

Any problem with this?
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
Will relevance feedback work?

- Brittany Speers
- hígado
- Cosmonaut
RF assumes the user has sufficient knowledge for initial query

- Misspellings - Brittany Speers
- Cross-language information retrieval – hígado
- Mismatch of searcher’s vocabulary vs. collection vocabulary
  - Cosmonaut/astronaut
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
  - msn live.com
  - 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 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
- What is query drift?
Expanding the query

- We would like to suggest alternative query formulations to the user with the goal of:
  - increasing precision
  - increasing recall

- What are methods we might try to accomplish this?
Increasing precision

- Query assist:
  - Generally done by query log mining
  - Recommend frequent recent queries that contain partial string typed by user (or query typed)
Increasing precision…

http://googleblog.blogspot.com/2009/03/two-new-improvements-to-google-results.html
Increasing recall: query expansion

- Automatically expand the query with related terms and run through index
- Spelling correction can be thought of a special case of this

cosmonaut  cosmonaut astronaut space pilot

How might we come up with these expansions?
How do we augment the user query?

- Manual thesaurus
  - E.g. MedLine: physician, syn: doc, doctor, MD, medico
  - Wordnet

- 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
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.
- 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

- Given a large collection of documents, how might we determine if two words are synonyms?
- Two words are synonyms if they co-occur with similar words

I drive a car
I bought new tires for my car
can I hitch a ride with you in your car

I drive an automobile
I bought new tires for my automobile
can I hitch a ride with you in your automobile
Automatic thesaurus generation

- Given a large collection of documents, how might we determine if two words are synonyms?
- Two words are synonyms if they co-occur with similar words

I drive a car
I bought new tires for my car
can I hitch a ride with you in your car

I drive an automobile
I bought new tires for my automobile
can I hitch a ride with you in your automobile
### 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 slightly</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 powder</td>
</tr>
<tr>
<td>mediating</td>
<td>reconciliation negotiate cease conciliation</td>
</tr>
<tr>
<td>keeping</td>
<td>hoping bring wiping could some would otherwise</td>
</tr>
<tr>
<td>lithographs</td>
<td>drawings Picasso Dali sculptures Gauguin Dali</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