Relevance Feedback and Query Expansion

Debapriyo Majumdar
Information Retrieval – Spring 2015
Indian Statistical Institute Kolkata
Importance of Recall

- Academic importance
- Not only of academic importance
  - Uncertainty about availability of information: are the returned documents relevant at all?
  - Query words may return small number of documents, none so relevant
  - Relevance is not graded, but documents missed out could be more useful to the user in practice
- What could have gone wrong?
  - Many things, for instance …
  - Some other choice of query words would have worked better
  - Searched for **aircraft**, results containing only **plane** were not returned
The gap between the user and the system

User needs some information

A retrieval system tries to bridge this gap

Assumption: the required information is present somewhere

The gap

- The retrieval system can only rely on the query words (in the simple setting)
- Wish: if the system could get another chance …
The gap between the user and the system

User needs some information

A retrieval system tries to bridge this gap

Assumption: the required information is present somewhere

If the system gets another chance
- Modify the query to fill the gap better
- Usually more query terms are added $\rightarrow$ query expansion
- The whole framework is called relevance feedback
Relevance Feedback

- **User** issues a query
  - Usually short and simple query
- The **system** returns some results
- 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.
  - It may be difficult to formulate a good query when you don’t know the collection well, so iterate
Example: similar pages

Old time Google

If you (the user) tell me that this result is relevant, I can give you more such relevant documents
Example 2: 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 some relevant documents with “+”
<table>
<thead>
<tr>
<th>Weight</th>
<th>Term</th>
<th>Weight</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.074</td>
<td>new</td>
<td>15.106</td>
<td>space</td>
</tr>
<tr>
<td>30.816</td>
<td>satellite</td>
<td>5.660</td>
<td>application</td>
</tr>
<tr>
<td>5.991</td>
<td>nasa</td>
<td>5.196</td>
<td>eos</td>
</tr>
<tr>
<td>4.196</td>
<td>launch</td>
<td>3.972</td>
<td>aster</td>
</tr>
<tr>
<td>3.516</td>
<td>instrument</td>
<td>3.446</td>
<td>arianespace</td>
</tr>
<tr>
<td>3.004</td>
<td>bundespost</td>
<td>2.806</td>
<td>ss</td>
</tr>
<tr>
<td>2.790</td>
<td>rocket</td>
<td>2.053</td>
<td>scientist</td>
</tr>
<tr>
<td>2.003</td>
<td>broadcast</td>
<td>1.172</td>
<td>earth</td>
</tr>
<tr>
<td>0.836</td>
<td>oil</td>
<td>0.646</td>
<td>measure</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
The theoretically best query

The information need is best “realized” by the relevant and non-relevant documents

Optimal query

X non-relevant documents
O relevant documents
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

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

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**

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

- 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
Rocchio Algorithm (SMART system)

- 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
- 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

- Known non-relevant documents: X
- Known relevant documents: O

Initial query

Revised query
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 ($\gamma=0$).
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
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
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: 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
Relevance Feedback on the Web

- Some search engines offer a similar/related pages feature (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.
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
  - We will study methods to build automatic thesaurus later in the course
Sources and Acknowledgements

- Several slides are adapted from the slides by Prof. Nayak and Prof. Raghavan for their course in Stanford University