Advanced topics in Computer Science

Jiaheng Lu
Department of Computer Science
Renmin University of China
www.jiahenglu.net
Putting it all together

Documents

- Parsing linguistics
- Indexers
  - Zone and field indexes
  - Inexact top $K$ retrieval
  - Tiered inverted positional index
  - $k$-gram

Indexes

User query

- Freetext query parser
- Spell correction
- Scoring and ranking

Results page

Heap

Training set

Scoring parameters

Putting it all together
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…
  - Focus on relevance feedback
  - The complete landscape
    - Global methods
      - Query expansion
      - Thesauri
      - Automatic thesaurus generation
    - Local methods
      - Relevance feedback
      - Pseudo relevance feedback
Query expansion
Relevance Feedback

- Relevance feedback: user feedback on relevance of docs in initial set of results
  - User issues a (short, simple) query
  - The user marks returned documents 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.

Shopping related 607,000 images are indexed and classified in the database
Only One keyword is allowed!!!

bike

Designed by Baris Sumengen and Shawn Newsam

Powered by JLAMP2000 (Java, Linux, Apache, Mysql, Perl, Windows2000)
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" alt="Motorcycle" /></td>
<td><img src="image2" alt="Motorcycle" /></td>
<td><img src="image3" alt="Motorcycle" /></td>
<td><img src="image4" alt="Motorcycle" /></td>
<td><img src="image5" alt="Bike" /></td>
<td><img src="image6" alt="Bike" /></td>
<td><img src="image7" alt="Bike" /></td>
<td><img src="image8" alt="Motorcycle" /></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>(144456, 253693)</td>
<td>(144473, 16249)</td>
</tr>
<tr>
<td>0.54182</td>
<td>0.56319296</td>
<td>0.584279</td>
<td>0.64301</td>
<td>0.650275</td>
<td>0.66709197</td>
<td>0.675018</td>
<td>0.6721</td>
</tr>
<tr>
<td>0.231944</td>
<td>0.267304</td>
<td>0.280881</td>
<td>0.354519</td>
<td>0.411745</td>
<td>0.358033</td>
<td>0.4639</td>
<td>0.393922</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>0.211118</td>
<td>0.278178</td>
</tr>
</tbody>
</table>
Relevance feedback on initial query

- **x** known non-relevant documents
- **o** known relevant documents
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.
Probabilistic relevance feedback

- Rather than reweighting in a vector space…
- If user has told us some relevant and irrelevant documents, then we can proceed to build a classifier, such as a Naive Bayes model:
  \[
  P(t_k|R) = \frac{|D_{rk}|}{|D_r|}
  \]
  \[
  P(t_k|NR) = \frac{(N_k - |D_{rk}|)}{(N - |D_r|)}
  \]
  - \( t_k \) = term in document; \( D_{rk} \) = known relevant doc containing \( t_k \); \( N_k \) = total number of docs containing \( t_k \)
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.
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

- Why do most search engines not use relevance feedback?
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 apply relevance feedback
Relevance Feedback Example:
Initial Query and Top 8 Results

- 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
Relevance Feedback Example:
Expanded Query

- 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
Top 8 Results After Relevance Feedback

- + 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
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
Relevance Feedback on the Web
[in 2003: now less major search engines, but same general story]

- 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
  - 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
Other Uses of Relevance Feedback

- Following a changing information need
- Maintaining an information filter (e.g., for a news feed)
- Active learning
  
  [Deciding which examples it is most useful to know the class of to reduce annotation costs]
Relevance Feedback

Summary

- Relevance feedback has been shown to be very effective at improving relevance of results.
  - Requires enough judged documents, otherwise it’s unstable ($\geq 5$ recommended)
  - Requires queries for which the set of relevant documents is medium to large
- Full relevance feedback is painful for the user.
- Full relevance feedback is not very efficient in most IR systems.
The complete landscape

- Global methods
  - Query expansion/reformulation
    - Thesauri (or WordNet)
    - Automatic thesaurus generation
  - Global indirect relevance feedback
- Local methods
  - Relevance feedback
  - Pseudo relevance feedback
Query Reformulation: Vocabulary Tools

- Feedback
  - Information about stop lists, stemming, etc.
  - Numbers of hits on each term or phrase

- Suggestions
  - Thesaurus
  - Controlled vocabulary
  - Browse lists of terms in the inverted index
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.
Query Expansion: Example

Also: see www.altavista.com, www.teoma.com
Types of Query Expansion

● Global Analysis: (static; of all documents in collection)
  ● Controlled vocabulary
    ● Maintained by editors (e.g., medline)
  ● Manual thesaurus
    ● E.g. MedLine: physician, syn: doc, doctor, MD, medico
  ● Automatically derived thesaurus
    ● (co-occurrence statistics)
  ● Refinements based on query log mining
    ● Common on the web

● Local Analysis: (dynamic)
Controlled Vocabulary
Thesaurus-based Query Expansion

- This doesn’t require user input
- 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”
  - And for updating it for scientific changes
Automatic Thesaurus Generation

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

- Two main approaches
  - Co-occurrence based (co-occurring words are more likely to be similar)
  - Shallow analysis of grammatical relations
    - Entities that are grown, cooked, eaten, and digested are more likely to be food items.

- 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 $C = A A^T$ where $A$ is term-document matrix.
- $w_{i,j} = (\text{normalized})$ weighted count $(t_{i,d_j})$

With integer counts – what do you get for a boolean cooccurrence 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 polish</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</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 awl</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 similar
  - False negatives: Words deemed dissimilar that are similar
- Since terms are highly correlated anyway, expansion may not retrieve many additional documents.
Query Expansion: Summary

- Query expansion is often effective in increasing recall.
  - Not always with general thesauri
  - Fairly successful for subject-specific collections
- In most cases, precision is decreased, often significantly.
- Overall, not as useful as relevance feedback; may be as good as pseudo-relevance feedback
Pseudo Relevance Feedback

- Automatic local analysis
- Pseudo relevance feedback attempts to automate the manual part of relevance feedback.
- Retrieve an initial set of relevant documents.
- Assume that top $m$ ranked documents are relevant.
- Do relevance feedback
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 user or query specific.
- This is the general area of clickstream mining
Resources


Resources

