Information Retrieval

Weike Pan

The slides are **adapted from those provided by Prof. Hinrich Schütze** at University of Munich (http://www.cis.lmu.de/~hs/teach/14s/ir/).

How can we improve recall in search?

- As an example consider query q: [aircraft] … and document d containing “plane”, but not containing “aircraft”
  - A simple IR system will not return d for q, even if d is the most relevant document for q

- We want to change this:
  - Return relevant documents even if there is no term match with the original query

Options for improving recall

- **Local**: Do a “local” on-demand analysis for a user query
  - Main local method: relevance feedback
- **Global**: Do a global analysis once (e.g., of collection) to produce thesaurus (同义词词典)
  - Use thesaurus for query expansion

Chapter 9 Relevance feedback & query expansion

- 9.1 Relevance feedback and pseudo relevance feedback
- 9.2 Global methods for query reformulation
- 9.3 References and further reading

Outline

- 9.1 Relevance feedback and pseudo relevance feedback
- 9.2 Global methods for query reformulation
- 9.3 References and further reading

9.1 Relevance feedback and pseudo relevance feedback

Relevance feedback: Basic idea

- The user issues a short and simple query.
- The search engine returns a set of documents.
- User marks some documents as relevant, some as nonrelevant.
- Search engine computes a new representation of the information need. Hope: better than the initial query.
- Search engine runs the new query and returns new results.

- We will use the term **ad hoc retrieval** to refer to regular retrieval without relevance feedback.

9.1 Relevance feedback and pseudo relevance feedback

A real example

- Initial query: [new space satellite applications]

	r		
+	1	0.539	NASA Hasn't Scrapped Imaging Spectrometer
+	2	0.533	NASA Scratches Environment Gear From Satellite Plan
	3	0.528	Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
	4	0.526	A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
	5	0.525	Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
	6	0.524	Report Provides Support for the Critics Of Using Big Satellites to Study Climate
	7	0.516	Arianespace Receives Satellite Launch Pact From Telesat Canada
+	8	0.509	Telecommunications Tale of Two Companies

- User then marks relevant documents with "+"

9.1 Relevance feedback and pseudo relevance feedback

Expanded query after relevance feedback

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

- Different from the original query: [new space satellite applications]

9.1 Relevance feedback and pseudo relevance feedback

Results for expanded query (old ranks in parentheses/括号)

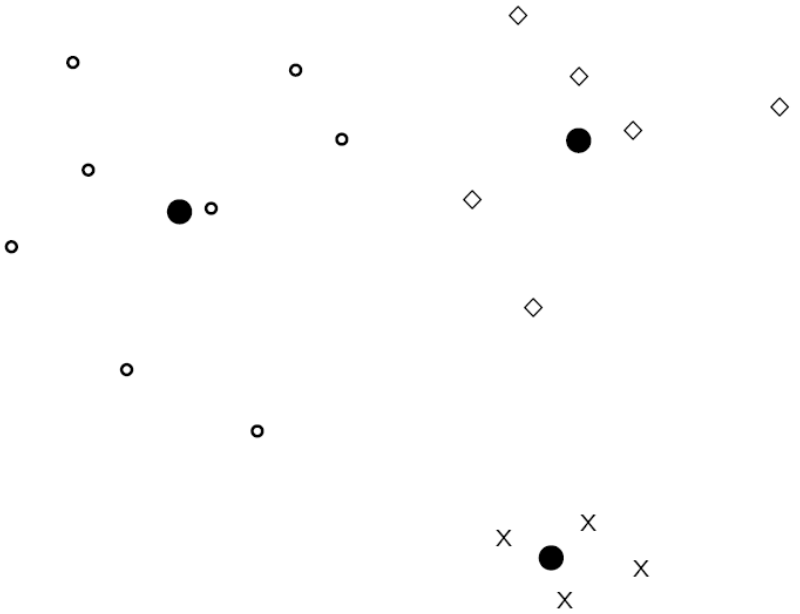
	r		
*	1 (2)	0.513	NASA Scratches Environment Gear From Satellite Plan
*	2 (1)	0.500	NASA Hasn't Scrapped Imaging Spectrometer
	3	0.493	When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
	4	0.493	NASA Uses 'Warm' Superconductors For Fast Circuit
*	5 (8)	0.492	Telecommunications Tale of Two Companies
	6	0.491	Soviets May Adapt Parts of SS-20 Missile For Commercial Use
	7	0.490	Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
	8	0.490	Rescue of Satellite By Space Agency To Cost $90 Million

9.1 Relevance feedback and pseudo relevance feedback

Key concept for relevance feedback: Centroid

- The centroid is the center of mass of a set of points.
- We represent the documents as points in a high-dimensional space.
- Thus: we can compute centroids of documents.

$$\vec{\mu}(D) = \frac{1}{|D|} \sum_{d \in D} \vec{v}(d)$$



9.1 Relevance feedback and pseudo relevance feedback

Rocchio algorithm

- The Rocchio algorithm implements relevance feedback in the vector space model.

Rocchio chooses the query $\vec{q}_{opt}$ that maximizes

$$\vec{q}_{opt} = \arg\max_{\vec{q}} [\text{sim}(\vec{q}, \mu(D_r)) - \text{sim}(\vec{q}, \mu(D_{nr}))]$$

D_r: set of relevant docs; D_{nr}: set of nonrelevant docs

- Separates relevant and nonrelevant docs maximally.

9.1 Relevance feedback and pseudo relevance feedback

Rocchio 1971 algorithm (SMART implementation)

$$
\begin{aligned}
\vec{q}_m &= \alpha\vec{q}_0 + \beta\mu(D_r) - \gamma\mu(D_{nr}) \\
&= \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j
\end{aligned}
$$

q_m: modified query vector; q_0: original query vector; D_r and D_{nr}: sets of known relevant and nonrelevant documents respectively; α, β, and γ: weights

- New query moves towards relevant documents and away from nonrelevant documents.

- Set negative term weights to 0, because "negative weight" for a term doesn't make sense in the vector space model.

9.1 Relevance feedback and pseudo relevance feedback

Positive vs. negative relevance feedback

- Positive feedback is more valuable than negative feedback.

- For example, set $\beta = 0.75$, $\gamma = 0.25$ to give higher weight to positive feedback.

- Many systems only allow positive feedback.

9.1 Relevance feedback and pseudo relevance feedback

Relevance feedback: Assumptions

- When can relevance feedback enhance recall?
- **Assumption A1**: The user knows the terms in the collection well enough for an initial query.
- **Assumption A2**: Relevant documents contain similar terms (so I can "hop" from one relevant document to a different one when giving relevance feedback).

9.1 Relevance feedback and pseudo relevance feedback

Violation of A1

- **Assumption A1**: The user knows the terms in the collection well enough for an initial query.

- Violation: Mismatch of searcher’s vocabulary and collection vocabulary, e.g., cosmonaut (宇航员) / astronaut (宇航员)

9.1 Relevance feedback and pseudo relevance feedback

Violation of A2

- **Assumption A2**: Relevant documents are similar.

- Example for violation: [contradictory (矛盾的) government policies]
- Several unrelated “prototypes”
  - Subsidies (补贴) for tobacco farmers vs. anti-smoking campaigns
  - Aid for developing countries vs. high tariffs (关税) on imports from developing countries

- Relevance feedback on tobacco docs will not help with finding docs on developing countries.

9.1 Relevance feedback and pseudo relevance feedback

Relevance feedback: Evaluation (1/3)

- Pick an evaluation measure, e.g., precision in top 10: P@10
- Compute P@10 for original query q0
- Compute P@10 for modified relevance feedback query q1
- In most cases: q1 is spectacularly (令人吃惊地) better than q0

- Is this a fair evaluation?

9.1 Relevance feedback and pseudo relevance feedback

Relevance feedback: Evaluation (2/3)

- Fair evaluation must be on “residual” collection: docs not yet judged by user.
- Studies have shown that relevance feedback is successful when evaluated this way.

- Empirically, one round of relevance feedback is often very useful. Two rounds are marginally useful.

9.1 Relevance feedback and pseudo relevance feedback

Relevance feedback: Evaluation (3/3)

- 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.

9.1 Relevance feedback and pseudo relevance feedback

Relevance feedback: Problems

- Relevance feedback is expensive
  - Relevance feedback creates long modified queries
  - Long queries are expensive to process

- Users are reluctant to provide explicit feedback.

- It’s often hard to understand why a particular document was retrieved after applying relevance feedback.

9.1 Relevance feedback and pseudo relevance feedback

Pseudo-relevance feedback

- Pseudo-relevance feedback automates the "manual" part of true relevance feedback.
- Pseudo-relevance feedback algorithm:
  - Step 1: Retrieve a ranked list of hits for the user's query
  - Step 2: Assume that the top k documents are relevant.
  - Step 3: Do relevance feedback (e.g., Rocchio algorithm)

- It works very well on average. But can go horribly wrong for some queries because of query drift.

Outline

- 9.1 Relevance feedback and pseudo relevance feedback
- 9.2 Global methods for query reformulation
- 9.3 References and further reading

9.2 Global methods for query reformulation

Types of user feedback

- User gives feedback on documents.
  - More common in relevance feedback
- User gives feedback on words or phrases.
  - More common in query expansion

9.2 Global methods for query reformulation

Query expansion

- Query expansion is another method for increasing recall.
- We use “global query expansion” to refer to “global methods for query reformulation”.

- In global query expansion, the query is modified based on some global resource, i.e., a resource that is not query-dependent.

- Main information we use: (near-)synonymy

9.2 Global methods for query reformulation

“Global” resources used for query expansion

- A publication or database that collects (near-)synonyms is called a thesaurus (同义词词典).
  - Manual thesaurus (maintained by editors, e.g., PubMed)
  - Automatically derived thesaurus (e.g., based on co-occurrence statistics)
  - Query-equivalence based on query log mining

9.2 Global methods for query reformulation

Thesaurus-based query expansion

- For each term t in the query, expand the query with words the thesaurus lists as semantically related with t.
- Example: hospital → medical
- Generally increases recall
- May significantly decrease precision, particularly with ambiguous terms
  - E.g., interest rate (利率风险) → interest rate fascinate (利率魅力)

- Widely used in specialized search engines for science and engineering
- It’s very expensive to create a manual thesaurus and to maintain it over time

9.2 Global methods for query reformulation

Automatic thesaurus generation

- Generate a thesaurus by analyzing the distribution of words in documents

- Definition 1: Two words are similar if they **co-occur with similar words**.
  - E.g., “car” ≈ “motorcycle” because both occur with “road”, “gas” and “license”, so they must be similar.

- Definition 2: Two words are similar if they **occur in a given grammatical relation** with the same words.
  - E.g., You can harvest, peel (削皮) and eat apples and pears, so apples and pears must be similar.

- Co-occurrence is more robust, grammatical relations are more accurate.

9.2 Global methods for query reformulation

Co-occurrence based thesaurus: Examples

Word	Nearest neighbors
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs
makeup	repellent lotion glossy sunscreen skin gel
mediating	reconciliation negotiate case conciliation
keeping	hoping bring wiping could some would
lithographs	drawings Picasso Dali sculptures Gauguin
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate

9.2 Global methods for query reformulation

Query expansion at search engines

- Main source of query expansion at search engines: query logs
- Example 1: After issuing the query [herbs] (药草), users frequently search for [herbal remedies] (草药疗法). → "herbal remedies" is a potential expansion of "herb".
- Example 2: Users searching for [flower pix] frequently click on the URL photobucket.com/flower. Users searching for [flower clipart] frequently click on the same URL. → "flower clipart" (花冠) and "flower pix" (花瓣) are potential expansions of each other.

Summary

- 9.1 Relevance feedback and pseudo relevance feedback
- 9.2 Global methods for query reformulation
- 9.3 References and further reading