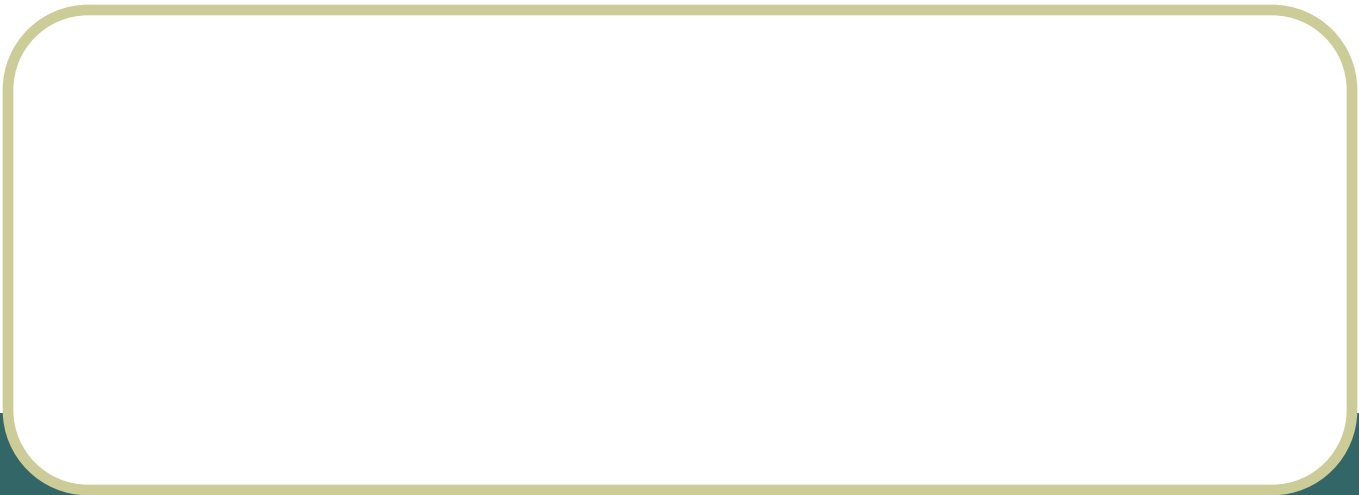


Relevance Feedback, Query Expansion and Pseudo Relevance Feedback



In this lecture

- Improving search 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

The screenshot shows a Mozilla Firefox browser window with the title 'regan - Google Search - Mozilla Firefox'. The address bar displays the search URL: <http://www.google.com/search?q=regan&btnG=Search&hs=UnN&hl=en&lr=&client=firefox-a>. The search bar contains the text 'regan'. The page shows search results for 'regan', with a total of 13,200,000 results found in 0.07 seconds. The results are categorized under 'Web'.

Web Results 1 - 10 of about 13,200,000 for **regan**. (0.07 seconds)

Brian Regan: The Official Site
Brian **Regan** is one of the best comedians performing today. His comedy, big enough for everyone, sharp enough for you, keeps audiences coming back time and ...
www.brianregan.com/ - 13k - [Cached](#) - [Similar pages](#)

reganmusic.com
Color.
www.reganmusic.com/ - 2k - [Cached](#) - [Similar pages](#)

Regan Nursery Bare Root Roses
We offer over 1100 bareroot roses from one of the largest selections of Grade 1 bareroot roses in the US, including David Austin roses, Hybrid Tea roses, ...
www.regannursery.com/ - 14k - [Cached](#) - [Similar pages](#)

See results for: **ronald reagan**

Biography of Ronald Reagan
Biography of **Ronald Reagan**, the fortieth President of the United States (1981-1989).
www.whitehouse.gov/history/presidents/r40.html

Ronald Reagan - Wikipedia, the free encyclopedia
Ronald Reagan visiting Nancy **Reagan** on the set of her movie Donovan's Brain, 1953.
... **Ronald Reagan** on the cover of TIME as "Man of the Year," 1980 ...
en.wikipedia.org/wiki/Ronald_Reagan

RonaldReagan.com
Provides in-depth biographical information, message boards, video clips, and transcripts of historic speeches.
www.ronaldreagan.com/

Regan Family Genealogy Forum
Margaret Ann **Regan** 1878 Providence RI godmother - Barbara Glassel 3/05/06. James **Regan** Clarendon School Canton OH 1930s - Gregory Winters 3/05/06 ...
genforum.genealogy.com/regan/ - 28k - [Cached](#) - [Similar pages](#)

Sponsored Links

Regan
Looking for **Regan**?
Find exactly what you want today.
www.eBay.com

Ronald Reagan Shirt
100% Cotton Button-Down Shirts
Hand Printed with Ron Reagan \$35
ILoveReagan.com

Find: Find Next Find Previous Highlight all Match case

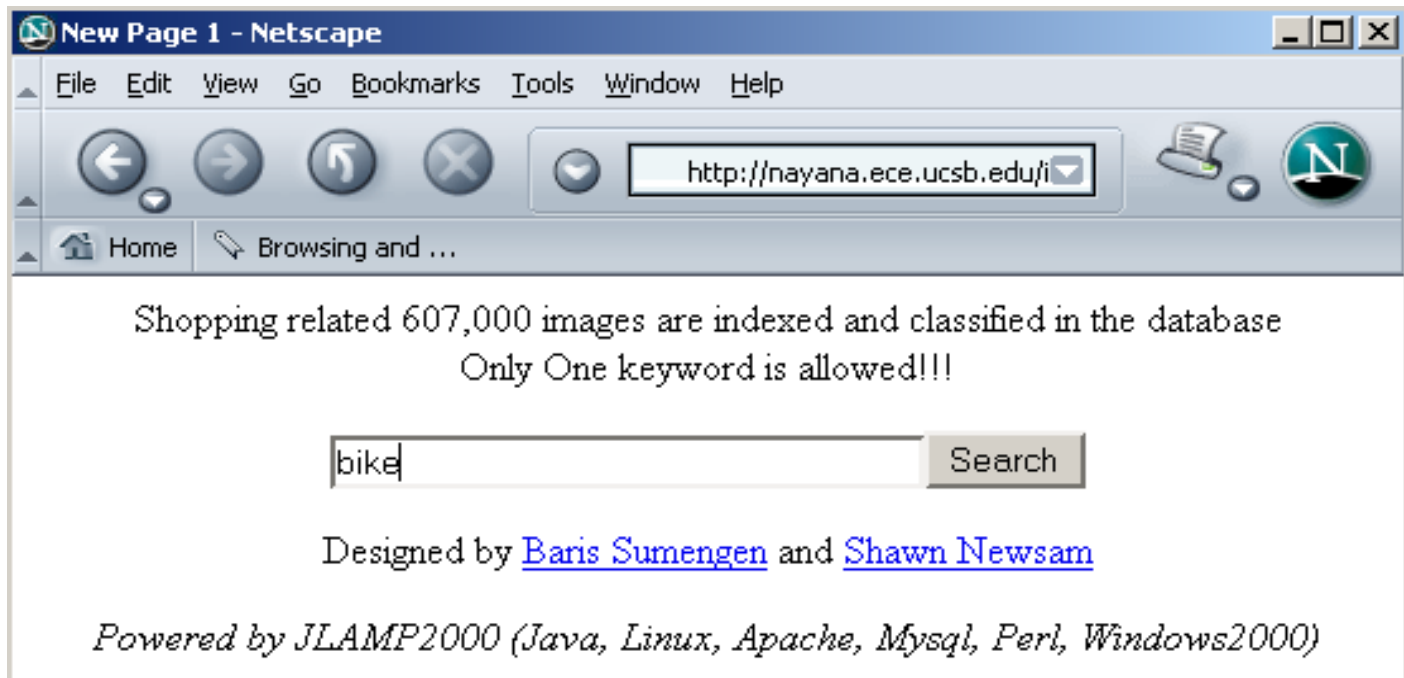
Done

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.html>



Results for Initial Query

[Browse](#)
[Search](#)
[Prev](#)
[Next](#)
[Random](#)


(144473, 16458)

0.0

0.0

0.0



(144457, 252140)

0.0

0.0

0.0



(144456, 262857)

0.0

0.0

0.0



(144456, 262863)

0.0

0.0

0.0



(144457, 252134)

0.0

0.0

0.0



(144483, 265154)

0.0

0.0

0.0



(144483, 264644)

0.0

0.0

0.0



(144483, 265153)

0.0

0.0

0.0



(144518, 257752)

0.0

0.0

0.0



(144538, 525937)

0.0

0.0

0.0



(144456, 249611)

0.0

0.0

0.0



(144456, 250064)

0.0

0.0

0.0

Relevance Feedback

Browse

Search

Prev

Next

Random



(144473, 16458)

0.0

0.0

0.0



(144457, 252140)

0.0

0.0

0.0



(144456, 262857)

0.0

0.0

0.0



(144456, 262863)

0.0

0.0

0.0



(144457, 252134)

0.0

0.0

0.0



(144483, 265154)

0.0

0.0

0.0



(144483, 264644)

0.0

0.0

0.0



(144483, 265153)

0.0

0.0

0.0



(144518, 257752)

0.0

0.0

0.0



(144538, 525937)

0.0

0.0

0.0



(144456, 249611)

0.0

0.0

0.0



(144456, 250064)

0.0

0.0

0.0

Results after Relevance Feedback

[Browse](#)
[Search](#)
[Prev](#)
[Next](#)
[Random](#)


(144538, 523493)
0.54182
0.231944
0.309876



(144538, 523835)
0.56319296
0.267304
0.295889



(144538, 523529)
0.584279
0.280881
0.303398



(144456, 253569)
0.64501
0.351395
0.293615



(144456, 253568)
0.650275
0.411745
0.23853



(144538, 523799)
0.66709197
0.358033
0.309059



(144473, 16249)
0.6721
0.393922
0.278178



(144456, 249634)
0.675018
0.4639
0.211118



(144456, 253693)
0.676901
0.47645
0.200451



(144473, 16328)
0.700339
0.309002
0.391337



(144483, 265264)
0.70170796
0.36176
0.339948



(144478, 512410)
0.70297
0.469111
0.233859

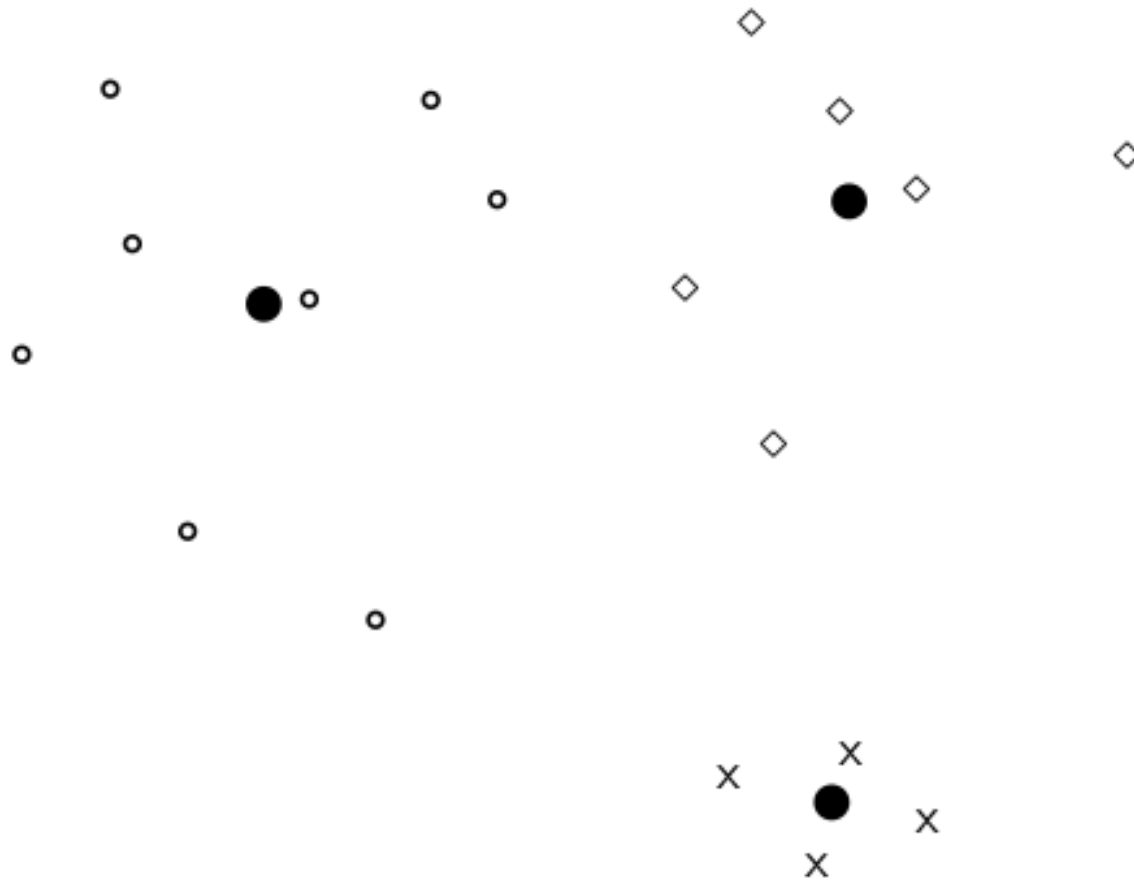
Key concept for relevance feedback: Centroid

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

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

where D is a set of documents and $\vec{v}(d) = \vec{d}$ is the vector we use to represent document d .

Centroid: Example



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

- Intent: \vec{q}_{opt} is the vector that separates relevant and nonrelevant docs maximally.
- Making some additional assumptions, we can rewrite \vec{q}_{opt} as:

$$\vec{q}_{opt} = \mu(D_r) + [\mu(D_r) - \mu(D_{nr})]$$

Rocchio' algorithm

- The optimal query vector is:

$$\begin{aligned}\vec{q}_{opt} &= \mu(D_r) + [\mu(D_r) - \mu(D_{nr})] \\ &= \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j + \left[\frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j \right]\end{aligned}$$

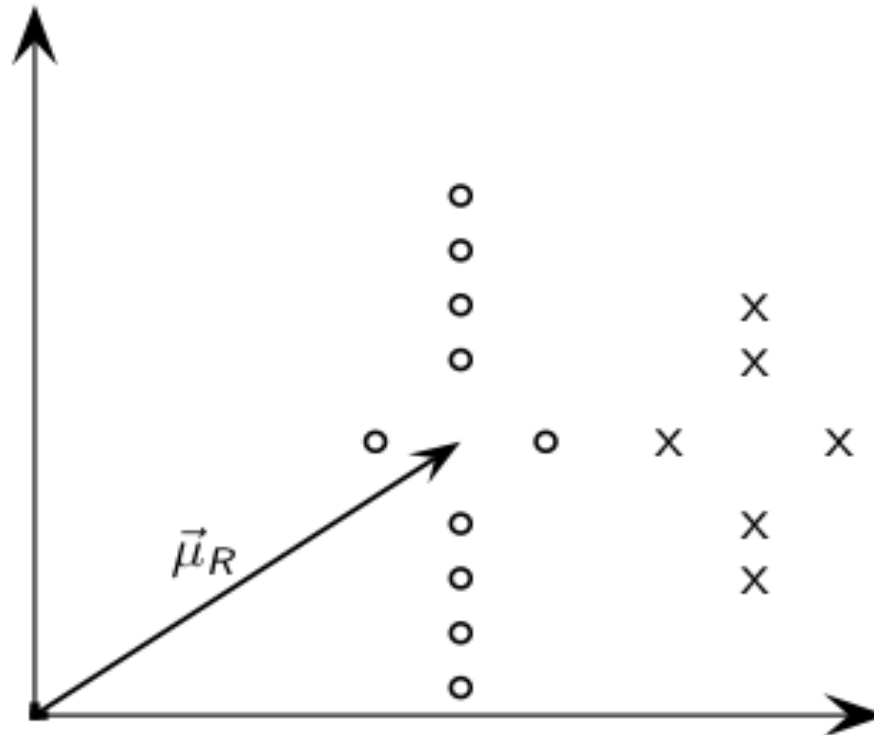
- We move the centroid of the relevant documents by the difference between the two centroids.

Exercise: Compute Rocchio' vector



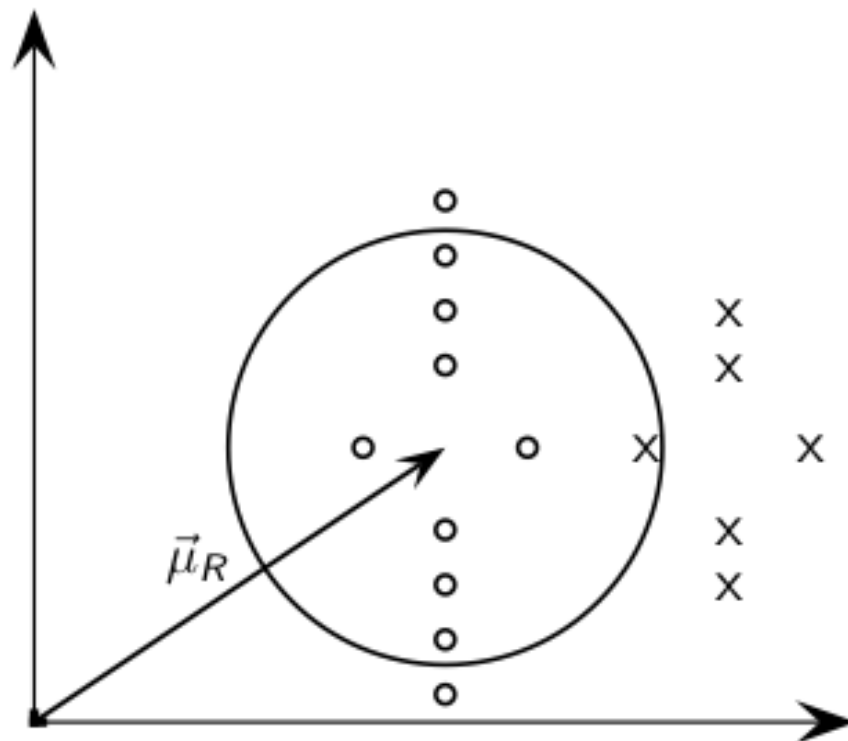
circles: relevant documents, Xs: nonrelevant documents

Rocchio' illustrated



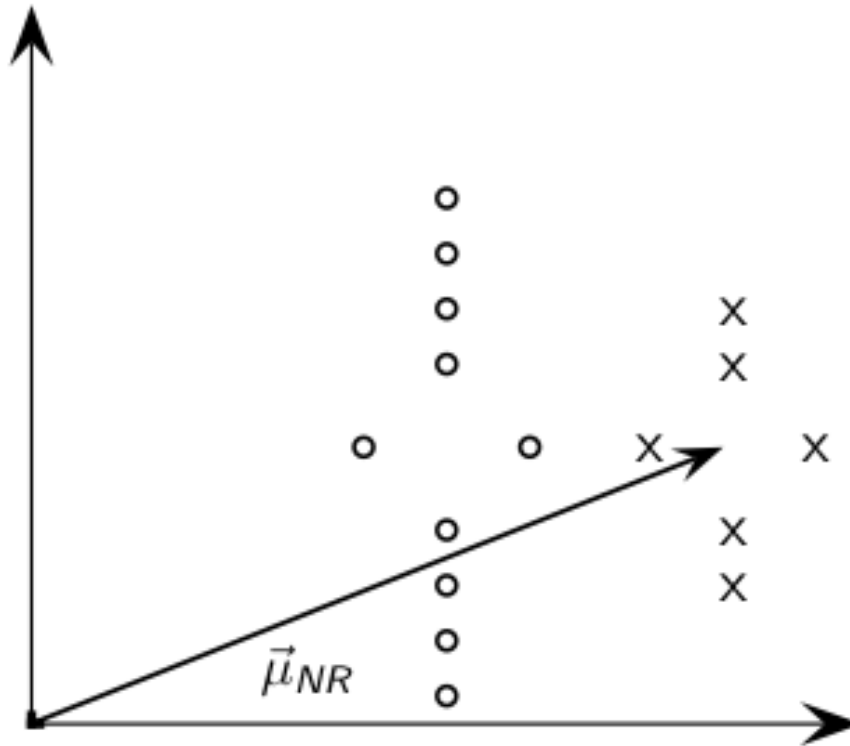
$\vec{\mu}_R$: centroid of relevant documents

Rocchio' illustrated



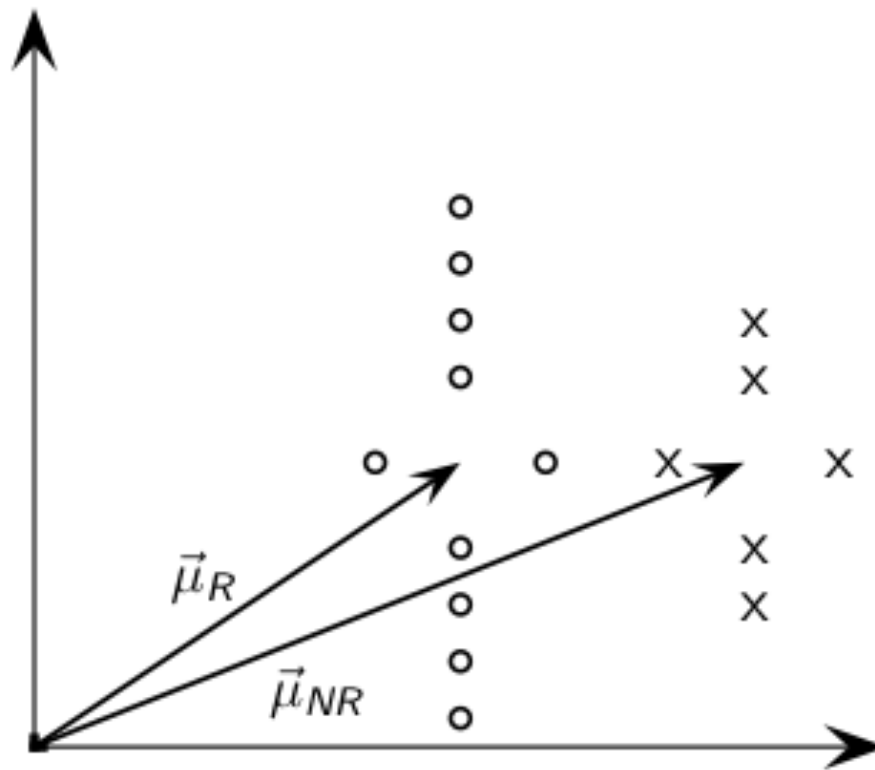
$\vec{\mu}_R$ does not separate relevant / nonrelevant.

Rocchio' illustrated

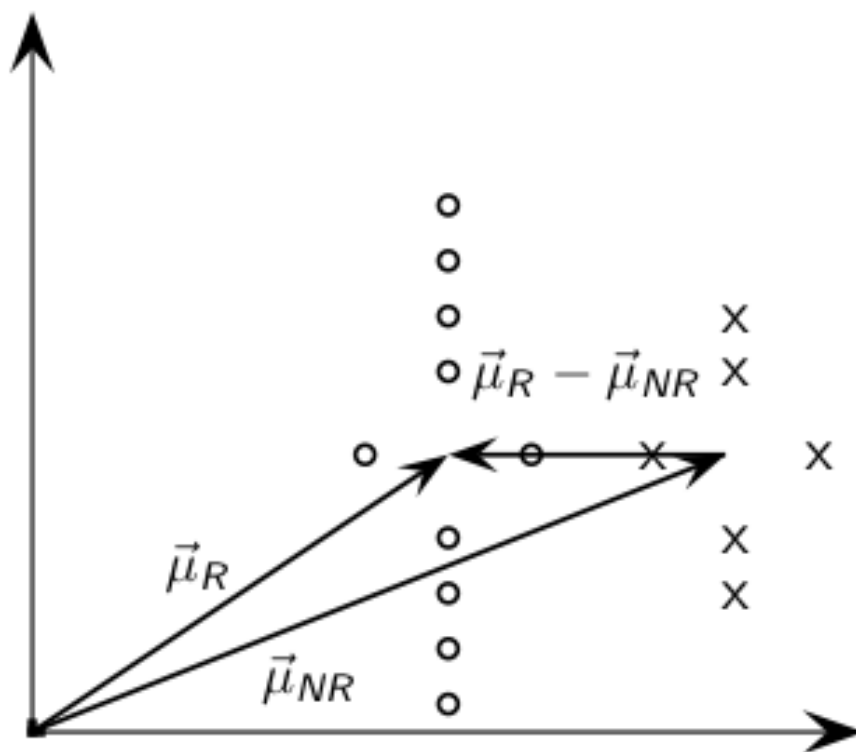


$\vec{\mu}_{NR}$: centroid of nonrelevant documents.

Rocchio' illustrated

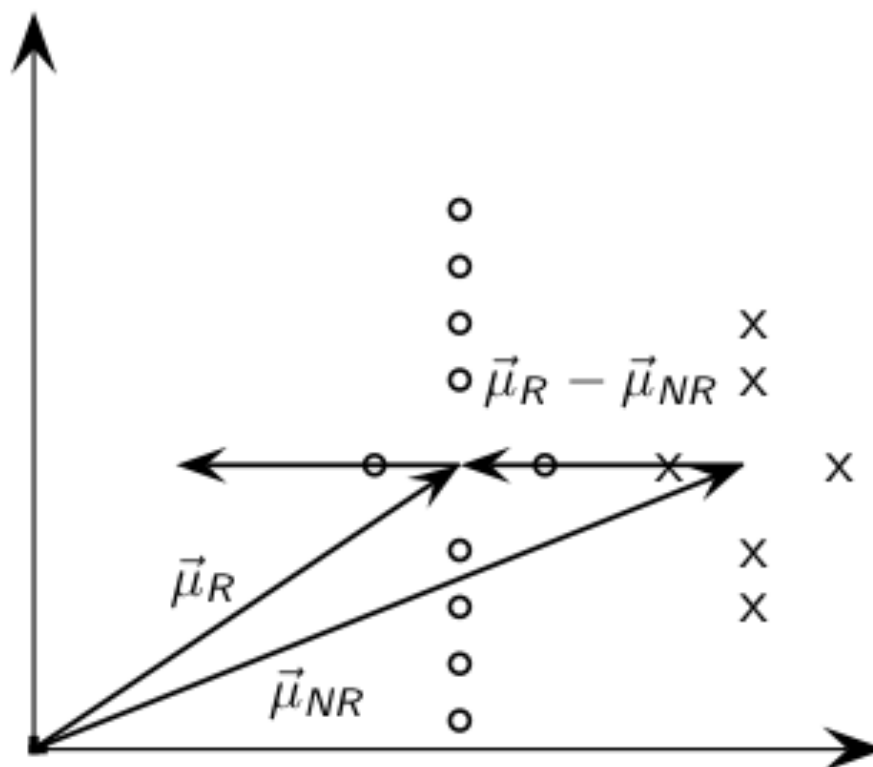


Rocchio' illustrated



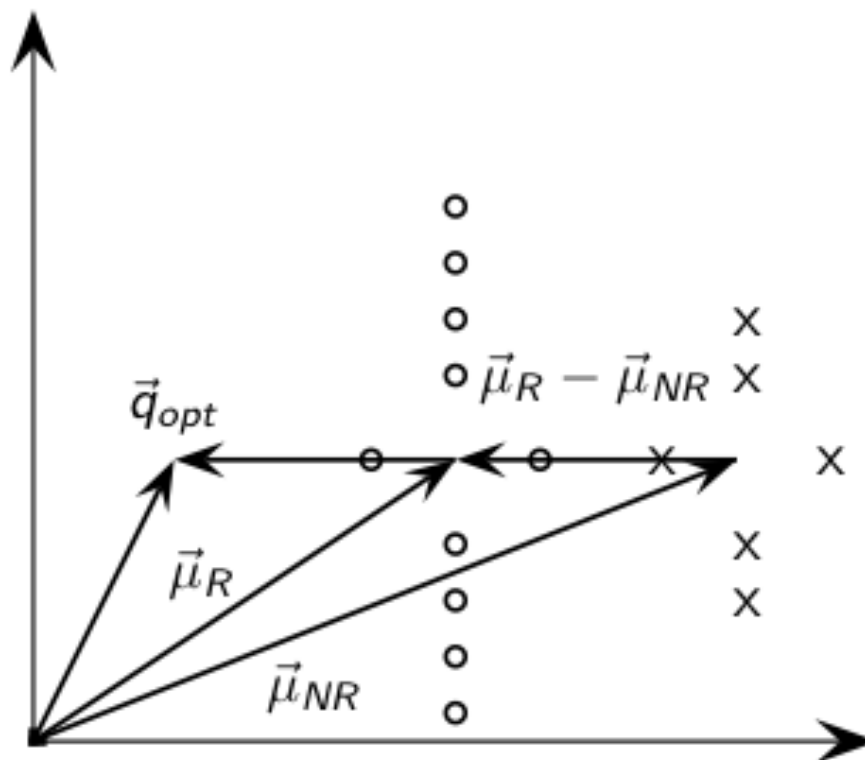
$\vec{\mu}_R - \vec{\mu}_{NR}$: difference vector

Rocchio' illustrated



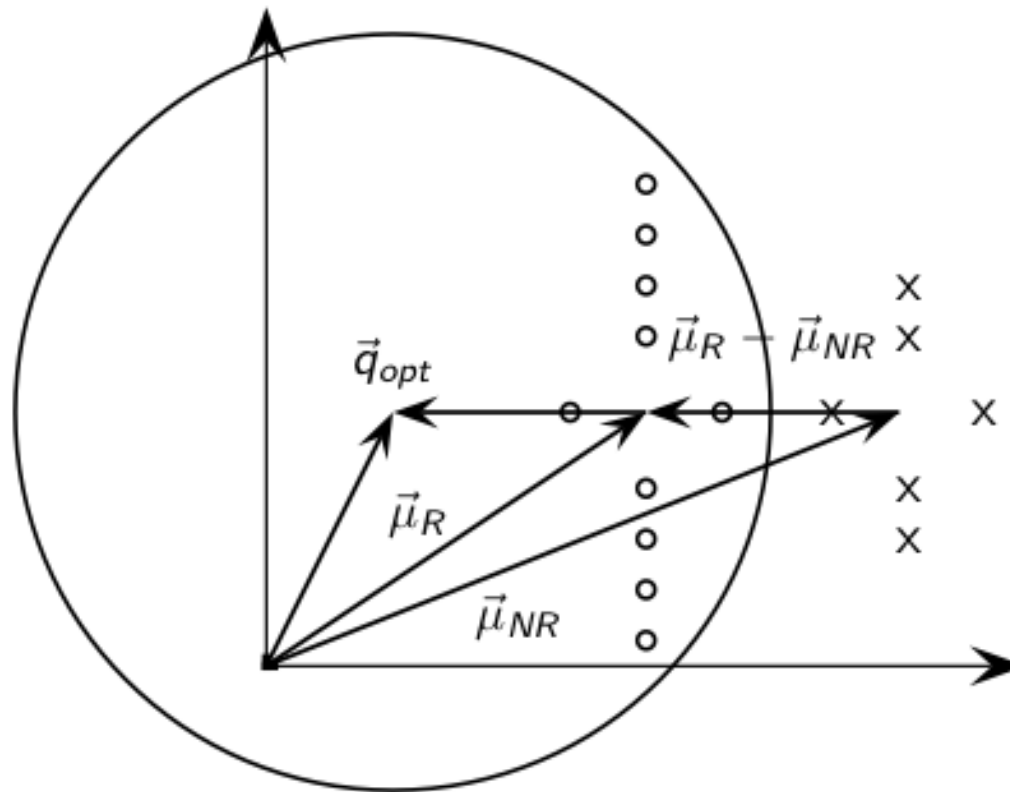
Add difference vector to $\vec{\mu}_R$...

Rocchio' illustrated



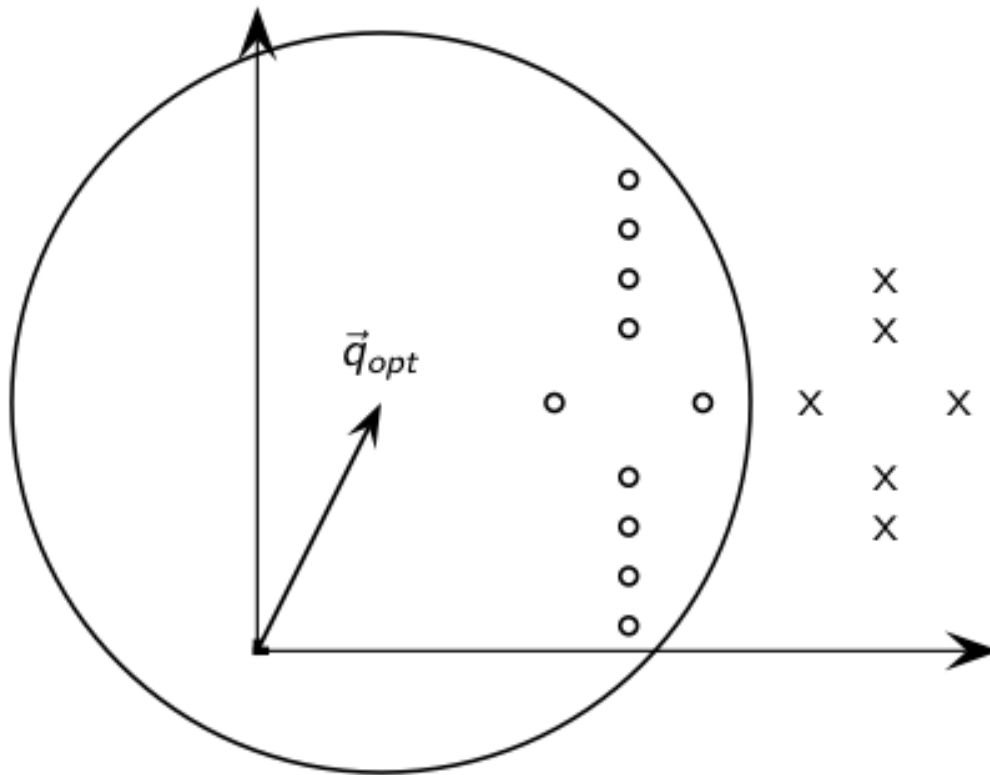
... to get \vec{q}_{opt}

Rocchio' illustrated



\vec{q}_{opt} separates relevant / nonrelevant perfectly.

Rocchio' illustrated



\vec{q}_{opt} separates relevant / nonrelevant perfectly.

Rocchio' and Rocchio (SMART implementation)

- Here the name Rocchio' for the theoretically better motivated original version of Rocchio.
- The implementation that is actually used in most cases is the SMART implementation – we use the name Rocchio (without prime) for that.

Rocchio 1971 algorithm (SMART)

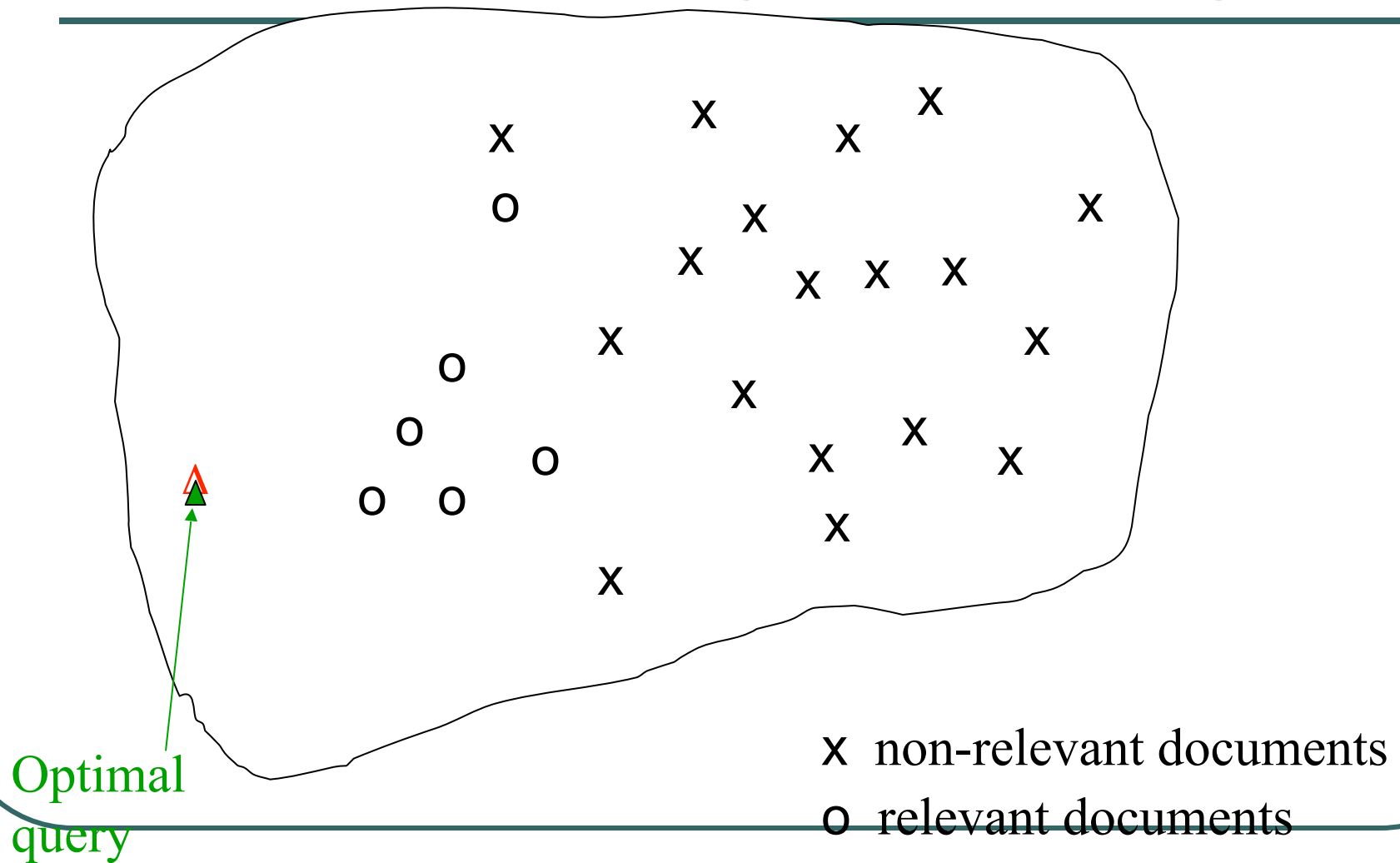
Used in practice:

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

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

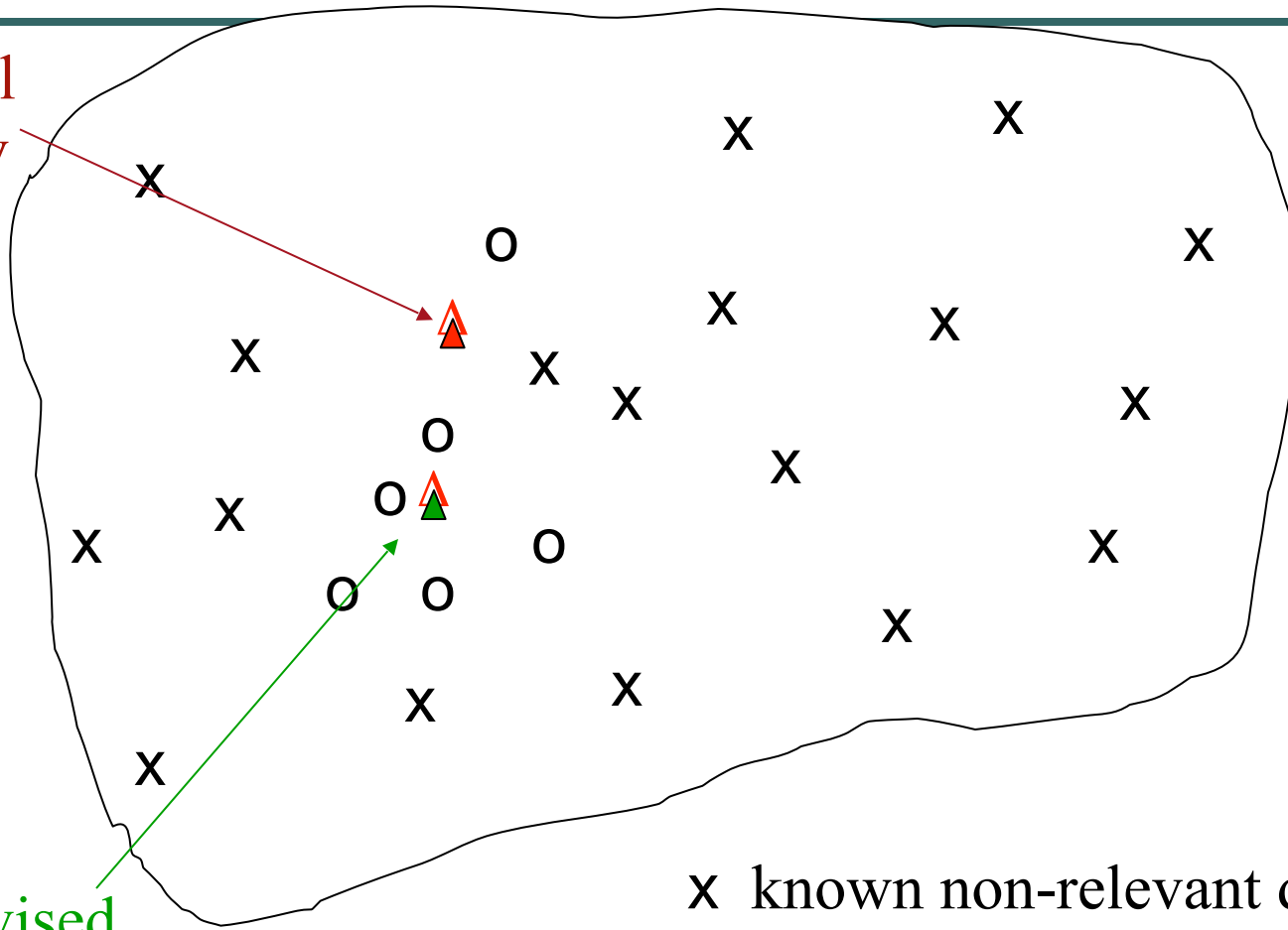
- New query moves towards relevant documents and away from nonrelevant documents.
- Tradeoff α vs. β/γ : If we have a lot of judged documents, we want a higher β/γ .
- Set negative term weights to 0.
- “Negative weight” for a term doesn’t make sense in the vector space model.

The Theoretically Best Query



Relevance feedback on initial query

Initial
query



Revised
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 ($\gamma=0$).



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
 - Mismatch of searcher's vocabulary vs. collection vocabulary
 - Cosmonaut/astronaut

Violation of A2

- There are several relevance prototypes.
- Examples:
 - Burma/Myanmar
 - Contradictory government policies
- 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

Evaluation of relevance feedback strategies

• Use q_0 and compute precision and recall graph

• Use q_m and compute precision recall graph

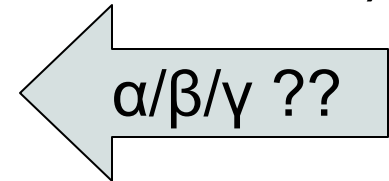
- Assess on all documents in the collection
 - Spectacular improvements, but ... it's cheating!
 - Partly due to known relevant documents ranked higher
 - Must evaluate with respect to documents not seen by user
- Use documents in residual collection (set of documents minus those assessed relevant)
 - 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

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)

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 (≥ 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.
- Other types of interactive retrieval may improve relevance by as much with less work.

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

YOU ARE HERE > [Home](#) > [My InfoSpace](#) > [Meta-Search](#) > Web Search Results

Web Search Results

Your Search

Select: ▼

☐ [Yellow Pages](#) ☐ [White Pages](#) ☐ [Classifieds](#)

Are you looking for?

[Jacksonville Jaguars](#)

[Jaguar Car](#)

[Black Jaguar](#)

[Jaguar Xk8](#)

[Wild Jaguars](#)

[Jaguar](#)

[Jaguar Accessories](#)


[Jaguar Automobile](#)


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)
 - Analysis of documents in result set

Controlled Vocabulary





National Library of Medicine 

PubMedNucleotideProteinGenomeStructurePopSetTaxonomy

SearchPubMed ▾ forcancerGoClear

LimitsPreview/IndexHistoryClipboardDetails

About Entrez

Text Version

Entrez PubMed

Overview

Help | FAQ

Tutorial

New/Noteworthy

E-Utilities

PubMed Services

Journals Database

MeSH Browser

Single Citation

Metabrowser

PubMed Query:

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

SearchURL

Thesaurus-based Query Expansion

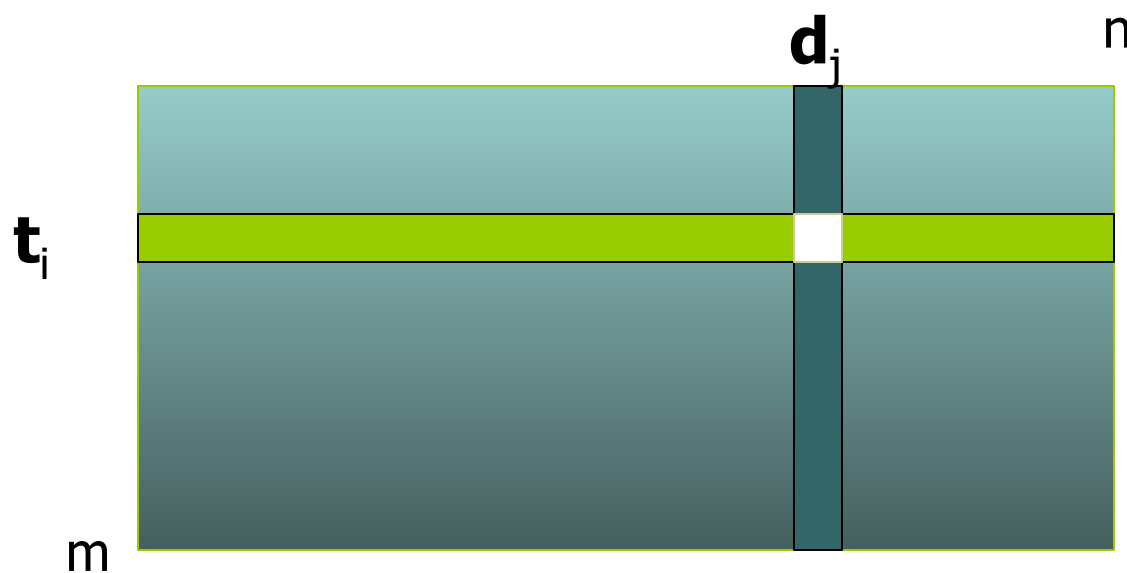
- This doesn't require user input
- For each term, t , in a query, expand the query with synonyms and related words of t from the thesaurus
 - feline → feline cat
- May weight added terms less than original query terms.
- Generally increases recall.
- Widely used in many science/engineering fields
- May significantly decrease precision, particularly with ambiguous terms.
 - “interest rate” → “interest rate fascinate evaluate”
- There is a high cost of manually producing a thesaurus
 - 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 $C = AA^T$ where A is term-document matrix.
- $w_{i,j}$ = (normalized) weighted count (t_i, d_j)



Automatic Thesaurus Generation

Example

word	ten nearest neighbors
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed slig
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs gazed
Makeup	repellent lotion glossy sunscreen Skin gel p
mediating	reconciliation negotiate cease conciliation p
keeping	hoping bring wiping could some would othe
lithographs	drawings Picasso Dali sculptures Gauguin l
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate awl

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.

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

- Mostly works (perhaps better than global analysis!)
 - Found to improve performance in TREC ad-hoc task
 - Danger of query drift
- pseudo relevance feedback is the relevance feedback without user intervention
- Automatic local analysis
- Pseudo relevance feedback attempts to automate the manual part of relevance feedback.

Pseudo Relevance Feedback

- Retrieve an initial set of relevant documents.
- *Assume* that top m ranked documents are relevant.
- Do relevance feedback

Two approaches

There are two well-known models for applying relevance feedback:

- (i) Ide's method [69]: This method adds all the terms of relevant documents of the set provided for relevance feedback and removes the terms of first irrelevant document in the set. Modified query vector is constructed as:

$$q_{new} = q_{old} + \sum_{d \in R} d_i - s$$

where q_{old} is the original query vector, q_{new} is the new query, d_i is the vector of relevant document R and s is the vector of the first irrelevant document in the feedback set.

-
- (ii) Rocchio's method [127]: Rocchio's method consists of moving the initial query vector toward the centroid of the relevant documents and away from the centroid of the non-relevant documents. It attempts to estimate "the optimal" user query through relevance feedback. This can be described by the following equation:

$$q_{new} = \alpha \times q_{old} + \beta \times \sum_{d_i \in R} d_i - \gamma \times \sum_{d_i \in I} d_i$$

Where, d_i is relevant or irrelevant document obtained by manual or automatic feedback during initial retrieval, I is the set of irrelevant documents and α , β and γ are coefficients. These coefficients are set by trial and error method.

$\alpha = 1$, $\beta = .75$ and $\gamma = .15$

-
- In pseudo-relevance feedback, the modified query vector is calculated by dropping the negative terms appearing in the Ide's and Rocchio's equation.

query drift

- Query expansion following retrieval feedback may degrade the performance if the top ranked documents retrieved during initial run are not relevant.
- Expanding query by adding terms from irrelevant documents, or adding terms from relevant documents that are not closely related to the query terms, will move the query representation away from what may be “optimal” query representation.
- This results in alteration of focus of the query i.e. ‘query drift’.

Pseudo relevance feedback: Cornell SMART at TREC 4

- Results show number of relevant documents out of top 100 for 50 queries (so out of 5000)
- Results contrast two length normalization schemes (L vs. I), and pseudo relevance feedback (PsRF) (done as adding 20 terms)

• Inc.Itc	3210
• Inc.Itc-PsRF	3634
• Lnu.Itu	3709
• Lnu.Itu-PsRF	4350

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