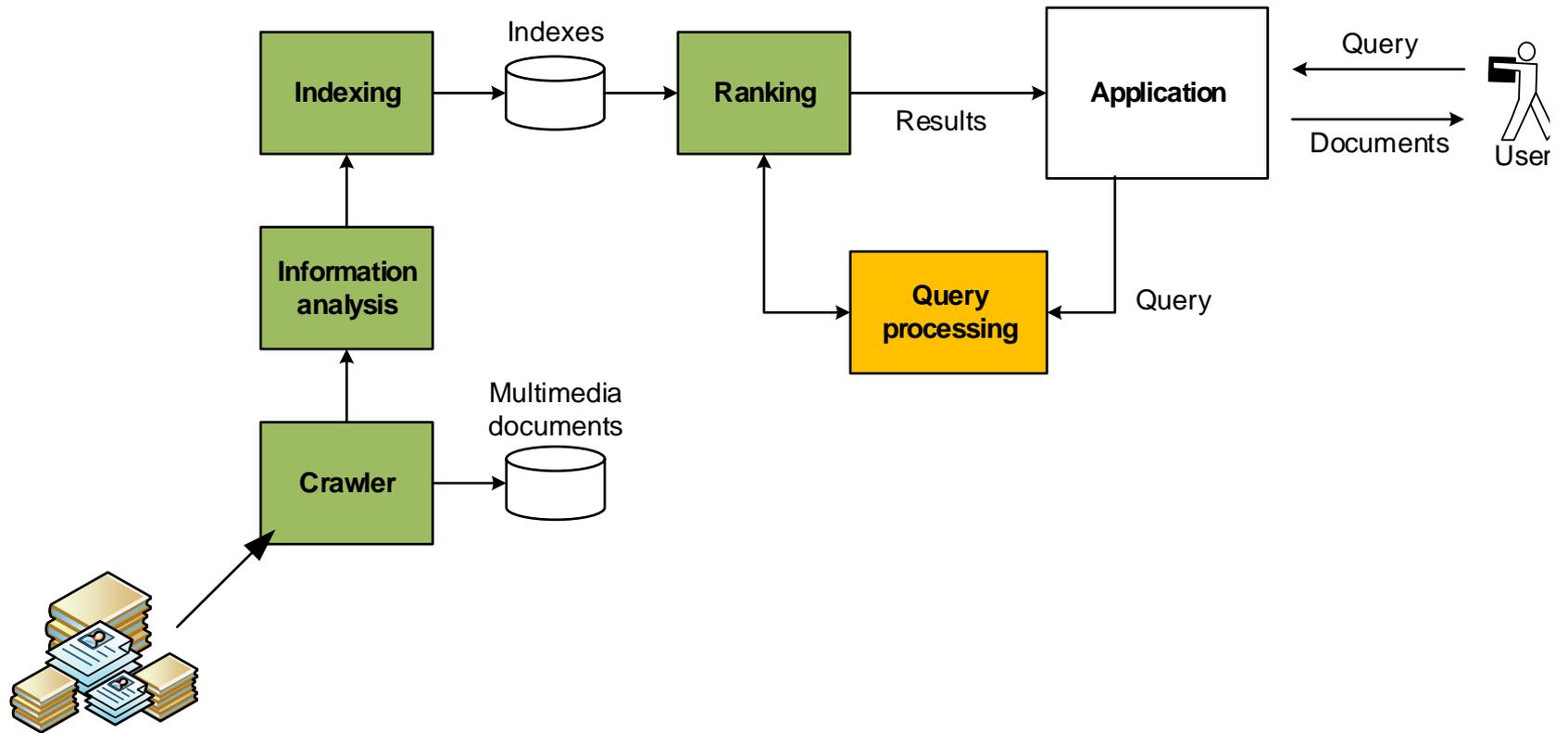


# Relevance-based Language Models

## Information Retrieval

# Overview



# Query



trump|

- trump
- trump **news**
- trump **cabinet**
- trump **tower**
- trump **executive orders**
- trump **memes**
- trump **impeachment**
- trump **russia**
- trumpet
- trump **latest**

Google Search    I'm Feeling Lucky

# Query assist

How can we revise the user query to improve search results?

The screenshot shows the Jitter search interface. The search bar contains the query "trump syria w". A dropdown menu is open, displaying suggestions: "wikileaks", "wall street", "warns", "white house", "washington post", "fake news", "new hampshire", and "#draintheswamp".

On the right, the "Top Tweets" section shows "About 6.3 million results (1.23 seconds)". The first tweet is from @warfeed, posted at 11:16 AM on 28 Sep 2016. It includes a bar chart showing the number of deaths in Aleppo, categorized by group and type of death.

Group	Civilians	Non-Civilians
Syrian Government	10	0
Russian Government	214	0
ISIS + Nusra	0	0
Opposition Armed Groups	0	0
Others	15	0

Below the chart, a quote states: "During one week in Aleppo, between September 21st and 26th, 2016, VDC documented 377 battle-related deaths, 369 are civilians and only 8 are non-civilians."

The second tweet is from @Skibabs, titled "Syria conflict: Assad hopes for 'anti-terror ally' in Trump: Syria's president says he hopes Donald Trump..." with a link to dlv.r.it/MgRz8L #Skibabs.

# Language Models

- Language Model given a document  $p(t|M_d)$ 
  - Computed from document statistics
- Language Model given a collection  $p(t|M_C)$ 
  - Computed from the collection statistics
- Language Model given a query  $p(t|Q)$ 
  - Computed from the top ranked documents
  - Based on a relevance estimation

# Relevance feedback



- Given the initial search results, the user marks some documents as important or non-important.
  - This information is used for a second search iteration where these examples are used to refine the results
- The characteristics of the positive examples are used to boost documents with similar characteristics
- The characteristics of the negative examples are used to penalize documents with similar characteristics

# Example: UX perspective

Results for initial query

Initial search results interface showing a grid of images and associated data. A mouse cursor is positioned over the 'BIKING' magazine cover. A 'User feedback' box is overlaid on the right side of the grid, containing a smaller version of the same grid and navigation buttons.

Navigation buttons: Browse, Search, Prev, Next, Random

User feedback: Browse, Search, Prev, Next, Random

(144473, 16458) 0.0 0.0 0.0	(144457, 252140) 0.0 0.0 0.0	(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, 262857) 0.0 0.0 0.0
(144483, 264644) 0.0 0.0 0.0	(144483, 265153) 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, 523493) 0.0 0.0 0.0

Results after Relevance Feedback

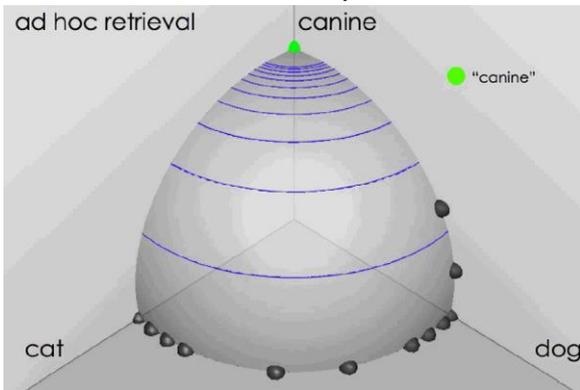
Relevance feedback results interface showing a refined grid of images and associated data. The grid is more focused on motorcycles and bicycles. Navigation buttons are present at the top.

Navigation buttons: Browse, Search, Prev, Next, Random

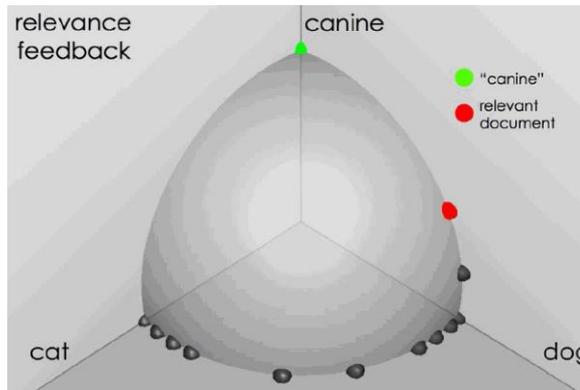
(144538, 523493) 0.54182 0.231944 0.309876	(144538, 523835) 0.56319296 0.267304 0.295889	(144538, 523529) 0.584279 0.280881 0.303398	(144456, 253569) 0.64501 0.351395 0.293615	(144456, 253568) 0.650275 0.411745 0.23853	(144538, 523799) 0.66709197 0.358033 0.309059
(144473, 16249) 0.6721 0.393922 0.278178	(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

# Example: geometric perspective

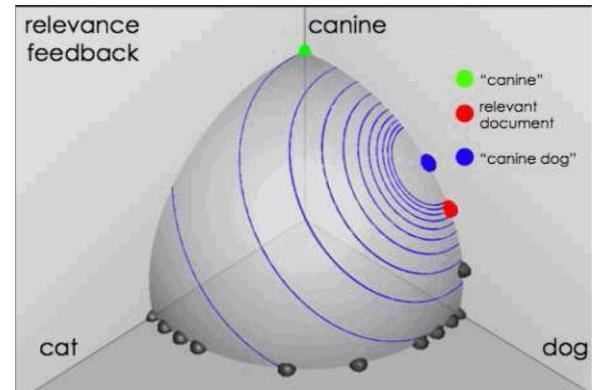
Results for Initial Query



User feedback



Results after Relevance Feedback



# Key concept: Centroid

Section 11

- The centroid is the center of mass of a set of points
  - Recall that we represent documents as points in a high-dimensional space
- The centroid of a set of documents  $C$  is defined as:

$$\bar{\mu}(C) = \frac{1}{|C|} \sum_{d \in C} \vec{d}$$

# Rocchio algorithm

- The Rocchio algorithm uses the vector space model to pick a relevance fed-back query
  - Rocchio seeks the query  $q_{opt}$  that maximizes

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

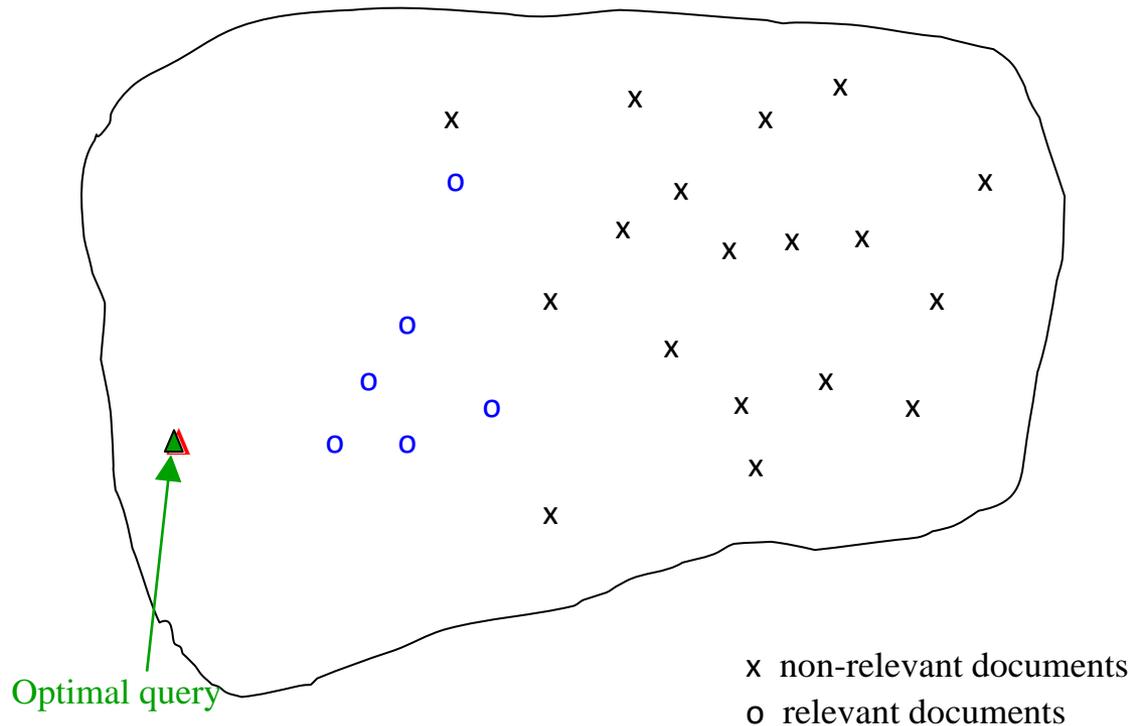
- Tries to separate documents marked as relevant and non-relevant

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\vec{d}_j \notin C_r} \vec{d}_j$$

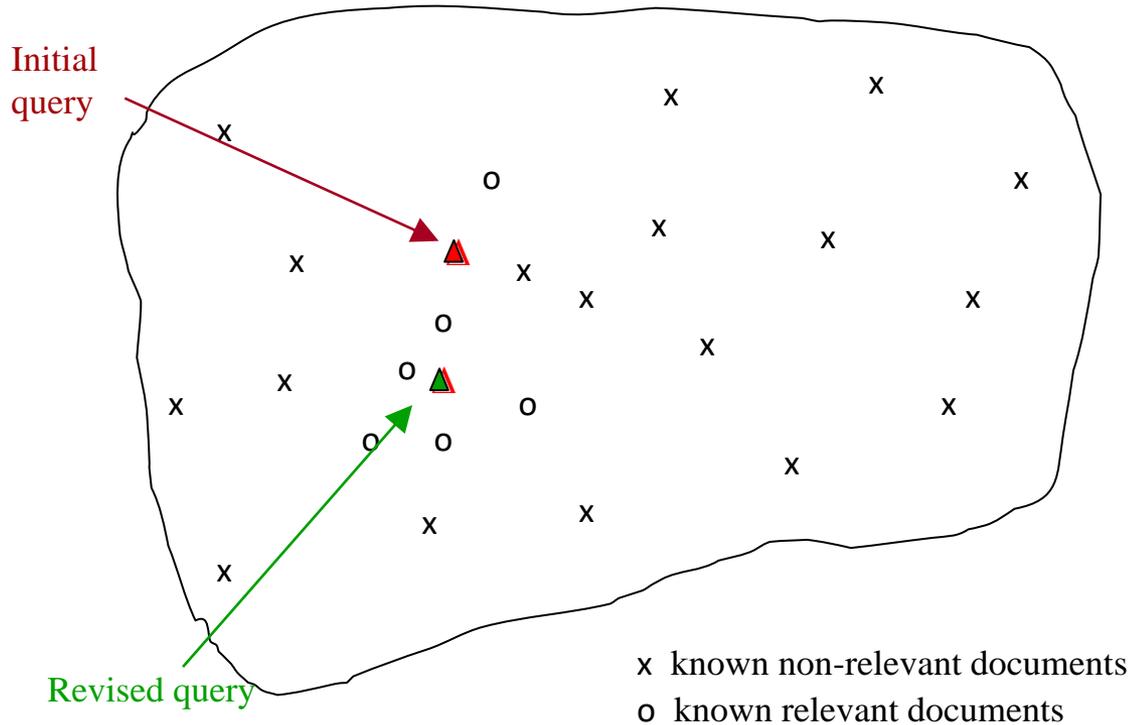
- Problem: we don't know the truly relevant docs

# The theoretically best query

Section 11



# Relevance feedback on initial query



# Rocchio 1971 Algorithm (SMART)

- Used in practice:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

- $D_r$  = 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)
- 
- The new query moves toward relevant documents and away from irrelevant documents

# Subtleties to note

- 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)

# Google A/B testing of relevance feedback



The image shows a screenshot of a Google search results page. At the top left is the Google logo. To its right is a search input field containing the text "olympic medal summary". Further right is a "Search" button and a link for "Advanced Search Preferences". Below the search bar is a navigation bar with "Web" selected and "News" as an option. On the right side of this bar, it says "Results 1 - 10 of ...".

The first search result is titled "Overall **Medal** Standings - The official website of the BEIJING 2008 ...". The description reads: "The Official Website of the Beijing 2008 **Olympic** Games August 8-24, 2008. COMPETITION INFORMATION. Schedules & Results - **Medals**; Athletes & Teams ...". The URL is "results.beijing2008.cn/WRM/ENG/INF/GL/95A/GL0000000.shtml" with a size of 54k. It includes "Cached" and "Similar pages" links.

The second result is titled "Olympics — Infoplease.com". The description reads: "Summary of gold medal winners for Summer and Winter **Olympic** Games. Summer **Olympics** Through The Years. Comprehensive historical section, including detailed ...". The URL is "www.infoplease.com/ipsa/A0114094.html" with a size of 28k. It includes "Cached" and "Similar pages" links.

The third result is titled "Facts About the **Olympic Medal**". The description reads: "Olympic medals since 1928 have featured the same design on the front: a Greek goddess, the **Olympic** Rings, the coliseum of ancient Athens, a Greek vase known ...". The URL is "www.cviog.uga.edu/Projects/olympix/answer.htm" with a size of 3k. It includes "Cached" and "Similar pages" links.

The fourth result is titled "OLYMPIC STATISTICS". The description reads: "In some cases, you will find 'half **medals**': in the early **Olympics**, some people had ...".

# Relevance feedback: Why is it not used?

- Users are often reluctant to provide explicit feedback
- Implicit feedback and user session monitoring is a better solution
- RF works best when relevant documents form a cluster
- In general negative feedback doesn't hold a significant improvement

# 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

Section 1.1

- 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

Sec 9.1.1

- 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: Caveat

Section 15.1

- True evaluation of usefulness must compare to other methods taking the same amount of time.

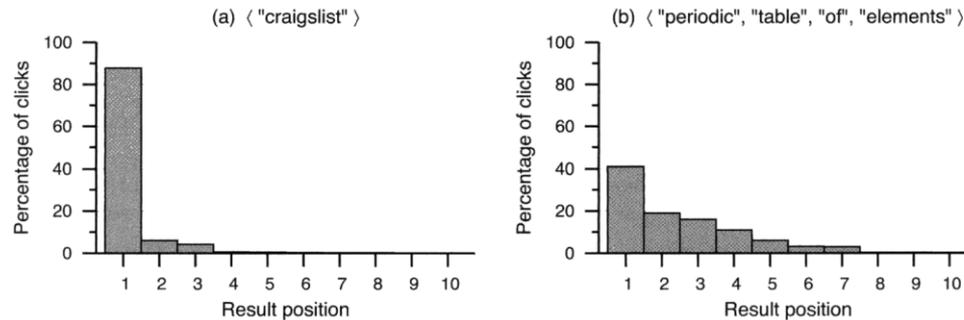


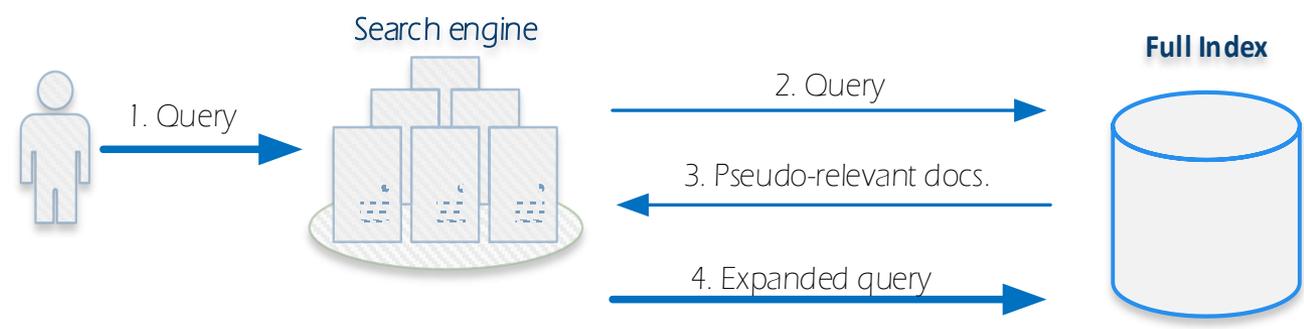
Figure 15.5 Clickthrough curve for a typical navigational query ("craigslist") and a typical informational query ("periodic", "table", "of", "elements").

- There is no clear evidence that relevance feedback is the “best use” of the user’s time

Users may prefer revision/resubmission  
to having to judge relevance of documents.

# Pseudo-relevance feedback

- Top documents are our “best guess” ...
- Given the initial query search results,
  - take **pseudo-relevant documents** from the top of this rank, and
  - generate an **expanded query** with these positive examples.



# Pseudo-relevance feedback

- The most frequent terms of all top documents are considered the pseudo-relevant terms:

$$topDocTerms = \sum_{i=1}^{\#topDocs} d_{retDocId(q_0,i)}$$

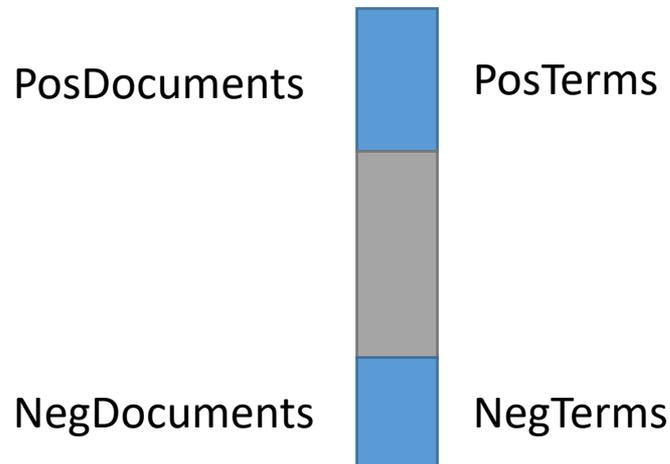
$$prfterms_i = \begin{cases} topDocTerms_i & topDocTerms_i < th \\ 0 & topDocTerms_i > th \end{cases}$$

$$, s. t. \|prfterms\|_0 = \#topterms$$

- The expanded queries then become:  $q = \gamma \cdot q_0 + (1 - \gamma) \cdot prfterms$
- Other strategies can be thought to automatically select “possibly” relevant documents

# Negative feedback

- The parameters are critical:
  - #TopDocuments
  - #TopTerms
- Excluding words from the less relevant documents also improve the selection of the expansion terms.



# Language Models

- Language Model given a document  $p(t|M_d)$ 
  - Computed from document statistics
- Language Model given a collection  $p(t|M_C)$ 
  - Computed from the collection statistics
- Language Model given a query  $p(t|Q)$ 
  - Computed from the top ranked documents
  - Based on a relevance estimation

# Relevance based language models

- Relevance based language models aims to estimate the relevance of each word for a given **query Q** and the **set  $\Theta$  of documents** retrieved with that query Q

$$p(w|Q, \Theta)$$

- This lets the system expand the initial query with new words captured from the set of documents  $\Theta$ :

$$\text{Expanded query} = \underbrace{\{w_1 \ w_2 \ \dots \ w_n\}}_{\substack{\text{Original query} \\ \text{words}}} \underbrace{\{w_p \ w_{p+1} \ w_{p+2} \ \dots \ w_{p+n}\}}_{\text{Expansion query words}}$$

# Expanding the query

- The relevance of each word in the expanded query is:

$$p(w|M'_Q) = (1 - \alpha) \cdot p(w|M_Q) + \alpha \cdot p_1(w|Q)$$

- $p(w|M_Q)$  is given by the original query.
- $p(w|Q)$  is given by the relevance-based model computed from the feedback documents.
- Words with higher probability  $p(w|M'_Q)$  will be used to generate the new expanded query.

# The expanded query

- The query vector of the expanded query is now a vector of probabilities:

$$\left[ \underbrace{p(w_1|M'_Q) \ p(w_2|M'_Q) \ \dots \ p(w_n|M'_Q)}_{\text{Original query words}} \ \underbrace{p(w_{n+1}|M'_Q) \ \dots \ p(w_{n+p}|M'_Q)}_{\text{Expansion query words}} \right]$$

- Words with probabilities below a given threshold should be zeroed.

# Relevance Model 3: i.i.d sampling

- The first approach assumes independence between query words:

$$p_{RM1}(w|Q) \propto \sum_{M_d \in \Theta} p(w|M_d)p(M_d) \prod_{i=1}^m p(q_i|M_d)$$

- The final relevance language model becomes:

$$p_{RM3}(w|M'_Q) = (1 - \alpha) \cdot p(w|M_Q) + \alpha \cdot p_{RM1}(w|Q)$$

- The  $\alpha$  parameter interpolates the original query with the new query.

# Relevance Model 4: conditional independence

- The second approach assumes conditional independence between query words and expansion words:

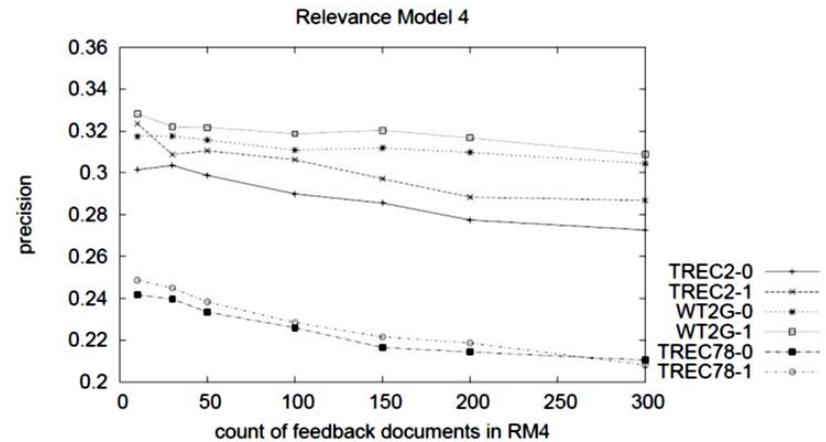
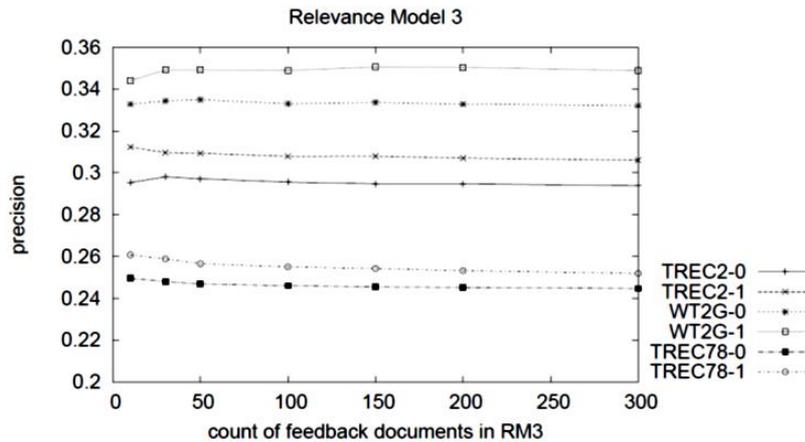
$$p_{RM2}(w|Q) \propto p(w) \prod_{i=1}^m \sum_{\theta_d \in \Theta} p(q_i|M_d) \frac{p(w|M_d)p(M_d)}{p(w)}$$

- The final relevance language model becomes:

$$p_{RM4}(w|M'_Q) = (1 - \alpha) \cdot p(w|M_Q) + \alpha \cdot p_{RM2}(w|Q)$$

- The  $\alpha$  parameter interpolates the original query with the new query.

# Comparison



Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01)*.

Yuanhua Lv and ChengXiang Zhai. 2009. A comparative study of methods for estimating query language models with pseudo feedback. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*.

Improve your query by selecting a topic

news

politics

business

Your query was expanded using news

- syria
- trump
- ruusia
- war
- latest
- cease
- envoy
- turkey
- aleppo
- fire
- civil
- says
- ground
- federal
- un
- deal
- calls
- northwest
- israeli
- ceasefire

Top Tweets

About 32.4 million results (5.91 seconds)



CatoTheYounger

@catoletters

@carlajo1947 US has Zero roll in Syria. Trump needs to get out of the way and let Russia and Syria handle ISIS in Syria

1:27 AM - 30 Nov 2016

1 retweet 1 like



International News

@Ghulam\_Rasool1

#pakistan#news Syria truce brings 'significant drop' in fighting: ALEPPO, Syria (AFP) - The UN's Syria envoy ... [bit.ly/2cFw0Jk](http://bit.ly/2cFw0Jk)

6:56 AM - 14 Sep 2016



Syria truce brings 'significant dr...

It is the latest bid to end a conflict that has killed more than 300,000 [dunyanews.tv](http://dunyanews.tv)

1 retweet 1 like



FRANCIS K S LIM

@cgnetwork

Syria government 'approves' US-Russia truce deal - state media: DAMASCUS, Syria - Syria's... [goo.gl/fb/HfqPrW](http://goo.gl/fb/HfqPrW)

8:08 PM - 10 Sep 2016



In the News



New York Times World

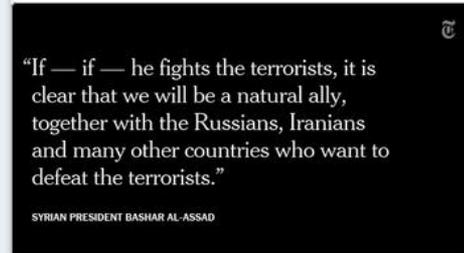
@nytimesworld

U.N.'s Syria envoy says Trump has limited window to work with Russia in Syria conflict.

[nyti.ms/2fjB967](http://nyti.ms/2fjB967)

2:28 AM - 22 Nov 2016

28 retweets 13 likes



The New York Times

@nytimes

Syria's leader calls Trump "a natural ally" in the fight against terror. Syria has called its opponents terrorists. [nyti.ms/2fwXFvS](http://nyti.ms/2fwXFvS)

3:10 PM - 16 Nov 2016

157 retweets 163 likes



The Associated Press

@AP

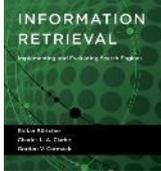
The Latest on war in Syria: Russian helicopter

# PRF iterations: Query drift

- Using multiple iterations of PRF may drift the interpretation of the original query, hurting results.

<b>Iteration</b>	<b>Expansion Terms</b>
1	dog, sniffing, canine, pooper, officers, metro, canines, police, animal, narcotics
2	dog, canine, pooper, sniffing, leash, metro, canines, animal, officers, narcotics
3	dog, canine, pooper, sniffing, leash, metro, canines, animal, owners, pets
4	dog, leash, animal, metro, canine, pooper, sniffing, canines, owners, pets
5	dog, leash, metro, canine, pooper, sniffing, canines, owners, animal, pets
6	dog, leash, metro, pooper, canines, owners, pets, animals, canine, scooper
7	dog, leash, metro, pooper, canines, owners, pets, animals, canine, scooper

# Experimental comparison



Method	TREC45				Gov2			
	1998		1999		2004		2005	
	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP
BM25	0.424	0.178	0.440	0.205	0.471	0.243	0.534	0.277
BM25+PRF	0.452	0.239	0.454	0.249	0.567	0.277	0.588	0.314

# Example with top 2 documents

- **Query:**
  - “Donald Trump”
- **Top retrieved doc1:**
  - “Donald Trump lashes out at figures who have been critical of him”
- **Top retrieved doc2:**
  - “Demi Lovato has been critical of Donald Trump”

# Summary

- PRF can improve top precision.
- It's often harder to understand why a particular document was retrieved after applying PRF.
- 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