Relevance Feedback

CS4201: Information Retrieval and Web Search

Dwaipayan Roy

IISER, Kolkata

Motivation

- Searching depends on matching keywords between user-query and document.

Motivation

- Searching depends on matching keywords between user-query and document.
- Searchers and document creators may use different keywords to denote same "concept".

Motivation

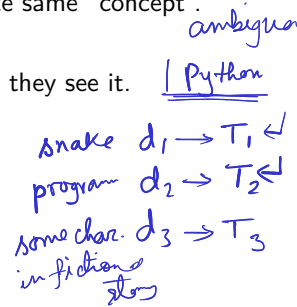
- Searching depends on matching keywords between user-query and document.
- Searchers and document creators may use different keywords to denote same "concept".
- Vocabulary mismatch $\Longrightarrow$ poor retrieval quality.

Motivation

- Searching depends on matching keywords between user-query and document.
- Searchers and document creators may use different keywords to denote same "concept".
- Vocabulary mismatch $\Longrightarrow$ poor retrieval quality.
- Users may not know what they are looking for, but they'll know when they see it.

Motivation

- Searching depends on matching keywords between user-query and document.
- Searchers and document creators may use different keywords to denote same "concept".
- Vocabulary mismatch $\Longrightarrow$ poor retrieval quality.
- Users may not know what they are looking for, but they'll know when they see it.
- Boost recall: "find me similar documents..."

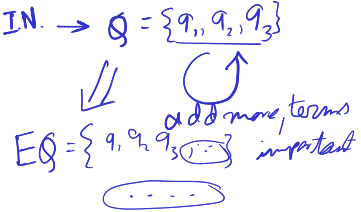


Motivation

- Searching depends on matching keywords between user-query and document.
- Searchers and document creators may use different keywords to denote same "concept".
- Vocabulary mismatch $\Longrightarrow$ poor retrieval quality.
- Users may not know what they are looking for, but they'll know when they see it.
- Boost recall: "find me similar documents..."
- Solution: try to make **a better representation** of the information need.

Motivation

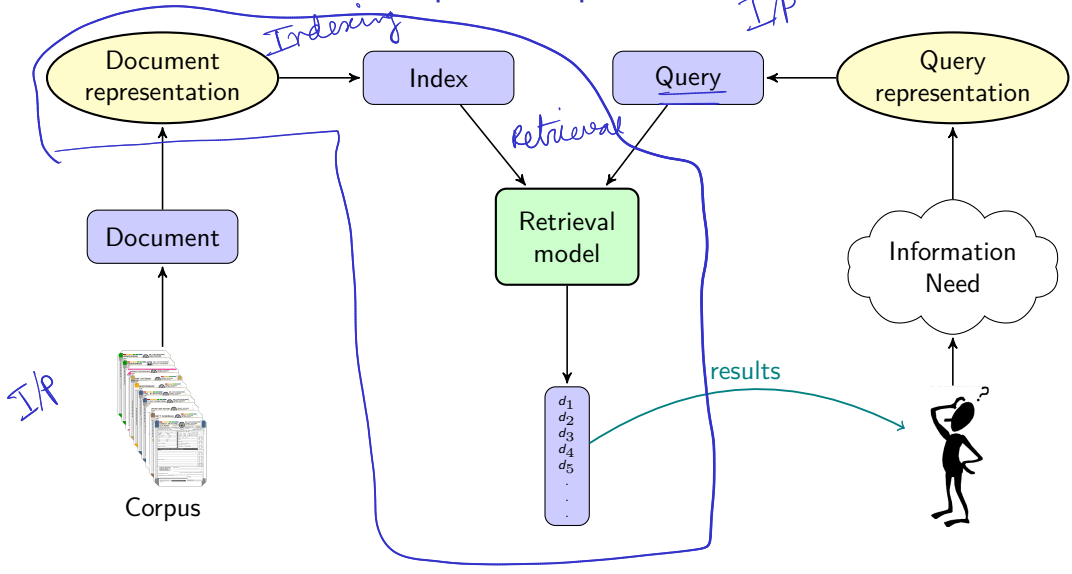
- Searching depends on matching keywords between user-query and document.
- Searchers and document creators may use different keywords to denote same "concept".
- Vocabulary mismatch $\Longrightarrow$ poor retrieval quality.
- Users may not know what they are looking for, but they'll know when they see it.
- Boost recall: "find me similar documents..."
- Solution: try to make **a better representation** of the information need.
  - expand the query by adding related words/phrases.



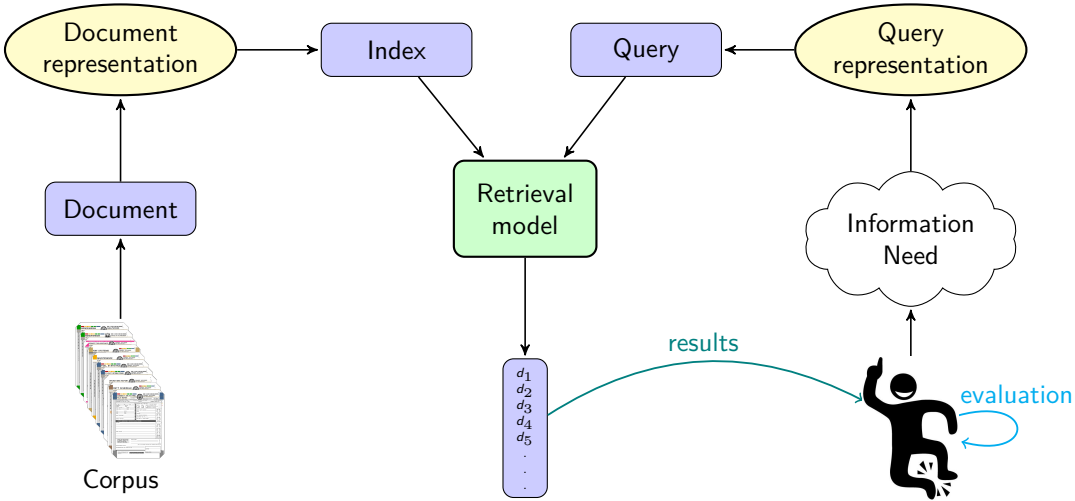
Motivation

- Searching depends on matching keywords between user-query and document.
- Searchers and document creators may use different keywords to denote same "concept".
- Vocabulary mismatch $\Longrightarrow$ poor retrieval quality.
- Users may not know what they are looking for, but they'll know when they see it.
- Boost recall: "find me similar documents..."
- Solution: try to make **a better representation** of the information need.
  - expand the query by adding related words/phrases.
- Issues:
  - **select** which terms to add to query
  - **calculate weights** for added terms

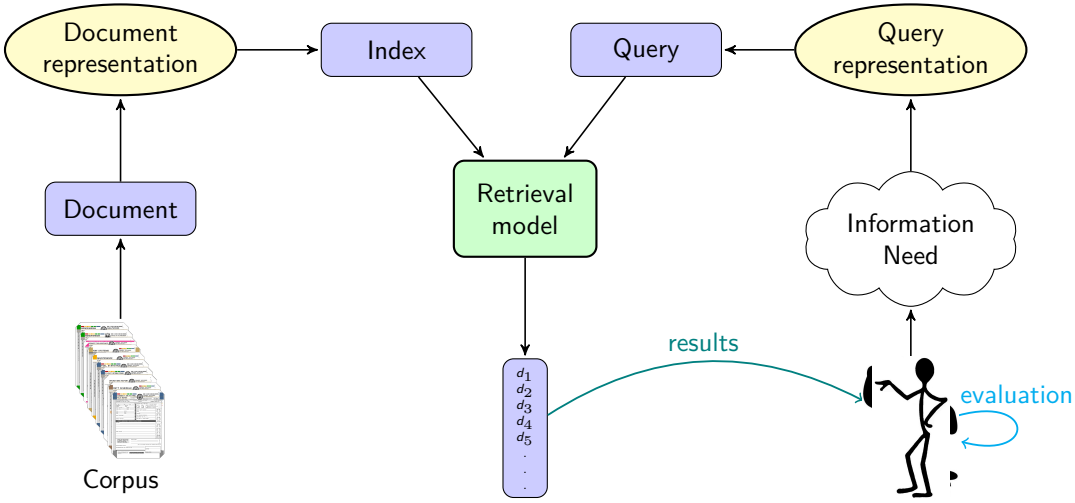
Relevance Feedback: A Graphical Representation



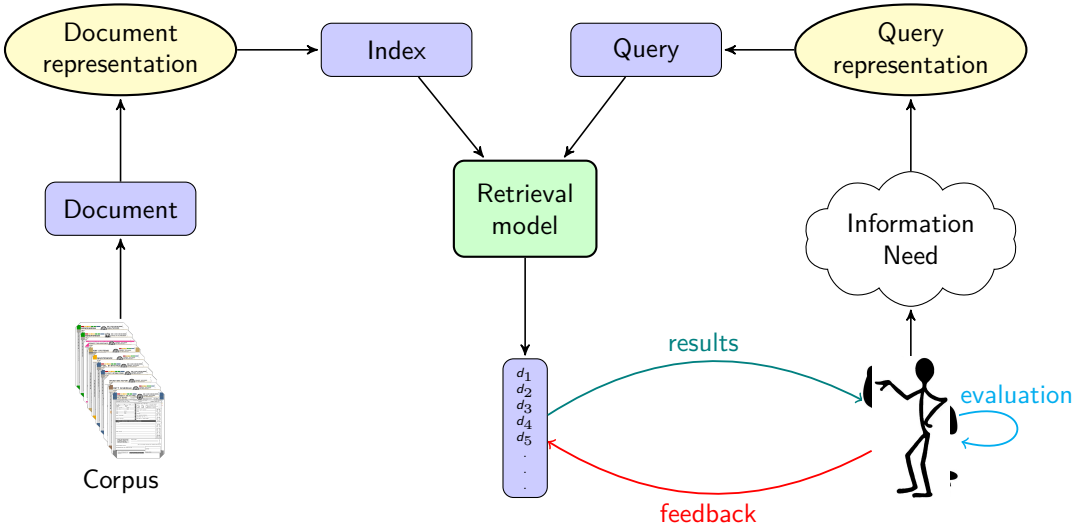
Relevance Feedback: A Graphical Representation



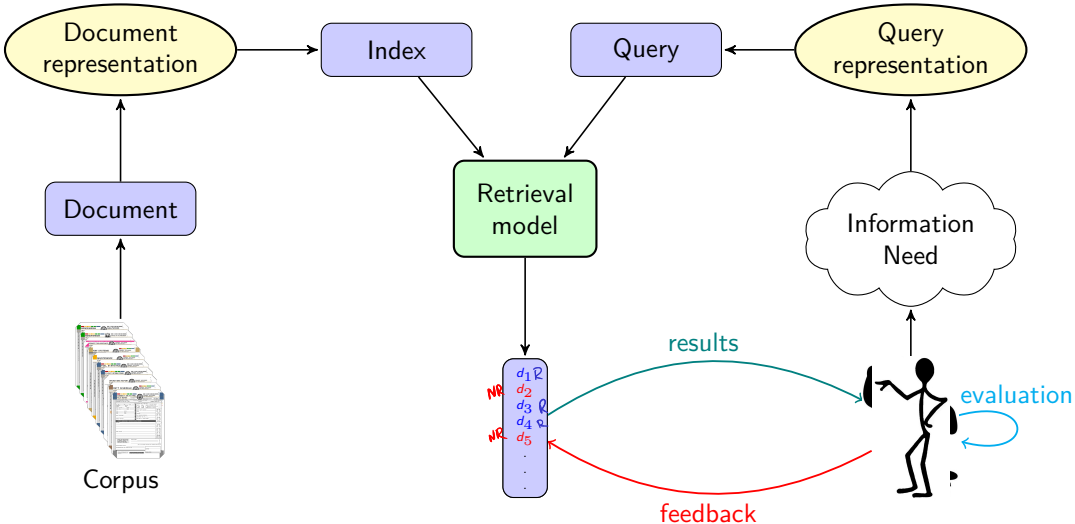
Relevance Feedback: A Graphical Representation



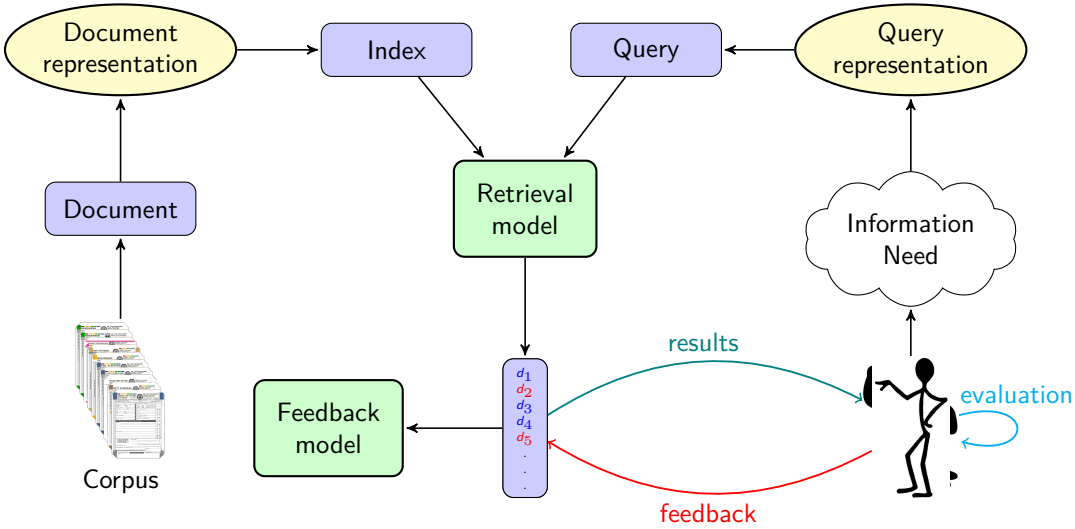
Relevance Feedback: A Graphical Representation



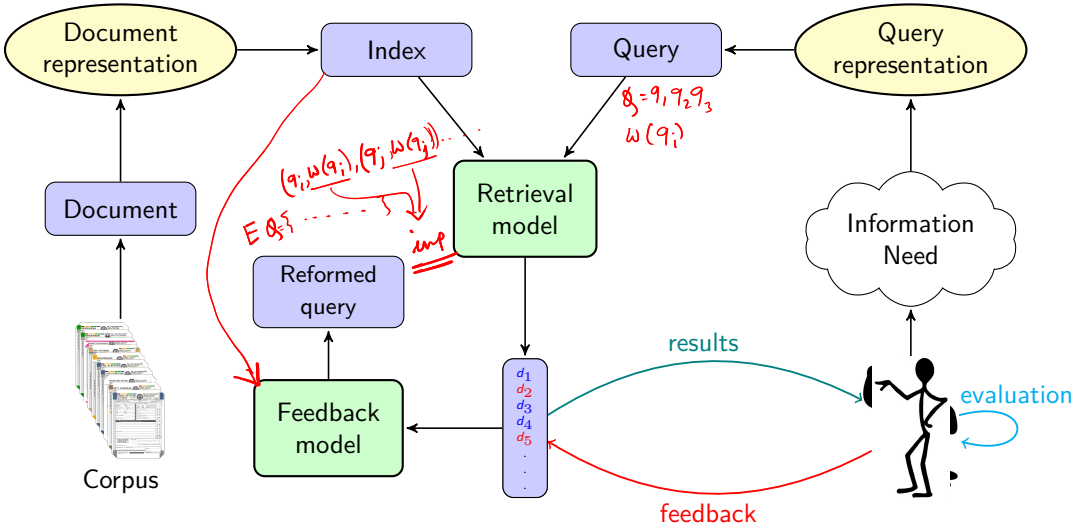
Relevance Feedback: A Graphical Representation



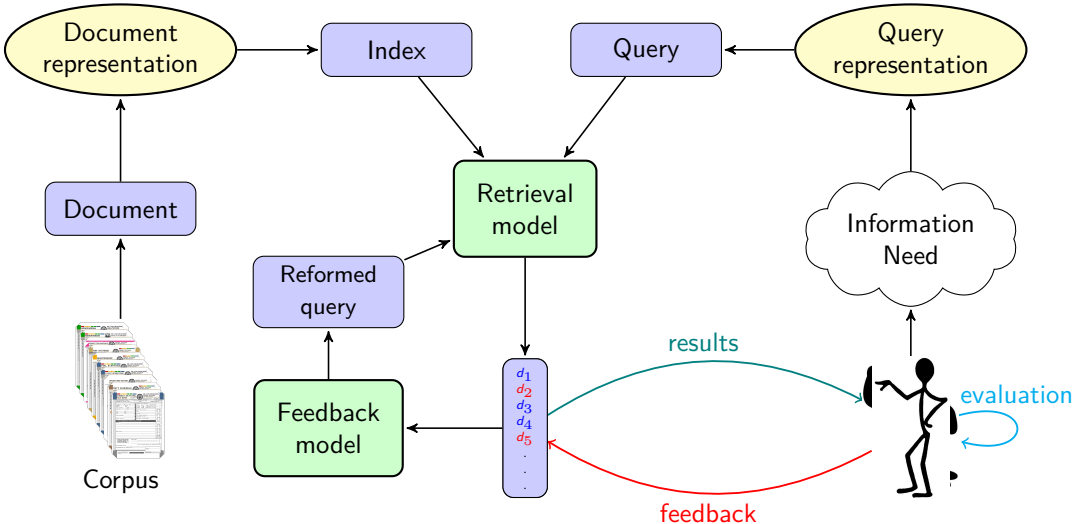
Relevance Feedback: A Graphical Representation



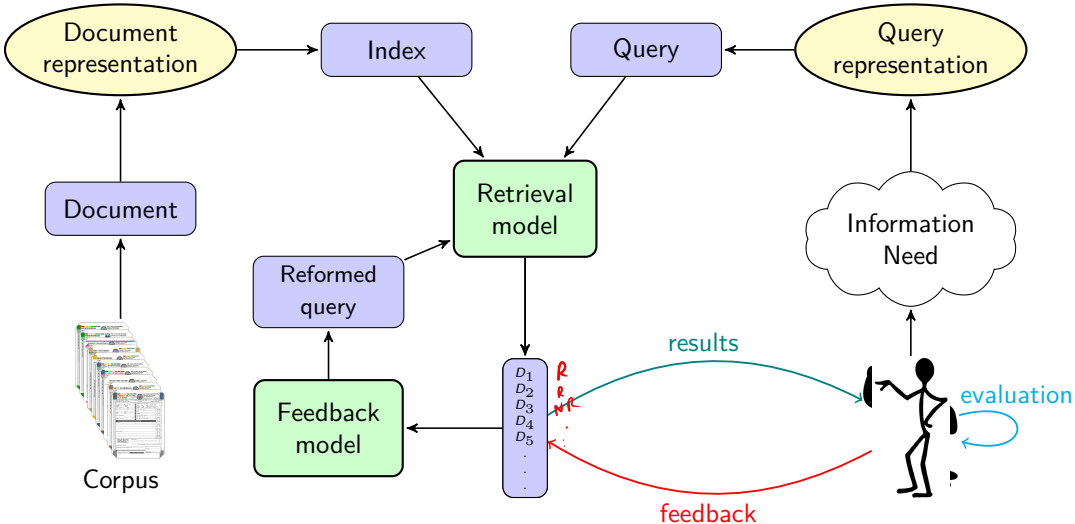
Relevance Feedback: A Graphical Representation



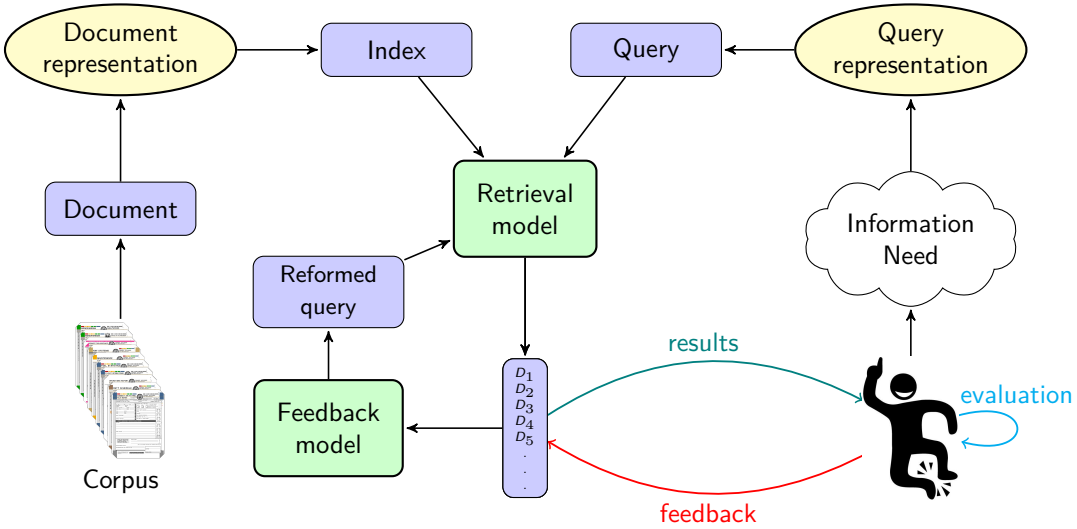
Relevance Feedback: A Graphical Representation



Relevance Feedback: A Graphical Representation



Relevance Feedback: A Graphical Representation



Relevance Feedback: Basic Idea

- User issues a query. 

Relevance Feedback: Basic Idea

- User issues a query.
- Search engine returns a set of documents.

Relevance Feedback: Basic Idea

- User issues a query.
- Search engine returns a set of documents.
- User marks some documents as relevant, some as non-relevant.

Relevance Feedback: Basic Idea

- User issues a query.
- Search engine returns a set of documents.
- User marks some documents as relevant, some as non-relevant.
- Search engine uses this feedback information, together with collection statistics to formulate **a better representation** of the information need.

Relevance Feedback: Basic Idea

- User issues a query.
- Search engine returns a set of documents.
- User marks some documents as relevant, some as non-relevant.
- Search engine uses this feedback information, together with collection statistics to formulate **a better representation** of the information need.
- Search engine performs retrieval with the reformulated query.
- The new set of documents returned by the engine, having (hopefully) better recall.

Relevance Feedback: Basic Idea

- User issues a query.
- Search engine returns a set of documents.
- User marks some documents as relevant, some as non-relevant.
- Search engine uses this feedback information, together with collection statistics to formulate **a better representation** of the information need.
- Search engine performs retrieval with the reformulated query.
- The new set of documents returned by the engine, having (hopefully) better recall.
- Can iterate this: several rounds of relevance feedback.

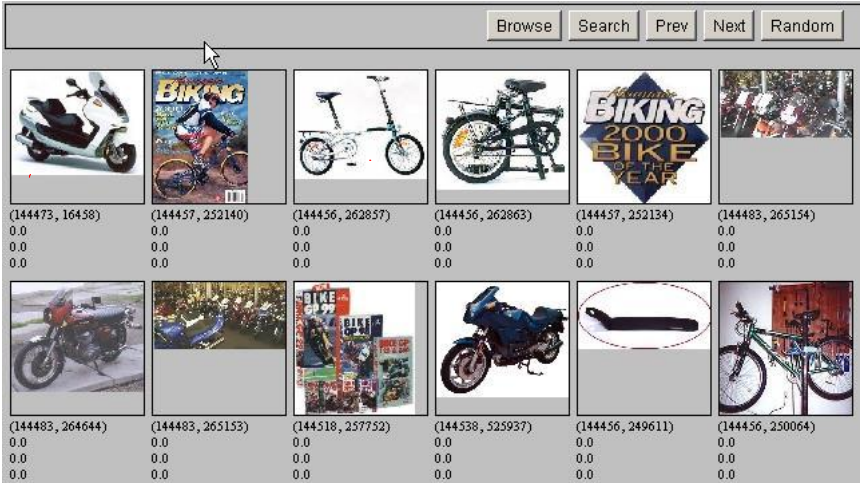
Relevance Feedback: Example 1

- Initial Query: bike

bicycle

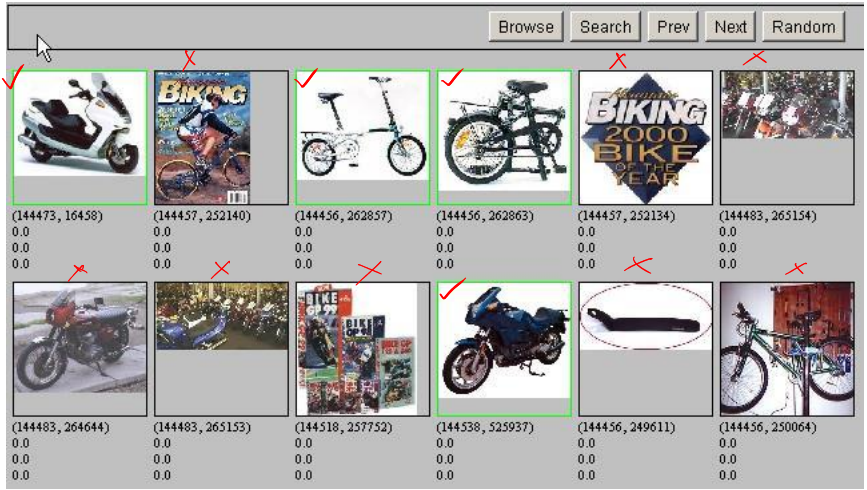
Relevance Feedback: Example 1

- Results for initial query



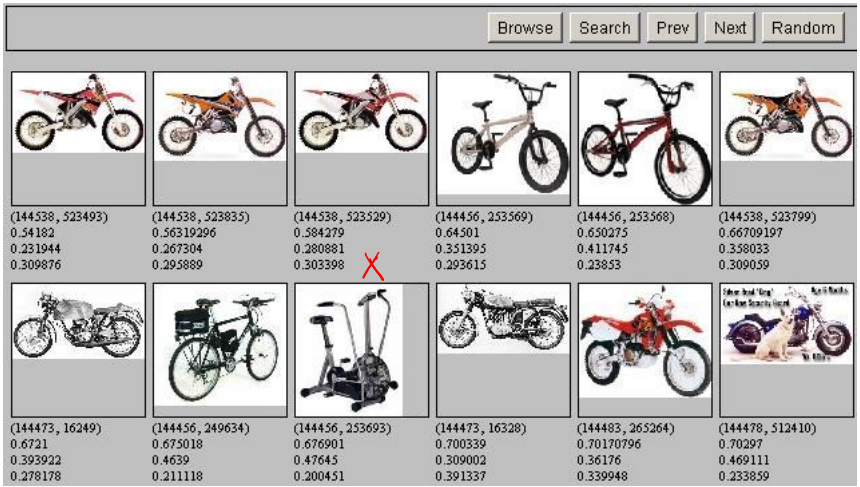
Relevance Feedback: Example 1

- Feedback: Relevant document selected



Relevance Feedback: Example 1

- Result after re-retrieval



Relevance Feedback: Example 2

- TREC 2 Topic - 113 : *New Space Satellite Applications*

1993
101 → 150

1992 → TREC1 51 → 100
TREC6 301 → 350
Topic
→ title
→ desc
→ narr

Relevance Feedback: Example 2

- TREC 2 Topic - 113 : *New Space Satellite Applications*

Rank	Score	Document title
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

Relevance Feedback: Example 2

- TREC 2 Topic - 113 : *New Space Satellite Applications*

	Rank	Score	Document title
✓	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

Relevance Feedback: Example 2

- Initial query: *new space satellite applications*
- 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

Relevance Feedback: Example 2

- TREC 2 Topic - 113 : *New Space Satellite Applications*

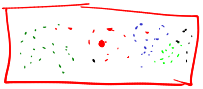
Rank	Score	Document title
1	0.513	NASA Scratches Environment Gear From Satellite Plan
2	0.500	NASA Hasn't Scrapped Imaging Spectrometer
3	0.493	When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
4	0.493	NASA Uses 'Warm' Superconductors For Fast Circuit
5	0.492	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

Relevance Feedback: Example 2

- TREC 2 Topic - 113 : *New Space Satellite Applications*

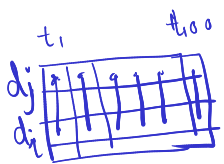
Old	Rank	Score	Document title
2	1	0.513	NASA Scratches Environment Gear From Satellite Plan
1	2	0.500	NASA Hasn't Scrapped Imaging Spectrometer
	3	0.493	When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
	4	0.493	NASA Uses 'Warm' Superconductors For Fast Circuit
8	5	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

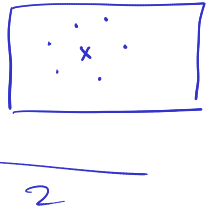
Key concept for relevance feedback: Centroid



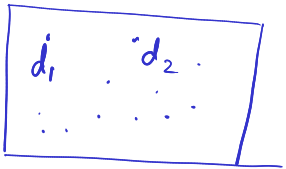
- The centroid is the center of mass of a set of points.
- Recall that we represent documents as points in a high-dimensional space.
- Thus: we can compute centroids of documents.
- The centroid of a set of documents D:

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





Rocchio's Algorithm for Relevance Feedback



- Standard algorithm used for relevance feedback (SMART, '70s).
- Integrates a measure of relevance feedback into Vector Space Model.
- Maximize the difference between average similarities for relevant and non-relevant documents.

Rocchio's Algorithm for Relevance Feedback

- Chooses the optimal query $\mathbf{Q_{opt}}$ that maximizes:

$$\mathbf{Q}_{opt} = argmax_{\mathbf{Q}} \left[sim(\mathbf{Q}, \mu(D_R)) - sim(\mathbf{Q}, \mu(D_{NR})) \right]$$

- D_R: Set of relevant documents
- D_{NR}: Set of non-relevant documents
- $\mu(D_X)$: Centroid of a set of documents X

Rocchio's Algorithm for Relevance Feedback

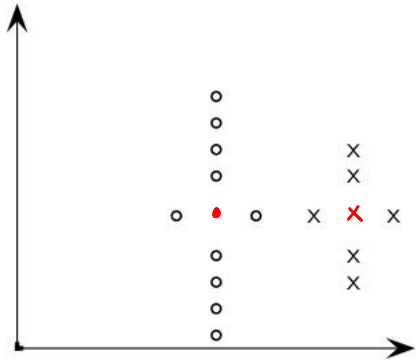
- Chooses the optimal query $\mathbf{Q_{opt}}$ that maximizes:

$$\mathbf{Q}_{opt} = argmax_{\mathbf{Q}}\Big[sim(\mathbf{Q}, \mu(D_R)) - sim(\mathbf{Q}, \mu(D_{NR}))\Big]$$

- D_R: Set of relevant documents
- D_{NR}: Set of non-relevant documents
- $\mu(D_X)$: Centroid of a set of documents X

To find the vector $\mathbf{Q}_{opt}$, that separates relevant and non-relevant documents maximally.

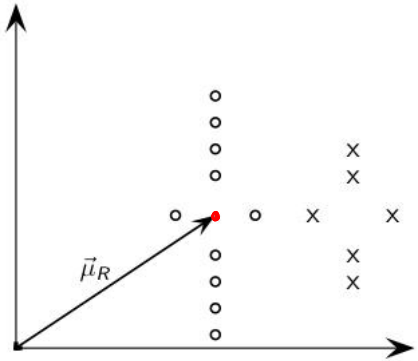
Rocchio's Algorithm Illustrated



○ relevant documents

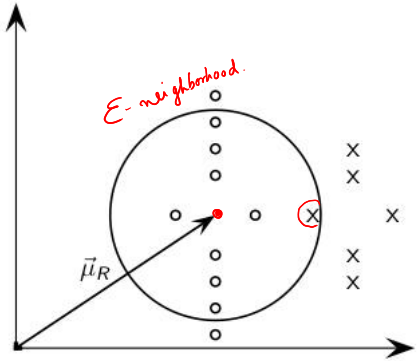
× non-relevant documents

Rocchio's Algorithm Illustrated



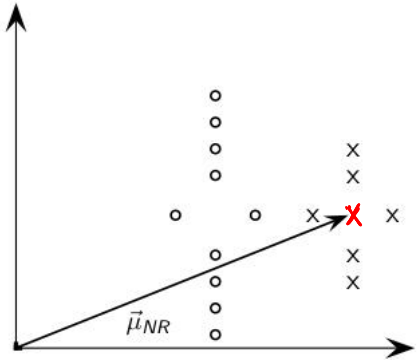
μ_R : centroid of relevant documents

Rocchio's Algorithm Illustrated



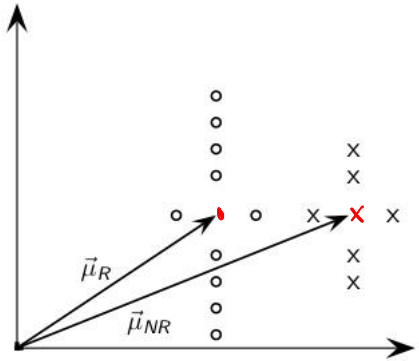
μ_R not seperating relevant documents from non-relevant

Rocchio's Algorithm Illustrated

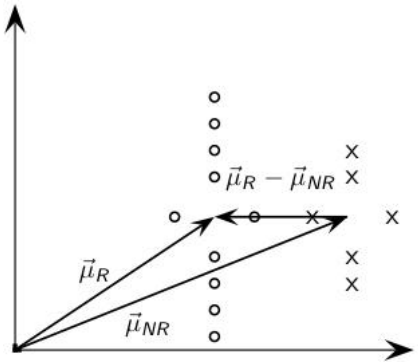


μ_{NR} : centroid of non-relevant documents

Rocchio's Algorithm Illustrated

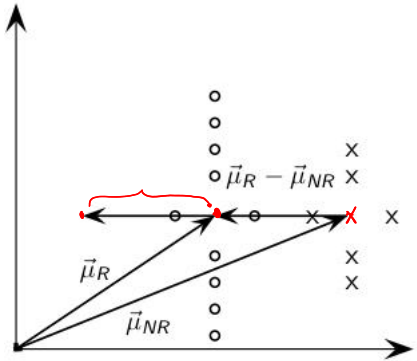


Rocchio's Algorithm Illustrated

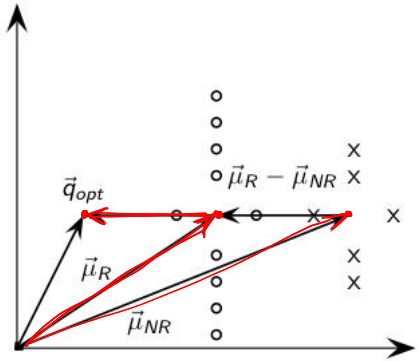


$\mu_R - \mu_{NR}$: difference vector

Rocchio's Algorithm Illustrated

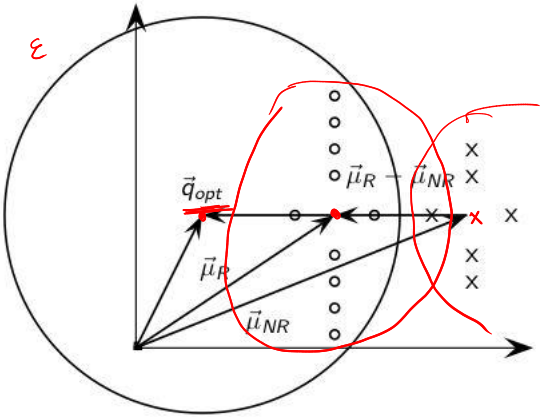


Rocchio's Algorithm Illustrated

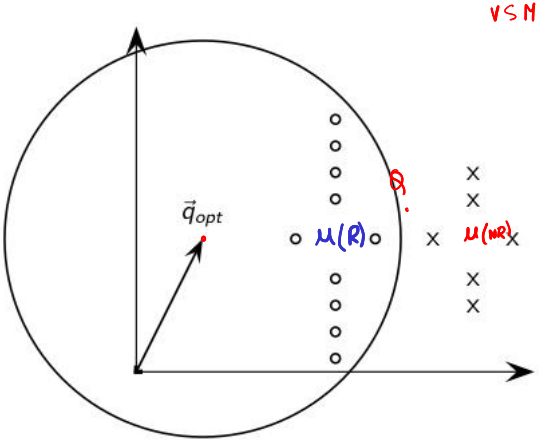


Add difference vector to μ_R to get μ_{opt}

Rocchio's Algorithm Illustrated



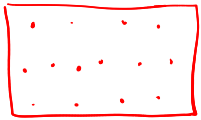
Rocchio's Algorithm Illustrated



$\mathbf{Q}_{opt}$ seperates relevant and non-relevant documents perfectly

Rocchio's Algorithm for Relevance Feedback

- The optimal query vector:



$$\mathbf{Q}_{opt} = \mu(D_R) + \left[\mu(D_R) - \mu(D_{NR})\right]$$

Rocchio's Algorithm for Relevance Feedback

- The optimal query vector:

$$\begin{aligned}\mathbf{Q}_{opt} &= \mu(D_R) + \Big[\mu(D_R) - \mu(D_{NR})\Big] \\ &= \frac{1}{|D_R|}\sum_{\mathbf{d_i}\in D_R}\mathbf{d_i} + \Big[\frac{1}{|D_R|}\sum_{\mathbf{d_i}\in D_R}\mathbf{d_i} - \frac{1}{|D_{NR}|}\sum_{\mathbf{d_j}\in D_{NR}}\mathbf{d_j}\Big]\end{aligned}$$

Rocchio's Algorithm for Relevance Feedback

- The optimal query vector:

$$\mathbf{Q}_{opt} = \mu(D_R) + \left[\mu(D_R) - \mu(D_{NR})\right]$$

$$= \frac{1}{|D_R|}\sum_{\mathbf{d_i}\in D_R}\mathbf{d_i} + \left[\frac{1}{|D_R|}\sum_{\mathbf{d_i}\in D_R}\mathbf{d_i} - \frac{1}{|D_{NR}|}\sum_{\mathbf{d_j}\in D_{NR}}\mathbf{d_j}\right]$$

The centroid of the relevant documents moved by the difference between the two centroids.

Rocchio's Algorithm

- In practice:

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

parameters

tune

original Q_0

Rel. Docs

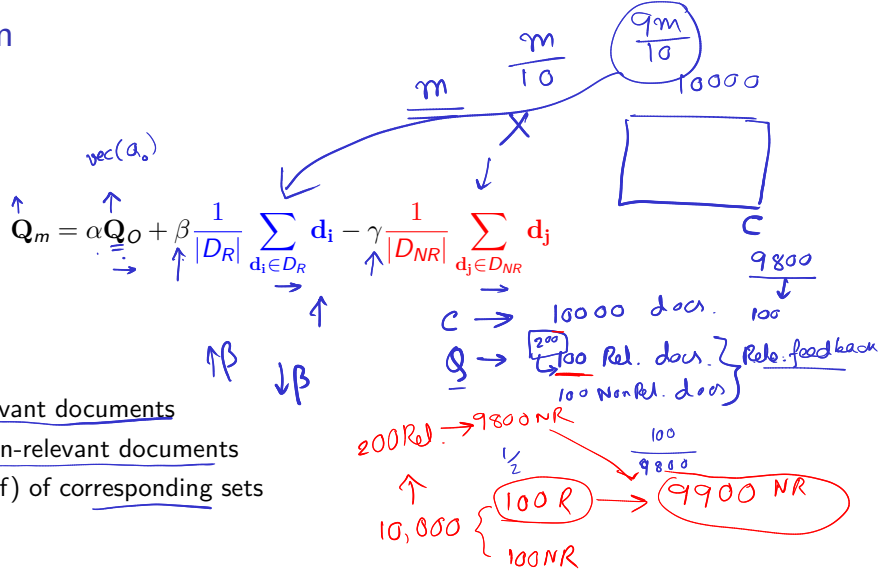
Non Rel Docs.

Rocchio's Algorithm

- In practice:

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

- Q_O: original query
- Q_m: modified query
- D_R: set of known relevant documents
- D_{NR}: set of known non-relevant documents
- α, β, γ: weights (belief) of corresponding sets



Rocchio's Algorithm: Parameter Settings

$$\underline{\underline{\mathbf{Q}_m}} = \alpha \mathbf{Q}_O + \underline{\beta} \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \underline{\gamma} \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

New (modified) query

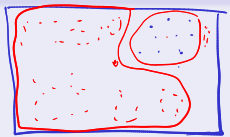
- Set higher $\underline{\beta}, \underline{\gamma}$ if there is a lot of judged documents.

Rocchio's Algorithm: Parameter Settings

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

New (modified) query

- Set higher β, γ if there is a lot of judged documents.
- Positive feedback more valuable than negative feedback. (why?)



Rocchio's Algorithm: Parameter Settings

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

New (modified) query

- Set higher β, γ if there is a lot of judged documents.
- Positive feedback more valuable than negative feedback. (why?)
- Can set negative term weight to 0 (negative weight doesn't make sense in vector space model).

Rocchio's Algorithm: Example

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

	t_1	t_2	t_4	t_6		
Original query	0	4	0	8	0	0
Positive feedback →	2	4	8	0	0	2
Negative feedback →	8	0	4	4	0	16

$\alpha = 1.0$ | $\beta = 0.5$ | $\gamma = 0.25$

0	4	0	8	0	0
1	2	4	0	0	1
2	0	1	1	0	4

-1	6	3	7	0	-3

Rocchio's Algorithm: Example

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

Original query	0	4	0	8	0	0	$\alpha = 1.0$	0	4	0	8	0	0
Positive feedback	2	4	8	0	0	2	$\beta = 0.5$	1	2	4	0	0	1
Negative feedback	8	0	4	4	0	16	$\gamma = 0.25$	2	0	1	1	0	4

expanded query

New query

t_1	t_2	t_3	t_4	t_5	t_6
-1	6	3	7	0	-3
0	6	3	7	0	0

new term

Rocchio's Algorithm: issues

What should be regarded as NR?

- Documents marked non-relevant by user?
  - too little information

Rocchio's Algorithm: issues

What should be regarded as *NR*?

- Documents marked non-relevant by user?
  - too little information
- All documents not known to be relevant?
  - Unseen relevant documents may affect term-weights

Rocchio's Algorithm: issues

What should be regarded as NR?

- Documents marked non-relevant by user?
  - too little information
- All documents not known to be relevant?
  - Unseen relevant documents may affect term-weights

How should terms be selected? How many?

Rocchio's Algorithm: issues

What should be regarded as *NR*?

- Documents marked non-relevant by user?
  - too little information
- All documents not known to be relevant?
  - Unseen relevant documents may affect term-weights

How should terms be selected? How many?

- Rank terms by # relevant documents they occur in.
- Add 50-100 terms.

Relevance Feedback: Problem

- Usually Expensive with respect to time.
- Users are reluctant to provide explicit feedback.

Indirect Relevance Feedback

- Uses *Evidences* rather than *Explicit* feedback.

Indirect Relevance Feedback

- Uses *Evidences* rather than *Explicit* feedback.
- Not specific to users.

Indirect Relevance Feedback

- Uses *Evidences* rather than *Explicit* feedback.
- Not specific to users.
- Suitable for Web search.

Indirect Relevance Feedback

- Uses *Evidences* rather than *Explicit* feedback.
- Not specific to users.
- Suitable for Web search.
- Click / Time spent on a particular retrieved document $\rightarrow$ inplicit indication of relevance.

Blind/Pseudo Relevance Feedback

- In absence of *true* feedback, assume top-ranked documents as relevant.

Blind/Pseudo Relevance Feedback

- In absence of *true* feedback, assume top-ranked documents as relevant.
- Obvious danger: if initial retrieval poor $\rightarrow$ feedback can aggravate the problem.

Blind/Pseudo Relevance Feedback

- In absence of *true* feedback, assume top-ranked documents as relevant.
- Obvious danger: if initial retrieval poor $\rightarrow$ feedback can aggravate the problem.
- Most groups at TREC found adhoc feedback useful on average.

Blind/Pseudo Relevance Feedback

Rocchio's algorithm

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

- Perform a retrieval.

Blind/Pseudo Relevance Feedback

Rocchio's algorithm

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

- Perform a retrieval.
- Consider top K ranked documents as relevant - D_R.

Blind/Pseudo Relevance Feedback

Rocchio's algorithm

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

- Perform a retrieval.
- Consider top K ranked documents as relevant - D_R.
- Which documents to be selected as non-relevant - D_{NR}?

Blind/Pseudo Relevance Feedback

Rocchio's algorithm

$$\mathbf{Q}_m = \alpha \mathbf{Q}_O + \beta \frac{1}{|D_R|} \sum_{\mathbf{d_i} \in D_R} \mathbf{d_i} - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d_j} \in D_{NR}} \mathbf{d_j}$$

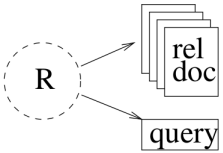
- Perform a retrieval.
- Consider top K ranked documents as relevant - D_R.
- Which documents to be selected as non-relevant - D_{NR}?
- Only consider D_R and set γ to 0.

Relevance based Language Model (RLM)

Lavrenko and Croft - SIGIR 2001

Relevance based Language Model (RLM)

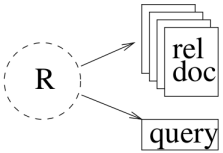
Lavrenko and Croft - SIGIR 2001



- Assumes that both query and relevant documents are sampled from a latent relevance model R.

Relevance based Language Model (RLM)

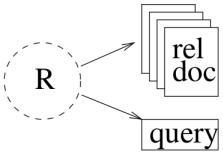
Lavrenko and Croft - SIGIR 2001



- Assumes that both query and relevant documents are sampled from a latent relevance model R.
- In absence of training data, top ranked documents considered as set of relevant documents.

Relevance based Language Model (RLM)

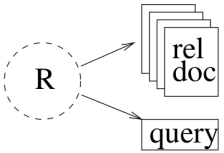
Lavrenko and Croft - SIGIR 2001



- Assumes that both query and relevant documents are sampled from a latent relevance model R.
- In absence of training data, top ranked documents considered as set of relevant documents.
- The query serves as the only absolute evidence about the relevance model.

Relevance based Language Model (RLM)

Lavrenko and Croft - SIGIR 2001



- Assumes that both query and relevant documents are sampled from a latent relevance model R.
- In absence of training data, top ranked documents considered as set of relevant documents.
- The query serves as the only absolute evidence about the relevance model.
- The task is to find (estimate) the density function for R.