

Relevance Feedback

CS4201: Information Retrieval and Web Search

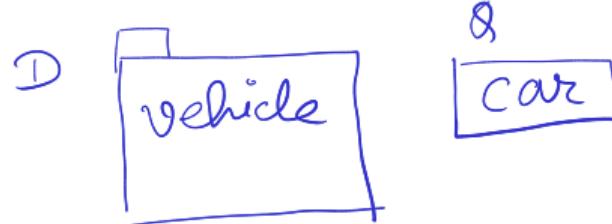
Dwaipayan Roy

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 \implies 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 \implies 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”.
ambiguous
- Vocabulary mismatch \Rightarrow poor retrieval quality.
- Users may not know what they are looking for, but they'll know when they see it.
I Python
- Boost recall: “find me similar documents...”

snake $d_1 \rightarrow T_1 \leftarrow$
program $d_2 \rightarrow T_2 \leftarrow$
some char. $d_3 \rightarrow T_3$
in fiction
story

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 \Rightarrow 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 \Rightarrow 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.

IN. $\rightarrow Q = \{q_1, q_2, q_3\}$

\Downarrow \uparrow

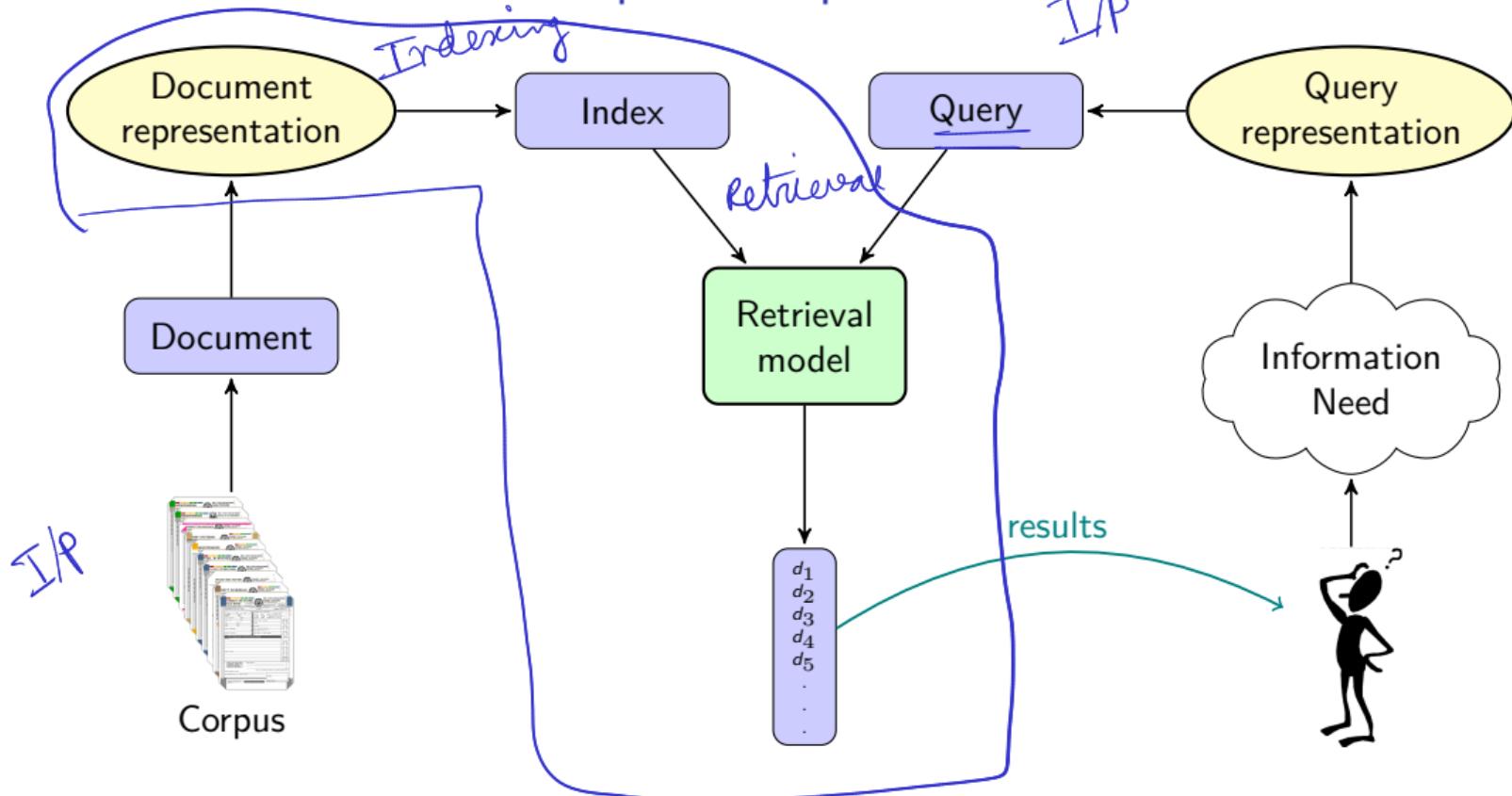
$EQ = \{q_1, q_2, q_3, \dots\}$ add more terms

important

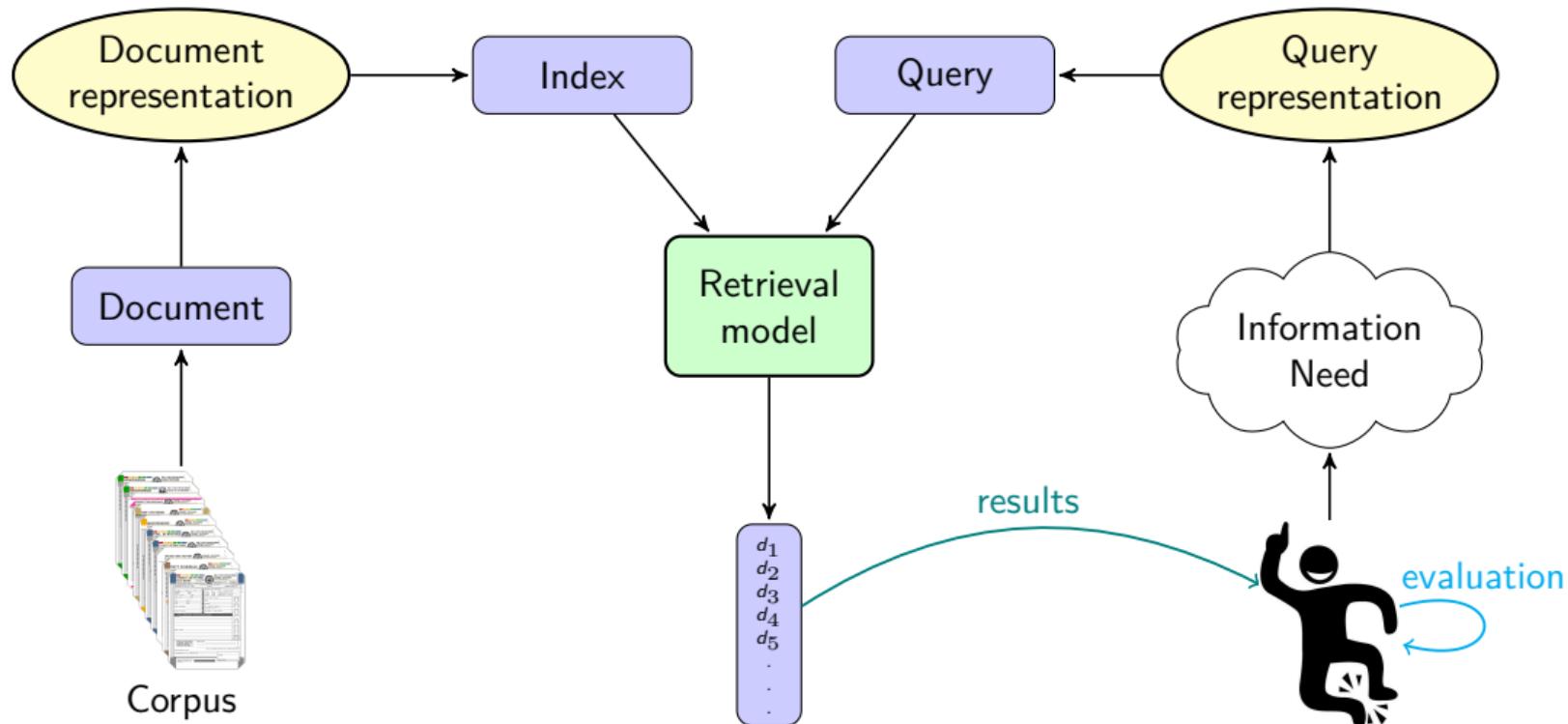
Motivation

- Searching depends on matching keywords between user-query and document.
- Searchers and document creators may use different keywords to denote same “concept”.
- Vocabulary mismatch \Rightarrow 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

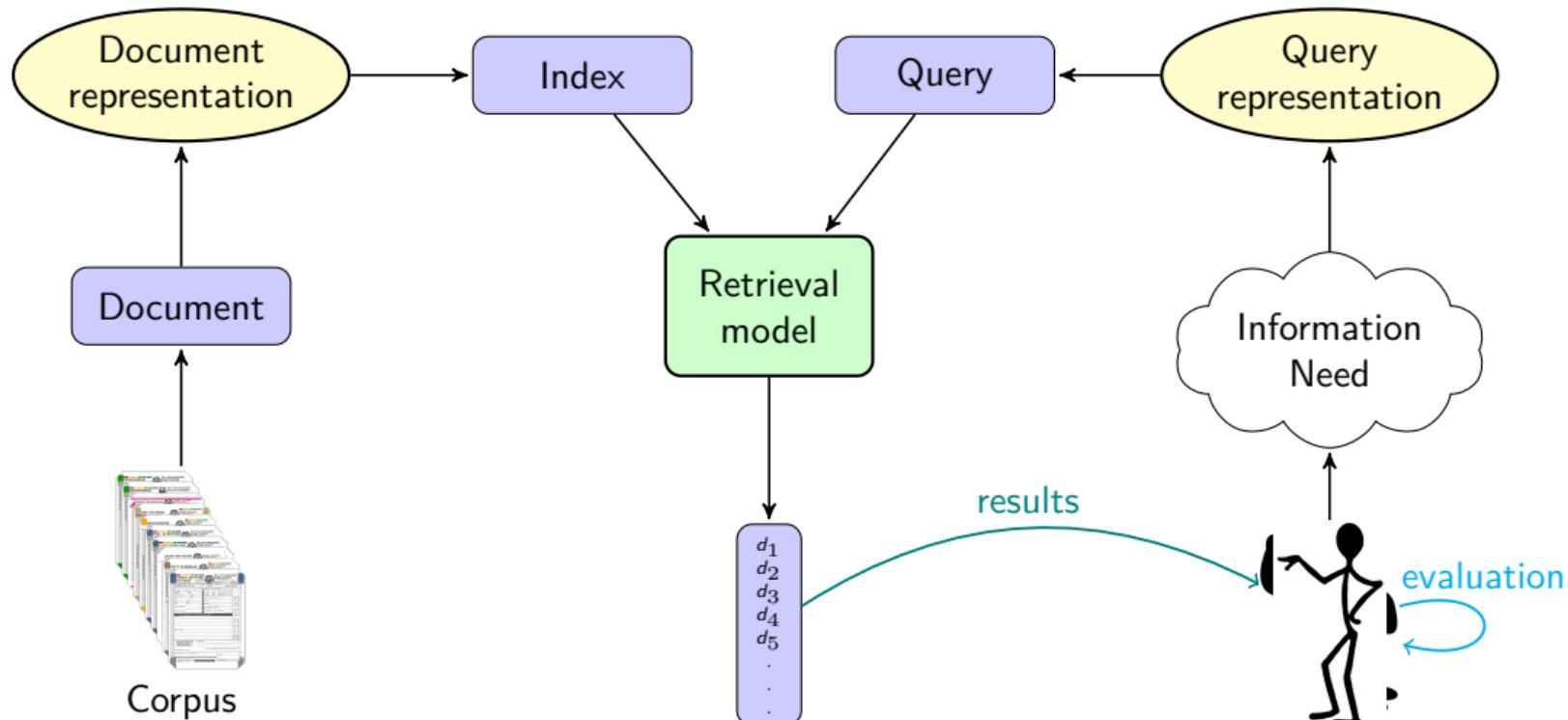
Relevance Feedback: A Graphical Representation



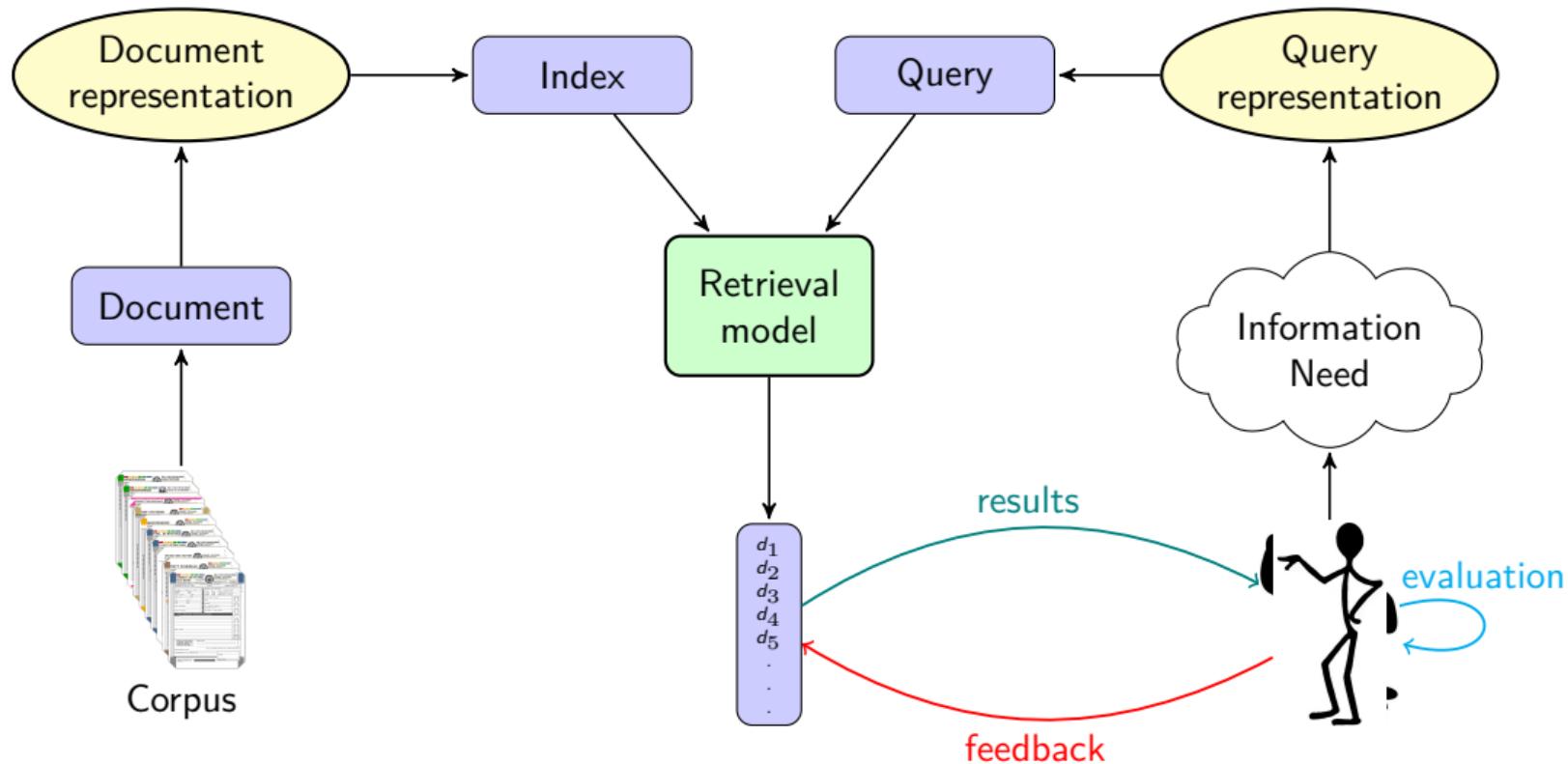
Relevance Feedback: A Graphical Representation



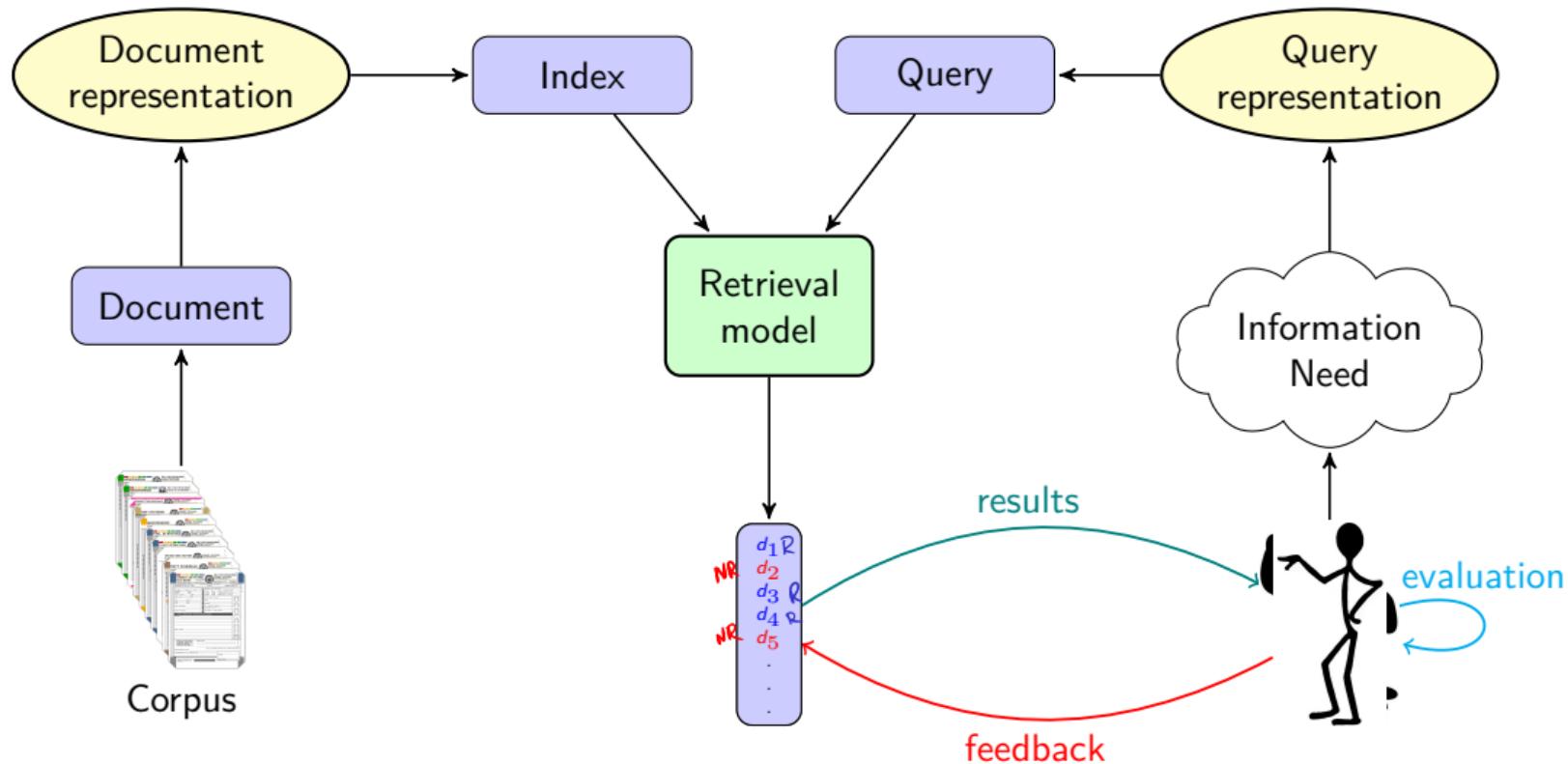
Relevance Feedback: A Graphical Representation



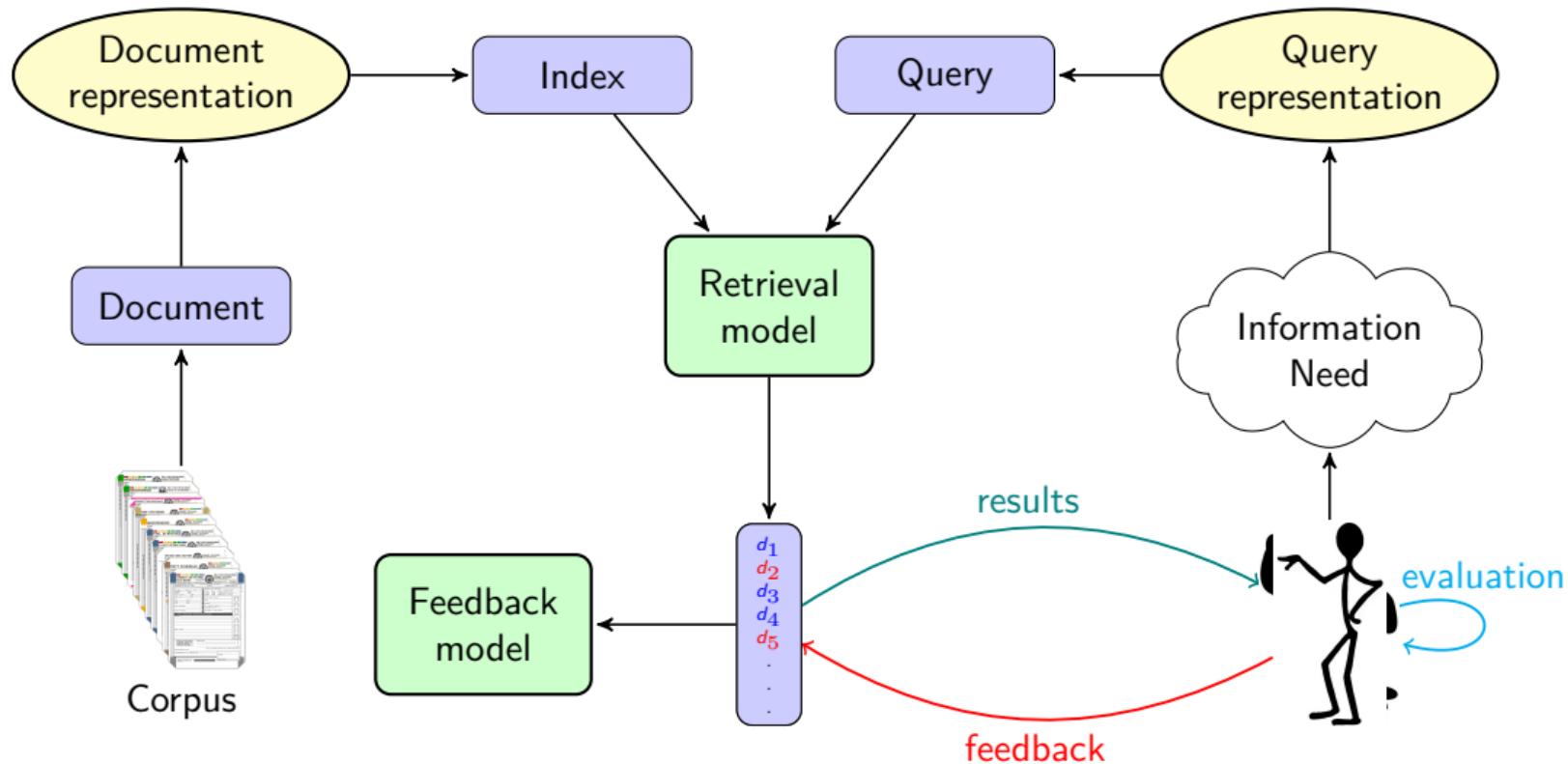
Relevance Feedback: A Graphical Representation



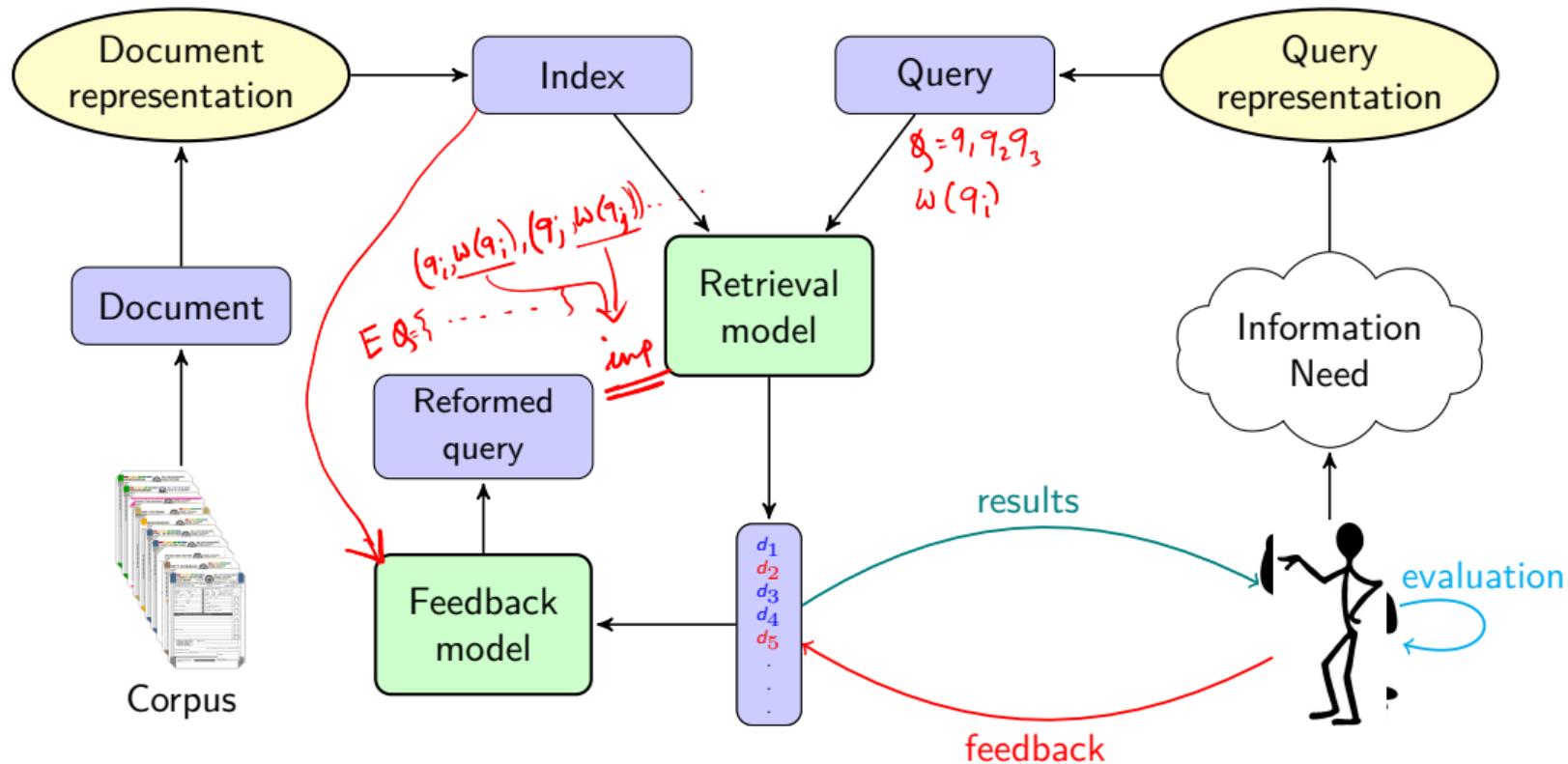
Relevance Feedback: A Graphical Representation



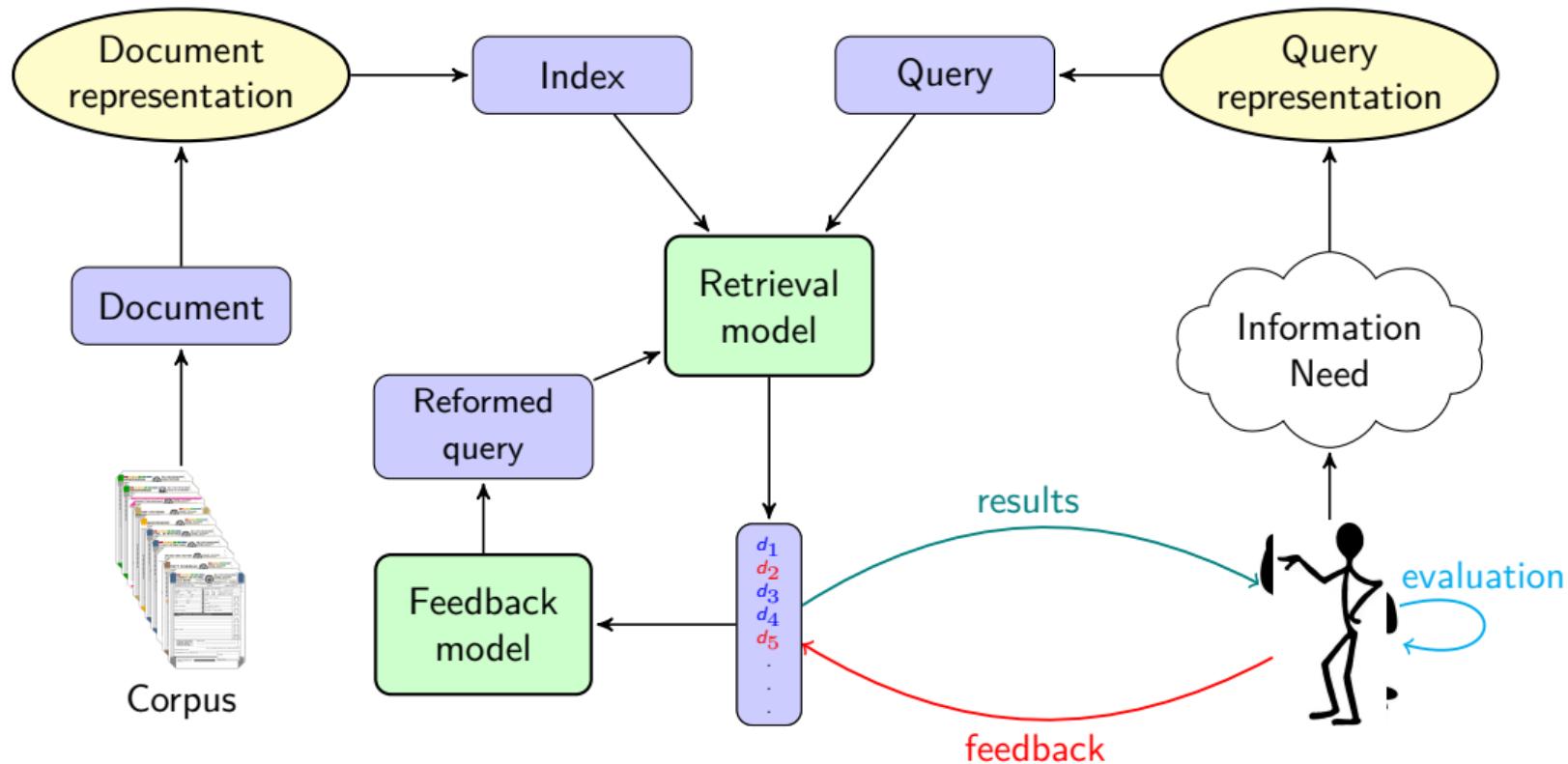
Relevance Feedback: A Graphical Representation



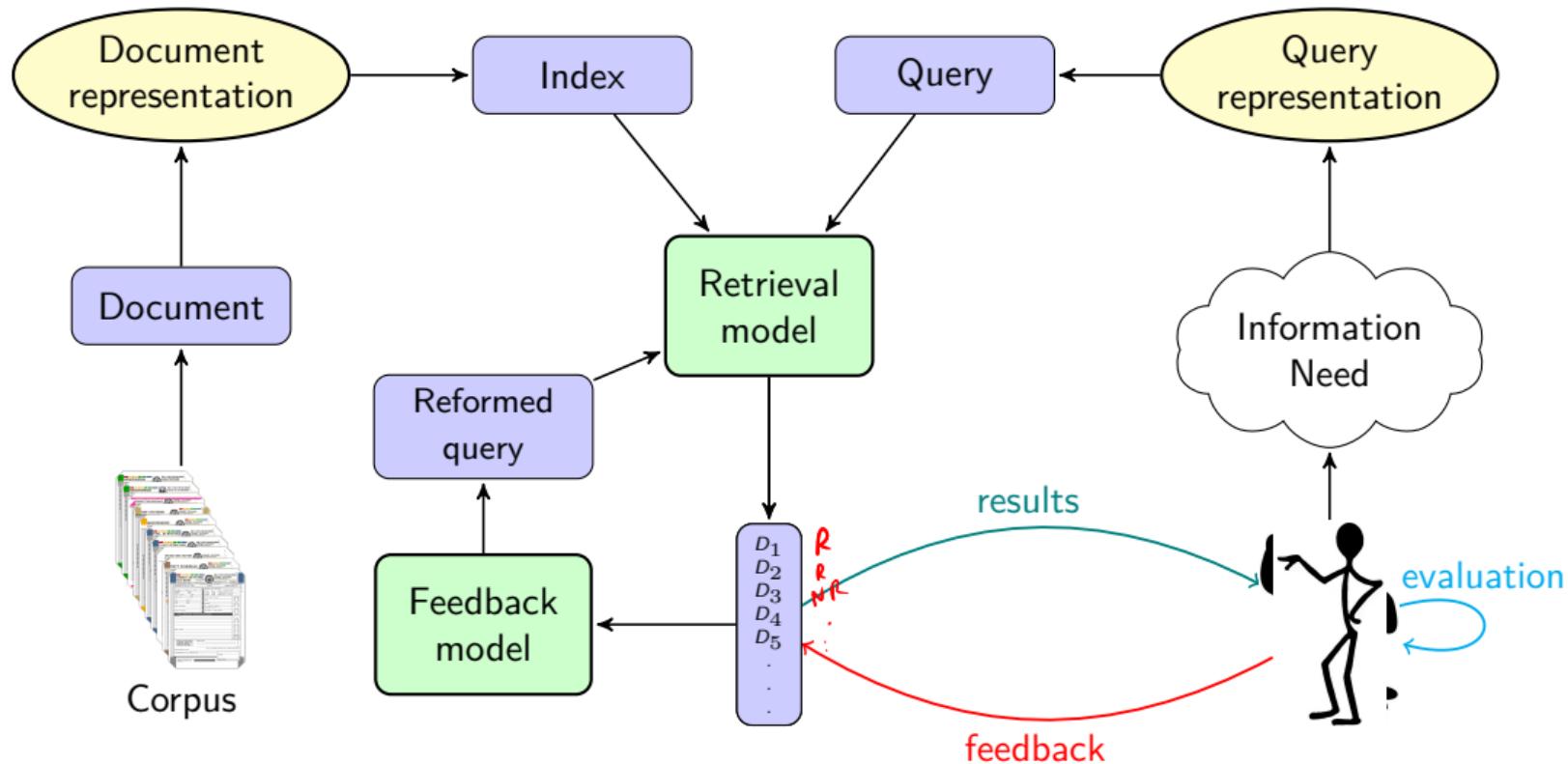
Relevance Feedback: A Graphical Representation



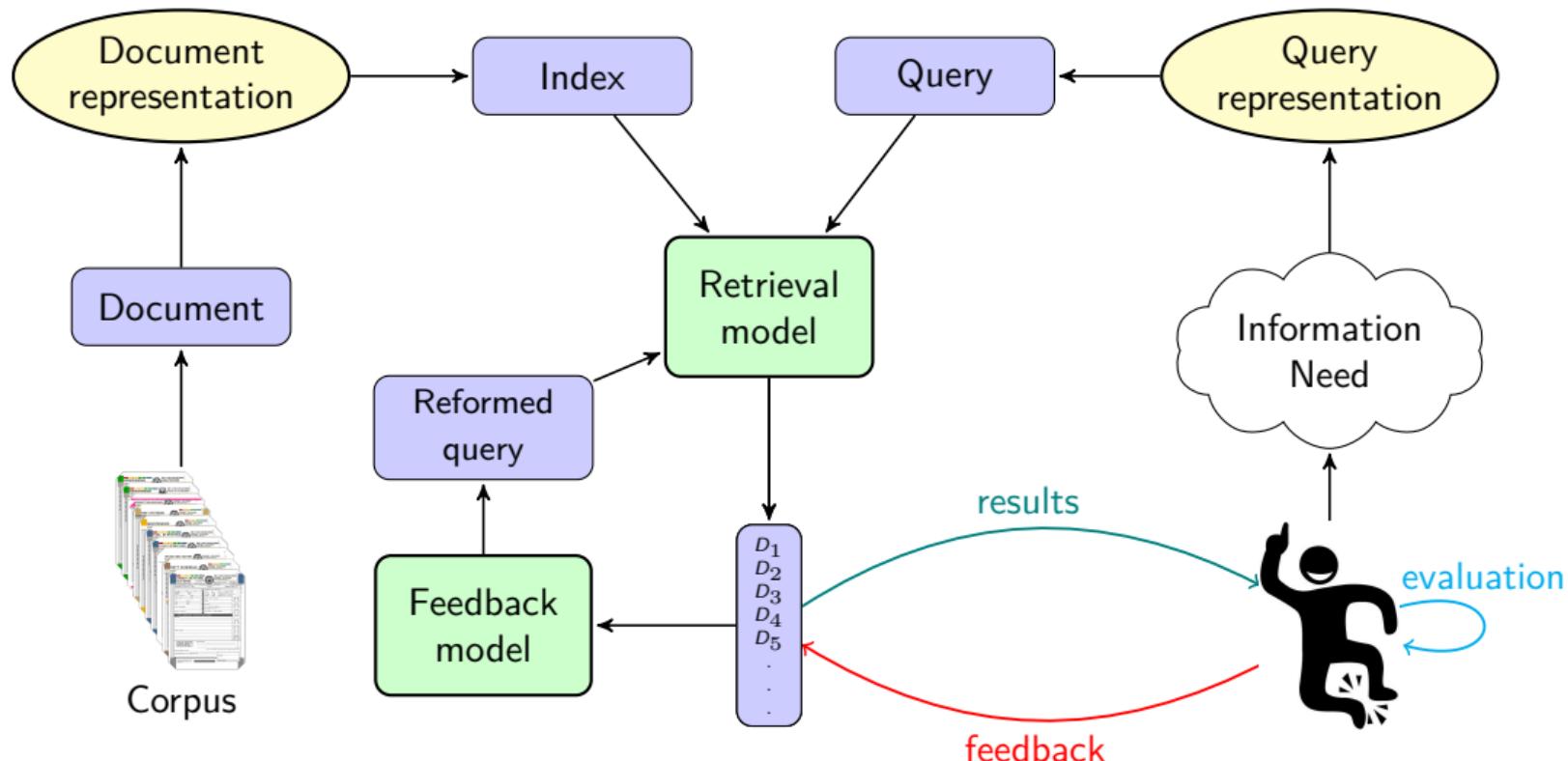
Relevance Feedback: A Graphical Representation



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

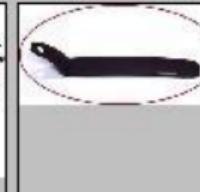
- Initial Query: bike

bicycle



Relevance Feedback: Example 1

- Results for initial query

						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: Example 1

- Feedback: Relevant document selected

Handwritten annotations in red:

- Top-left image: 100/100
- Second row, first column: 4/12
- Bottom-left image: 4/100

		Browse	Search	Prev	Next	Random
						
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)	
0.0	0.0	0.0	0.0	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	0.0	
						
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)	
0.0	0.0	0.0	0.0	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	0.0	

Relevance Feedback: Example 1

- Result after re-retrieval

						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
<p>11 / 100 Recall ↑ precision ↗</p> <p>X</p>											

Relevance Feedback: Example 2

- TREC 2 Topic - 113 : New Space Satellite Applications

1993
101 → 150

1992 → TREC1 51 → 100

TREC6 301 → 350

Topic



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*

4 term

- Expanded query after relevance feedback

weight	term	weight	term
2.074	<u>new</u>	15.106	<u>space</u>
<u>30.816</u>	<u>satellite</u>	<u>5.660</u>	<u>application</u>
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

why?

18
//

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

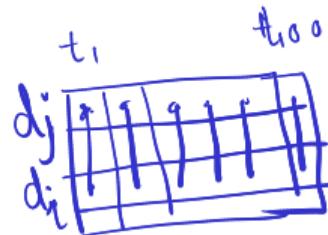
Relevance Feedback: Example 2

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

Old	Rank	Score	Document title
2	<u>1</u>	0.513	NASA Scratches Environment Gear From Satellite Plan
1	<u>2</u>	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

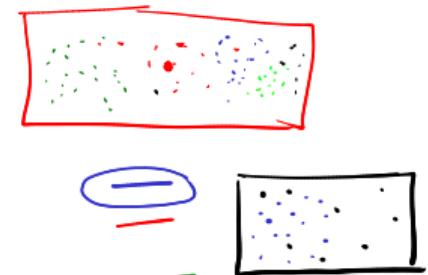
- The centroid is the center of mass of a set of points.
- Recall that we represent documents as points in a high-dimensional space.
- Thus: we can compute centroids of documents.
- The centroid of a set of documents D :



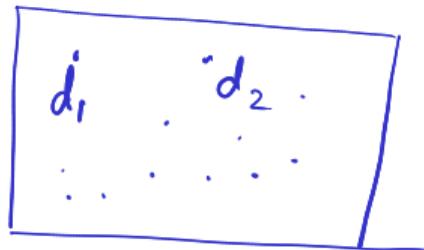
$$\mu(D) = \frac{1}{|D|} \sum_{d \in D} d$$

$\text{vec}(d_i) = 100\text{dim.}$

$$\text{vec}(d_i + d_j) = \begin{bmatrix} \vdots & \vdots & \vdots \\ \vdots & \times & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix}_2$$



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

sim(Q, centroid of NonRel)

- Chooses the optimal query \mathbf{Q}_{opt} that maximizes:

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

sim(Q, centroid of Rel doc)

sim(Q, centroid of NonRel)

The diagram shows the mathematical expression for the optimal query \mathbf{Q}_{opt} . A blue bracket underlines the term $\operatorname{argmax}_{\mathbf{Q}}$. Another blue bracket underlines the entire expression $sim(\mathbf{Q}, \mu(D_R)) - sim(\mathbf{Q}, \mu(D_{NR}))$. Above the first bracket, the word "Rel" is written above the term $sim(\mathbf{Q}, \mu(D_R))$, and above the second bracket, the words "NonRel" are written above the term $sim(\mathbf{Q}, \mu(D_{NR}))$. Below the expression, the text "sim(Q, centroid of Rel doc)" is written next to the first bracket, and "sim(Q, centroid of NonRel)" is written next to the second bracket.

- 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

Q

- Chooses the optimal query \mathbf{Q}_{opt} that maximizes:

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

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

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

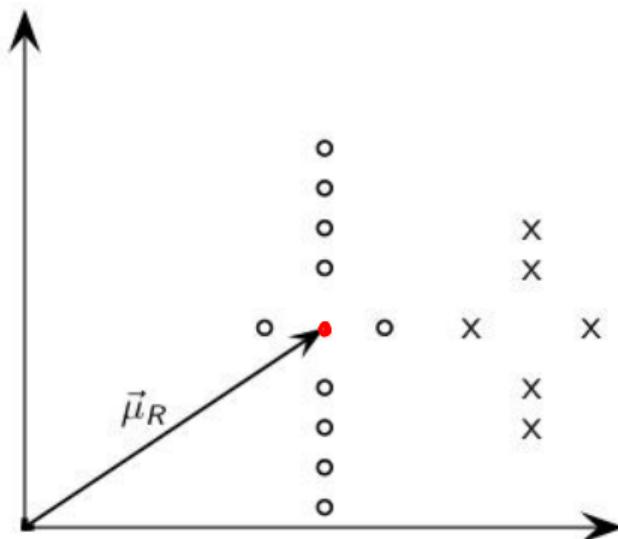
Rocchio's Algorithm Illustrated



○ relevant documents

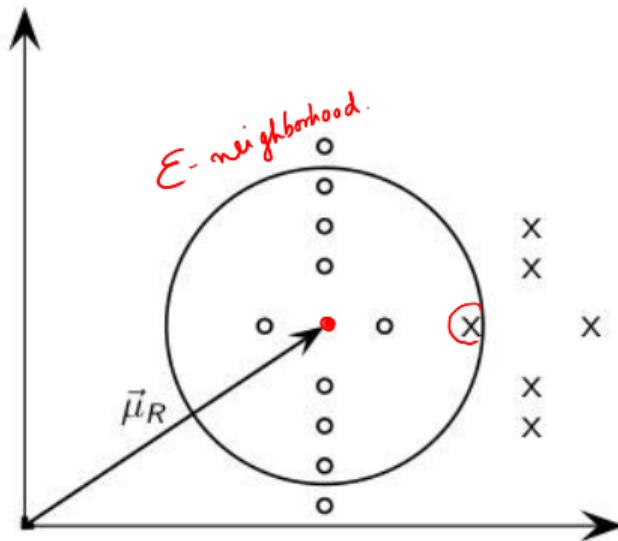
× non-relevant documents

Rocchio's Algorithm Illustrated



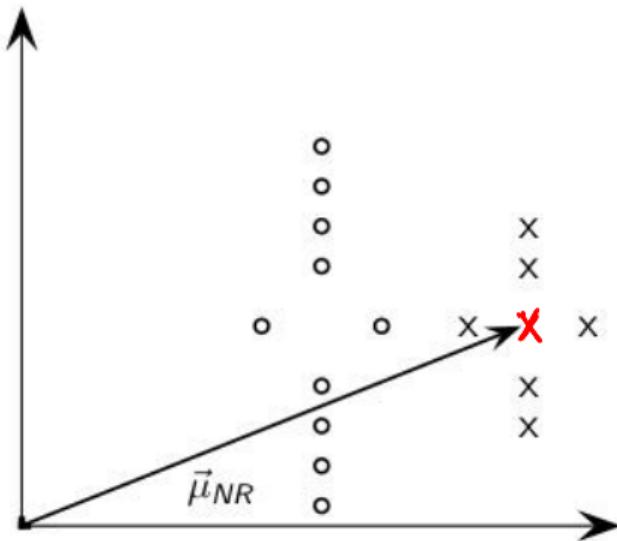
$\vec{\mu}_R$: centroid of relevant documents

Rocchio's Algorithm Illustrated



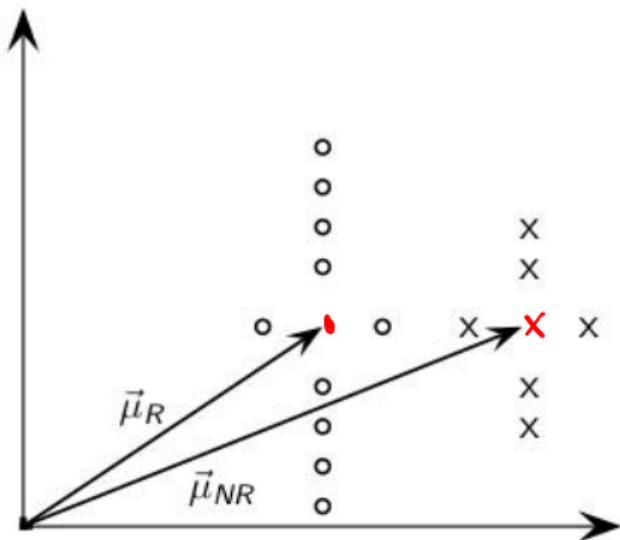
$\vec{\mu}_R$ not separating relevant documents from non-relevant

Rocchio's Algorithm Illustrated

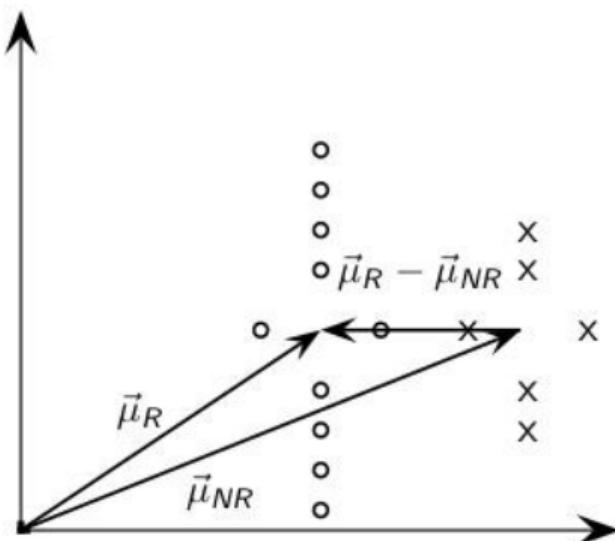


μ_{NR} : centroid of non-relevant documents

Rocchio's Algorithm Illustrated

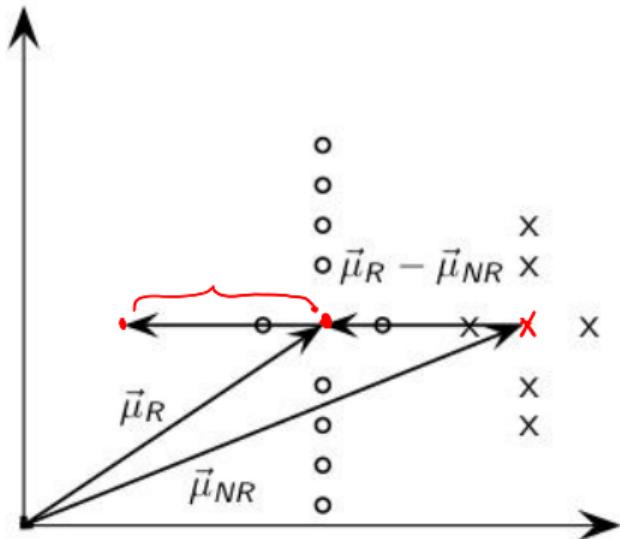


Rocchio's Algorithm Illustrated

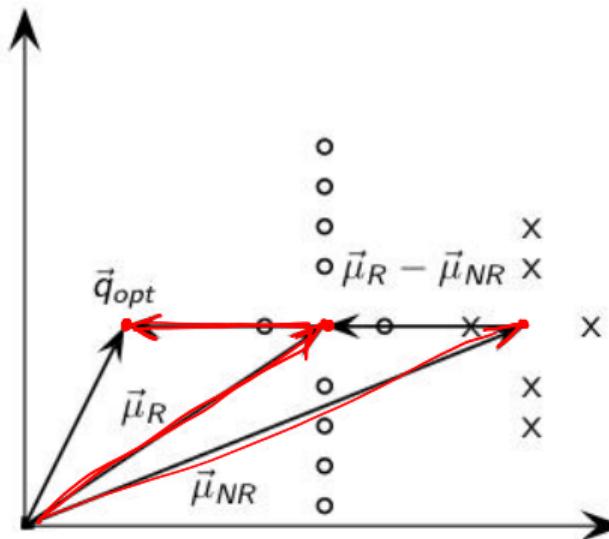


μ_R – μ_{NR} : difference vector

Rocchio's Algorithm Illustrated

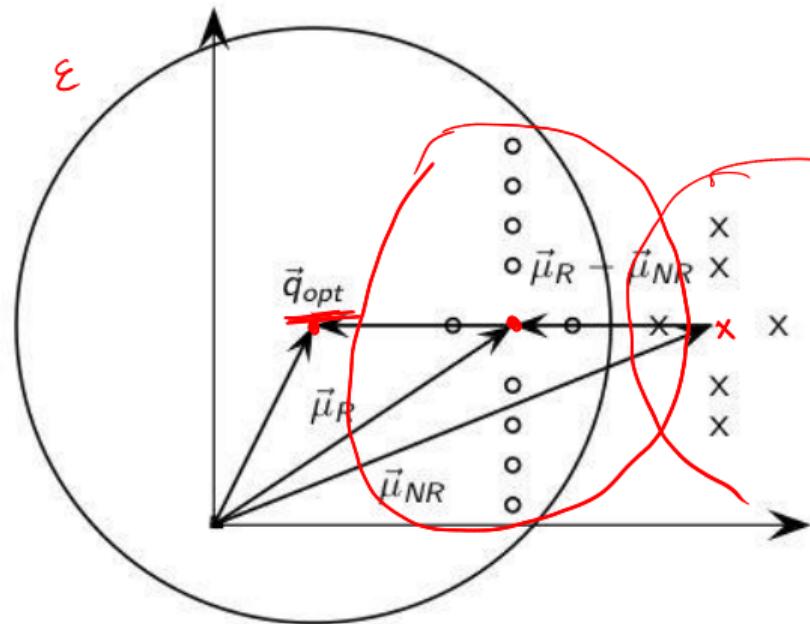


Rocchio's Algorithm Illustrated



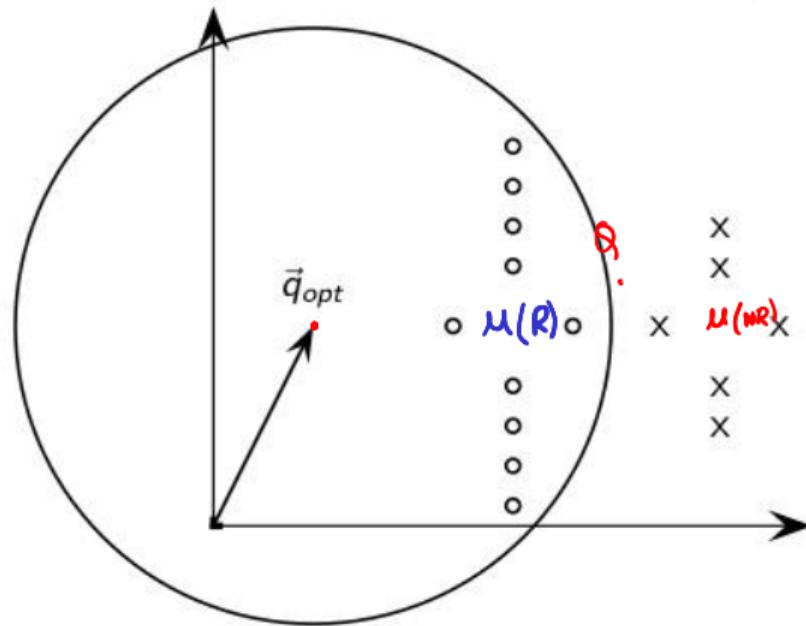
Add difference vector to μ_R to get μ_{opt}

Rocchio's Algorithm Illustrated



Rocchio's Algorithm Illustrated

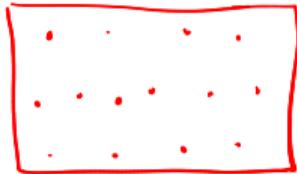
vsm



$\underline{\underline{Q_{opt}}}$ separates relevant and non-relevant documents perfectly

Rocchio's Algorithm for Relevance Feedback

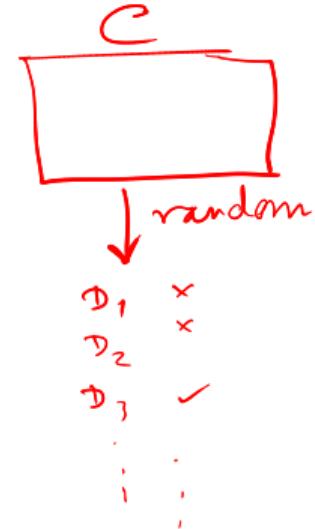
- The optimal query vector:



\underline{Q} $\xrightarrow{\text{init. ret.}}$
 $M(M) \Rightarrow \text{centroid of } M.$

$$Q_{opt} = \overrightarrow{\mu(D_R)} + \left[\overrightarrow{\mu(D_R)} - \overrightarrow{\mu(D_{NR})} \right]$$

✓ $\rightarrow D_R$
✗ $\rightarrow D_{NP}$
✓ $\rightarrow D_R$
✗ } $\rightarrow D_{NR}$



Rocchio's Algorithm for Relevance Feedback

- The optimal query vector:

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

Rocchio's Algorithm for Relevance Feedback

$$Q = \{ D_R \} \cup D_{NR}$$



- The optimal query vector:

$$\begin{aligned} Q_{opt} &= \mu(D_R) + [\mu(D_R) - \mu(D_{NR})] \\ &= \underbrace{\frac{1}{|D_R|} \sum_{d_i \in D_R} d_i}_{\text{set of rel. docr. of } D_R} + \underbrace{\left[\frac{1}{|D_R|} \sum_{d_i \in D_R} d_i - \frac{1}{|D_{NR}|} \sum_{d_j \in D_{NR}} d_j \right]}_{\text{set of rel. docr. of } Q} \end{aligned}$$

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

Rocchio's Algorithm

parameters

- In practice:

$$Q_m = \underbrace{\alpha Q_o}_{\text{original } Q} + \underbrace{\beta \frac{1}{|D_R|} \sum_{d_i \in D_R} d_i}_{\text{Rel. Docn}} - \underbrace{\gamma \frac{1}{|D_{NR}|} \sum_{d_j \in D_{NR}} d_j}_{\text{Non Rel. Docn.}}$$

tune

Rocchio's Algorithm

- In practice:

keyword query
 \downarrow
 $Q = \{q_1, q_2\}$

$\underline{Q_0}$: original query

$\underline{Q_m}$: modified query

Relevance feedback
 $\underline{D_R}$: set of known relevant documents

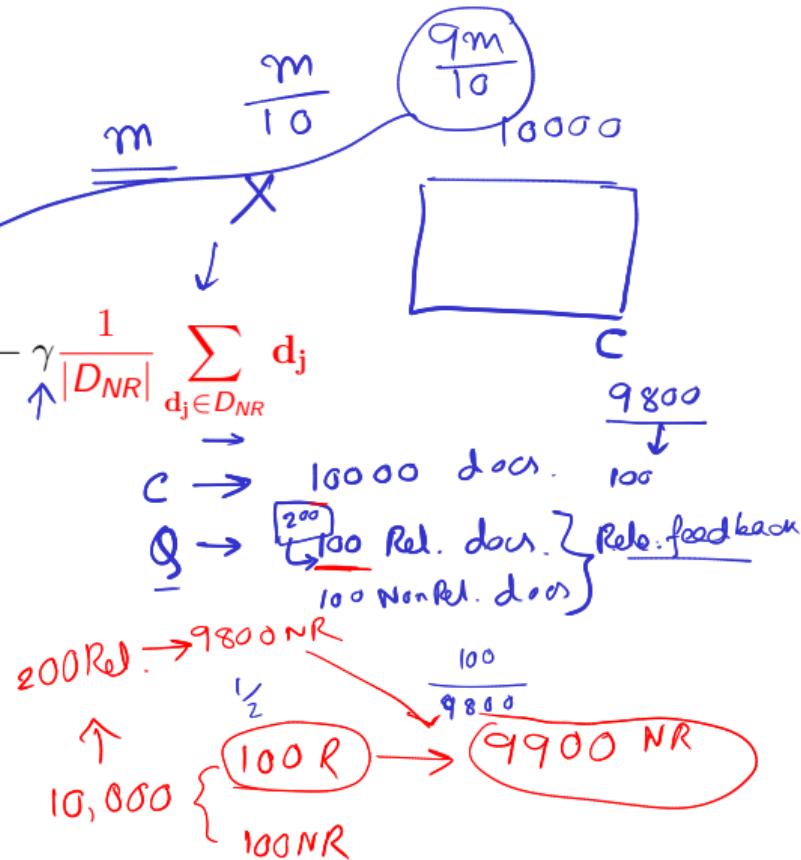
$\underline{D_{NR}}$: set of known non-relevant documents

$\underline{\alpha, \beta, \gamma}$: weights (belief) of corresponding sets

$$Q_m = \vec{Q}_0 + \beta \frac{1}{|D_R|} \sum_{d_i \in D_R} d_i - \gamma \frac{1}{|D_{NR}|} \sum_{d_j \in D_{NR}} d_j$$

\vec{Q}_0

\vec{Q}_m



Rocchio's Algorithm: Parameter Settings

$$\underline{\mathbf{Q}_m} = \alpha \mathbf{Q}_o + \beta \frac{1}{|D_R|} \sum_{\mathbf{d}_i \in D_R} \mathbf{d}_i - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d}_j \in D_{NR}} \mathbf{d}_j$$

New (modified) query

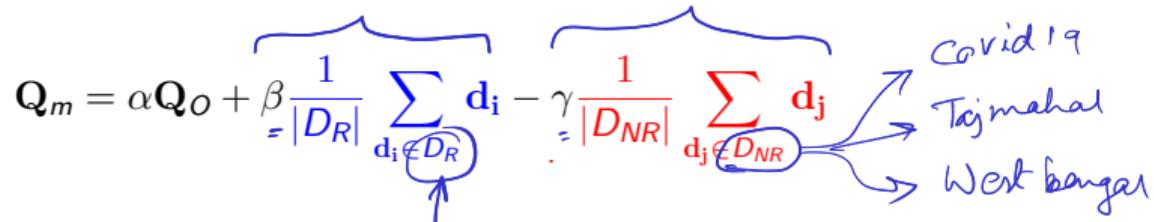
- Set higher β, γ if there is a lot of judged documents.

Rocchio's Algorithm: Parameter Settings

$Q \rightarrow$ Rel. doc. on the same topic

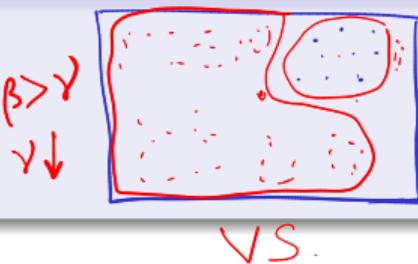
$$Q_m = \alpha Q_0 + \beta \frac{1}{|D_R|} \sum_{d_i \in D_R} d_i - \gamma \frac{1}{|D_{NR}|} \sum_{d_j \in D_{NR}} d_j$$

C = Wikipedia
Q = "German Football team"



New (modified) query

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



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

$M = 6$

$$Q_m = \underbrace{\alpha Q_o}_{\text{Original query}} + \beta \underbrace{\frac{1}{|D_R|} \sum_{d_i \in D_R} d_i}_{\mathcal{D}_R} - \gamma \underbrace{\frac{1}{|D_{NR}|} \sum_{d_j \in D_{NR}} d_j}_{\mathcal{D}_{NR}}$$

Original query

	t_1	t_2	t_3	t_4	t_5	t_6
	0	4	0	8	0	0

$$Q_o = \{t_2, t_4, t_5\}$$

Positive feedback \rightarrow

	t_1	t_2	t_3	t_4	t_5	t_6
	2	4	8	0	0	2

Negative feedback \rightarrow

	t_1	t_2	t_3	t_4	t_5	t_6
	8	0	4	4	0	16

$$\alpha = 1.0$$

0	4	0	8	0	0
---	---	---	---	---	---

$$Q_o = \alpha Q_o$$

+

$$\beta = 0.5$$

1	2	4	0	0	1.
---	---	---	---	---	----

\mathcal{D}_R

-

$$\gamma = 0.25$$

2	0	1	1	0	4
---	---	---	---	---	---

\mathcal{D}_{NR}

-1	6	3	7	0	-3
----	---	---	---	---	----

Rocchio's Algorithm: Example

$$|V| = M$$

$$\underline{\underline{Q_m}} = \alpha Q_O + \beta \frac{1}{|D_R|} \sum_{d_i \in D_R} d_i - \gamma \frac{1}{|D_{NR}|} \sum_{d_j \in D_{NR}} d_j$$

$\underline{\underline{Q_O}}$

Original query $\begin{bmatrix} 0 & 4 & 0 & 8 & 0 & 0 \end{bmatrix}$

$$\underline{\underline{Q_O}} = \{t_2, t_4\}$$

$\alpha = 1.0 \rightarrow \begin{bmatrix} 0 & 4 & 0 & 8 & 0 & 0 \end{bmatrix}$

Positive feedback $\rightarrow \begin{bmatrix} 2 & 4 & 8 & 0 & 0 & 2 \end{bmatrix}$

$\beta = 0.5 \rightarrow \begin{bmatrix} 1 & 2 & 4 & 0 & 0 & 1 \end{bmatrix}$

Negative feedback $\rightarrow \begin{bmatrix} 8 & 0 & 4 & 4 & 0 & 16 \end{bmatrix}$

$\gamma = 0.25 \rightarrow \begin{bmatrix} 2 & 0 & 1 & 1 & 0 & 4 \end{bmatrix}$

expanded query

$$\underline{\underline{Q_m}} = \{t_2, t_4, \underline{\underline{t_3}}\}$$

New query

$t_1 \ t_2 \ t_3 \ t_4 \ t_5 \ t_6$

$\begin{bmatrix} -1 & 6 & 3 & 7 & 0 & -3 \\ 0 & 6 & 3 & 7 & 0 & 0 \end{bmatrix}$

↑ ↑ ↑ new term

$t_i \quad t_j$

5 8

3 5 8

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

! Red. X

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 → implicit 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 → 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 → 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

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

- 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

$$\mathbf{Q}_m = \alpha \mathbf{Q}_o + \beta \frac{1}{|D_R|} \sum_{\mathbf{d}_i \in D_R} \mathbf{d}_i - \gamma \frac{1}{|D_{NR}|} \sum_{\mathbf{d}_j \in D_{NR}} \mathbf{d}_j$$

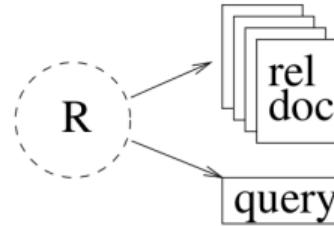
- Perform a retrieval.
- Consider top K ranked documents as **relevant** - D_R .
- Which documents to be selected as **non-relevant** - D_{NR} ?
- Only consider D_R and set γ to 0.

Relevance based Language Model (RLM)

Lavrenko and Croft - SIGIR 2001

Relevance based Language Model (RLM)

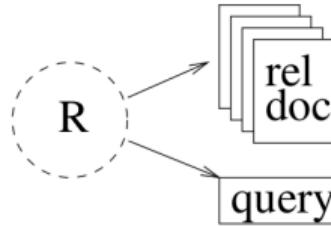
Lavrenko and Croft - SIGIR 2001



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

Relevance based Language Model (RLM)

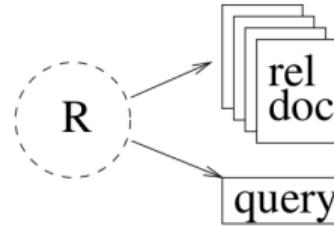
Lavrenko and Croft - SIGIR 2001



- 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)

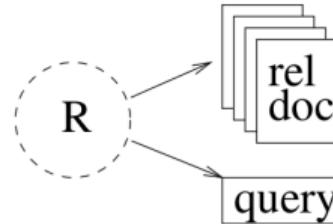
Lavrenko and Croft - SIGIR 2001



- 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 .