https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018

Indian Institute of Information Technology, Sri City

Searchers may not have a well-developed idea of what information they are searching for, they may not be able to express their conceptual idea of what information they want into a suitable query and they may not have a good idea of what information is available for retrieval. **– Ruthven and Lalmas, The Knowledge Engineering Review, 2003.**

Relevance Feedback

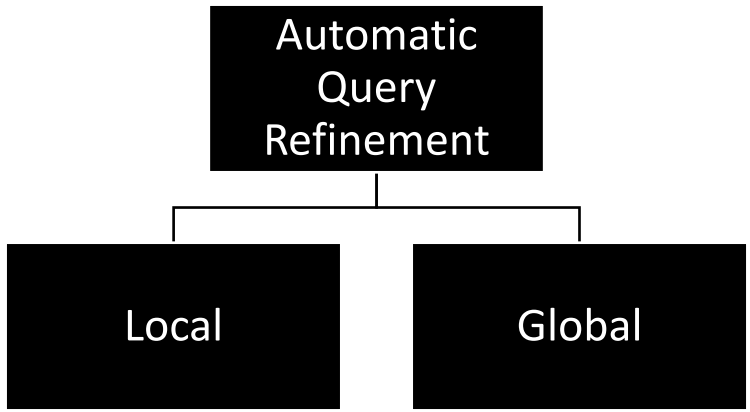
How to improve relevance?

Relevance Feedback
And
Query Expansion

The Problem of Synonymy

- What result do you expect for a query, “plane”?
- What if plane appears in this query, “plane from Delhi to Goa”?
- So many synonyms which will work for web search…
  - Flight
  - Aircraft
  - Airplane
  - Aeroplane
  - By Air
  - Fly
  - Flgt
  - Arcrft

How to ensure good results?



<u>Use</u> the **query** or the **results** for reformulating the query

We will study:

Relevance Feedback

Pseudorelevance

Indirect Relevance Feedback

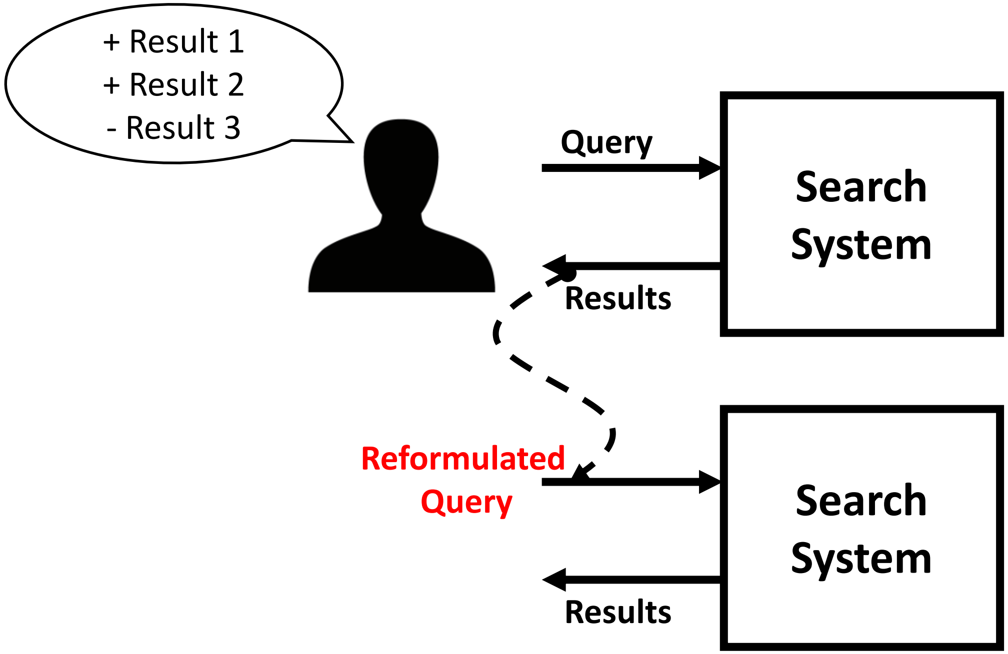
<u>Do not use</u> the **query** or the **results** for reformulating the query.

Eg:

Use *Thesaurus.*

Do Spelling Correction.

Relevance Feedback



An Example

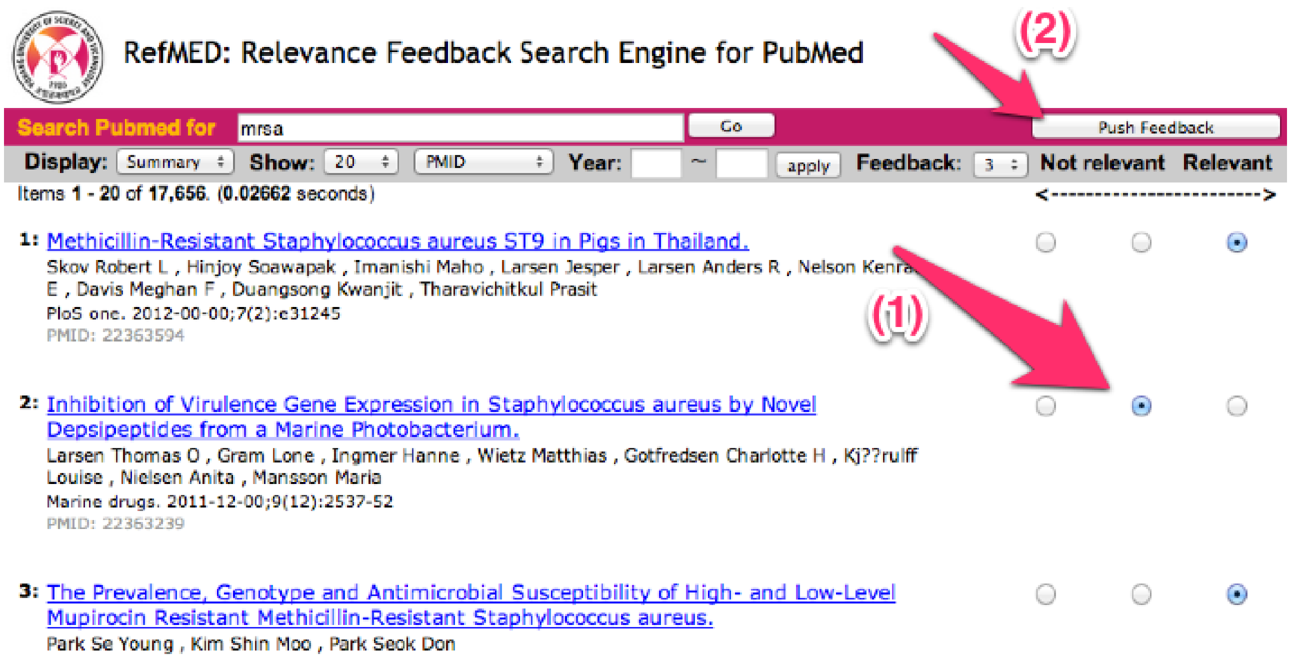
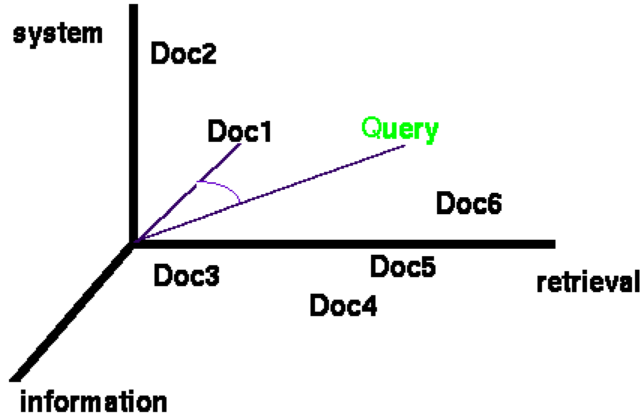


Image source: https://sites.google.com/site/postechdm/research

Interesting Characteristics

- Indexed content is unknown to the user.
- “Information Need” changes after looking at the results.
  - User visits youtube to listen to a specific set of songs.
  - After the first song, he changes his mind and listens to something else!

A Recap of Vector Space Models



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Image Source: https://fox.cs.vt.edu/talks/1995/KY95/

Rocchio Algorithm for Relevance Feedback

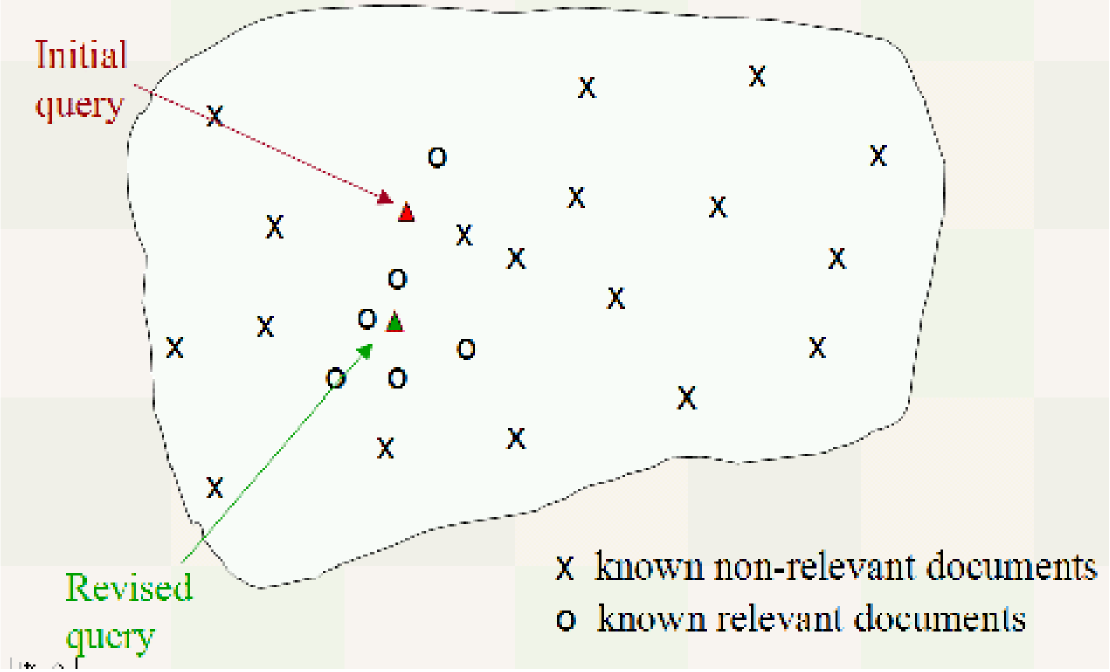


Image Source: https://nlp.stanford.edu/IR-book/

Moving the Centroid!

Modify the query (and therefore, the query vector from q0 to qm):

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

D_r = Set of known relevant documents
D_{nr} = Set of known nonrelevant documents
q_o = Initial query vector
q_m = Modified query vector

Rocchio relevance feedback - Example

- Given:
  - Initial query = “cheap CDs cheap DVDs extremely cheap CDs”.
  - d_{1} = “CDs cheap software cheap CDs” is judged as relevant.
  - d_{2} = “cheap thrills DVDs” is judged as nonrelevant
- What would the revised query vector be after relevance feedback?

Let us solve this together

Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Assume α = 1, β = 0.75, γ = 0.25.

Representing Initial Query in Vector Space

Initial query = "cheap CDs cheap DVDs extremely cheap CDs".

	cheap	CDs	DVDs	extremely	software	thrills
q_0	3	2	1	1	0	0

Rocchio relevance feedback - Example

Quiz: Can you complete the following table?

q_0 = "cheap CDs cheap DVDs extremely cheap CDs".

d_1 = "CDs cheap software cheap CDs".

d_2 = "cheap thrills DVDs".

	cheap	CDs	DVDs	extremely	software	thrills
q_0	3	2	1	1	0	0
d_1						
d_2						

Rocchio relevance feedback - Example

Quiz: Can you complete the following table?

q_0 = "cheap CDs cheap DVDs extremely cheap CDs".

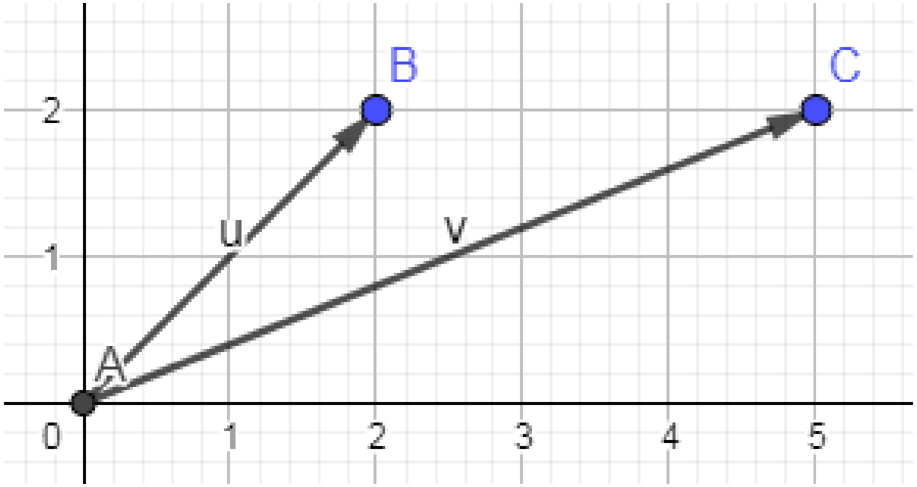
d_1 = "CDs cheap software cheap CDs".

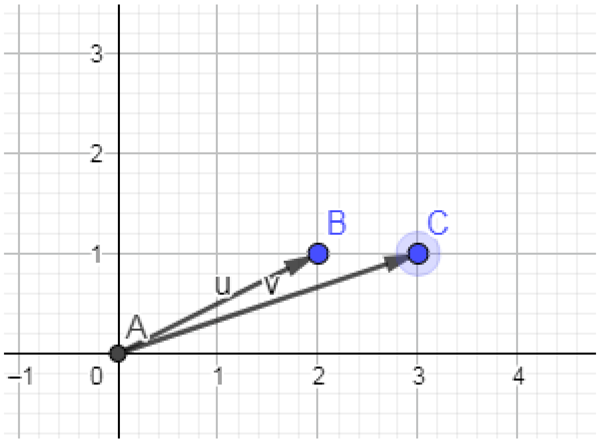
d_2 = "cheap thrills DVDs".

	cheap	CDs	DVDs	extremely	software	thrills
q_0	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1

Moving Vectors

- Move (2,2) to (5,2) by adding 3 to x.





Rocchio relevance feedback - Example

Quiz: How to calculate the modified query vector, q_m?

d_1 is judged as **relevant**. d_2 is judged as **non-relevant**.

Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

	cheap	CDs	DVDs	extremely	software	thrills
q_0	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1

q_m						

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

Rocchio relevance feedback - Example

Quiz: How to calculate the modified query vector, q_m?

d_1 is judged as relevant. d_2 is judged as nonrelevant.
Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

	cheap	CDs	DVDs	extremely	software	thrills
q_0	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1

$$q_m = q_0 + 0.75*d_1 - 0.25*d_2$$

q_m	4.25	3.5	0.75	1	0.75	0

Negative weight does not make sense. So, leave them as zero.

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

Pseudo (Blind) Relevance Feedback

- No User Judgment.
- Assume that the top-k ranked documents are relevant.

Initial query = “cheap CDs cheap DVDs extremely cheap CDs”.
$d_{1} =$ “CDs cheap software cheap CDs”.
$d_{2} =$ “cheap thrills DVDs”.
What would the revised query vector be after pseudo relevance feedback if top-1 document is considered as relevant?
Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Assume α = 1, β = 0.75, γ = 0.25.

- May lead to <u>query drift</u>.

Indirect (Implicit) Relevance Feedback

- No asking for judgments from users.
- No automatic feedback such as assuming top-k documents as relevant.

Clickstream Mining

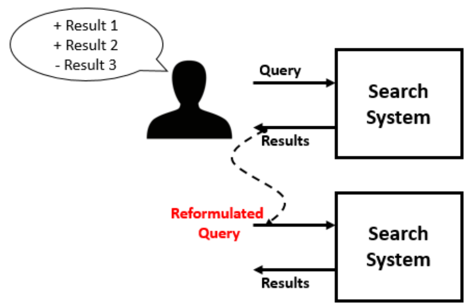
Recap

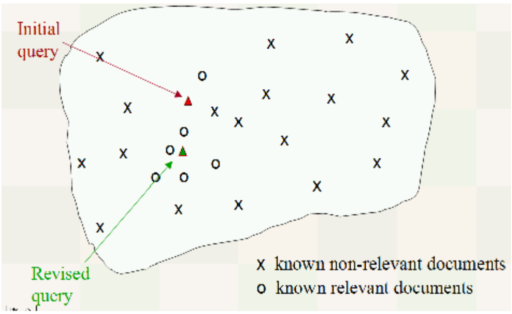
Automatic Query Refinement

Local | Global

Dealing with Local Query Refinement

Relevance Feedback





	cheap	CDs	DVDs	extremely	software	thrills
q_0	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1
$q_m = q_0 + 0.75*d_1 - 0.25*d_2$						
q_m	4.25	3.5	0.75	1	0.75	0

Negative weight does not make sense. So, leave them as zero.

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

Pseudo (Blind) Relevance Feedback

Assume top-k is relevant.

Indirect (Implicit) Relevance Feedback

Using clickstreams, query logs, etc.

Global (User/Result-Independent) Query Refinement

- Automatic Thesaurus Generation
  - Fast = rapid
  - Tall = height?
  - Sound = noise?
  - Restaurant = Hotel = Motel?
- How to handle domain specific phrases?
- Slangs!
- …

How to automate the thesaurus generation?

Co-occurrence Analysis

MAINFRAMES

Mainframes are primarily referred to large computers with rapid, advanced processing capabilities that can execute and perform tasks equivalent to many Personal Computers (PCs) machines networked together. It is characterized with high quantity Random Access Memory (RAM), very large secondary storage devices, and high-speed processors to cater for the needs of the computers under its service.

Consisting of advanced components, mainframes have the capability of

MAINFRAMES

Mainframes usually are referred those computers with fast, advanced processing capabilities that could perform by itself tasks that may require a lot of Personal Computers (PC) Machines. Usually mainframes would have lots of RAMs, very large secondary storage devices, and very fast processors to cater for the needs of those computers under its service.

Due to the advanced components mainframes have, these computers

Source: https://web.stanford.edu/~jurafsky

Co-occurrence Analysis

- Term-Document Matrix
  - How often does individual terms appear in a document?
- Term-Term Matrix
  - How often terms co-occur?

Quiz: Which books are similar?

	Book1	Book2	Book3	Book4
cricket	400	10	355	3
football	5	5	4	4
hockey	9	330	10	200
tennis	2	6	12	4

Co-occurrence Analysis

- Two documents are similar if the document vectors are similar.

	Book1	Book2	Book3	Book4
cricket	400	10	355	3
football	5	5	4	4
hockey	9	330	10	200
tennis	2	6	12	4

Book1 and Book3 seem to be on cricket.
Book2 and Book4 are about hockey.

Co-occurrence Analysis

- Two terms are similar if the term vectors are similar.

	Book1	Book2	Book3	Book4
boundary	400	310	355	389
four	515	225	390	400
movie	9	4	8	1
film	2	6	9	2

Remember, context is important!

Magic with Matrices

Transpose

- If A is as given below, what is A^T?

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \qquad \mathbf{A}^{\mathbf{T}} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

Co-occurrence Analysis

- Assume a Boolean term-document matrix A.
- What does AA^T mean?

$$
\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array}
\begin{array}{c} d_1\ d_2\ d_3\ d_4 \\ \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \end{array}
\begin{array}{c} d_1\ d_2\ d_3\ d_4 \\ \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \end{array}
=
\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array}
\begin{array}{c} t_1\ t_2\ t_3\ t_4 \\ \begin{pmatrix} 2 & 2 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 2 \end{pmatrix} \end{array}
$$

- Usually, weighted length-normalized tf in a sliding window is used to count co-occurrence.

A =

	D1	D2	D3	D4	D5	d6
T1	1	0	1	0	1	0
T2	1	1	0	1	0	0
T3	0	1	1	0	1	0
T4	0	1	0	0	1	0
T5	1	0	0	1	1	1
T6	1	0	1	0	1	0

Terms 1 & 6 appear in the same documents

A^T =

	T1	T2	T3	T4	T5	T6
D1	1	1	0	0	1	1
D2	0	1	1	1	0	0
D3	1	0	1	0	0	1
D4	0	1	0	0	1	0
D5	1	0	1	1	1	1
D6	0	0	0	0	1	0

AA^T =

	T1	T2	T3	T4	T5	T6
T1	3	1	2	1	2	3
T2	1	3	1	1	2	1
T3	2	1	3	2	1	2
T4	1	1	2	2	1	1
T5	2	2	1	1	4	2
T6	3	1	2	1	2	3

Term 6 seems to be best related to itself and term 1.

Thank You