https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018
Indian Institute of Information Technology, Sri City

Characteristic vectors representing code are often high dimensional, *e.g.*, there are more than two hundreds different syntactic element types in Java ASTs..

– Lee et al. From their paper, "Instant Code Clone Search".

Relevance Feedback

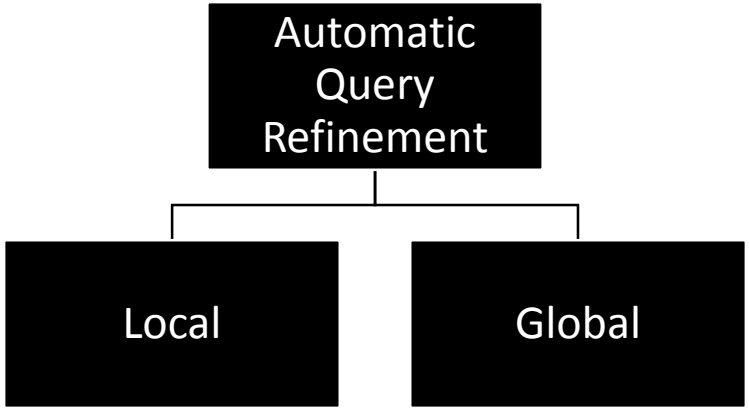
How to improve relevance?

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?



Use the query or the results for reformulating the query

We will study:

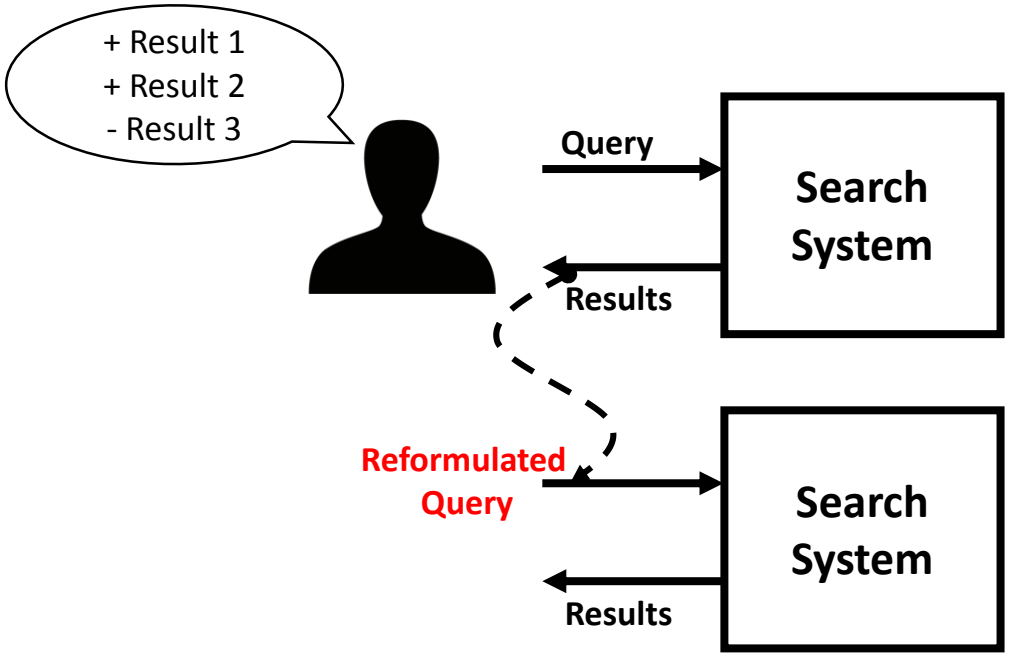
- *Relevance Feedback*
- *Pseudorelevance*
- *Indirect Relevance Feedback*

Do not use the query or the results for reformulating the query.

Eg:

- Use *Thesaurus.*
- *Do Spelling Correction.*

Relevance Feedback



An Example

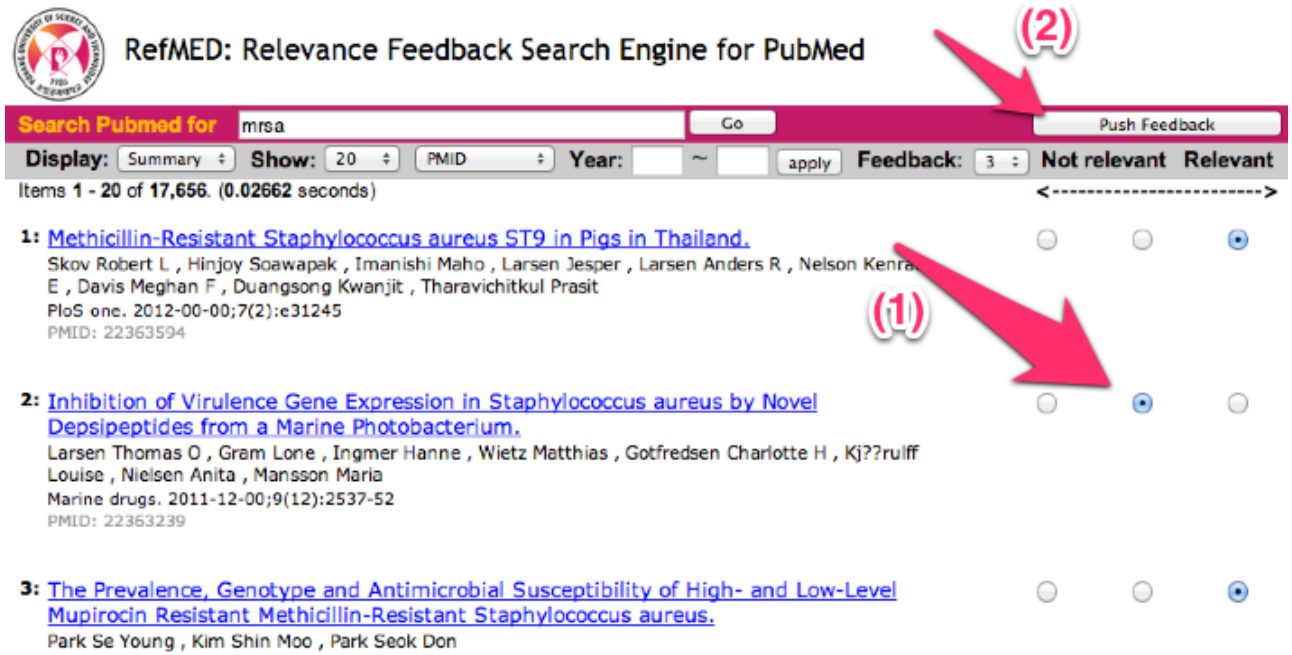
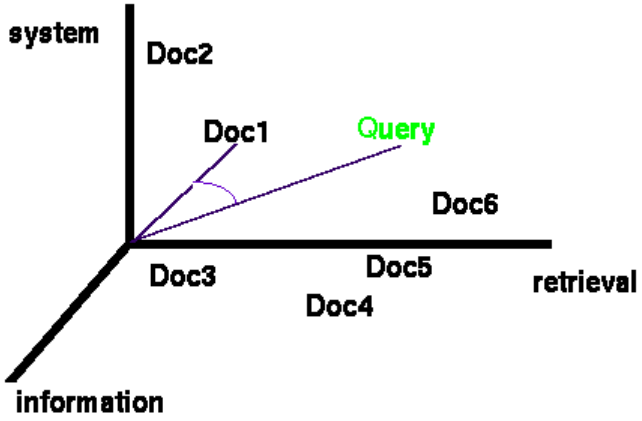


Image source: https://sites.google.com/site/postechdm/research

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

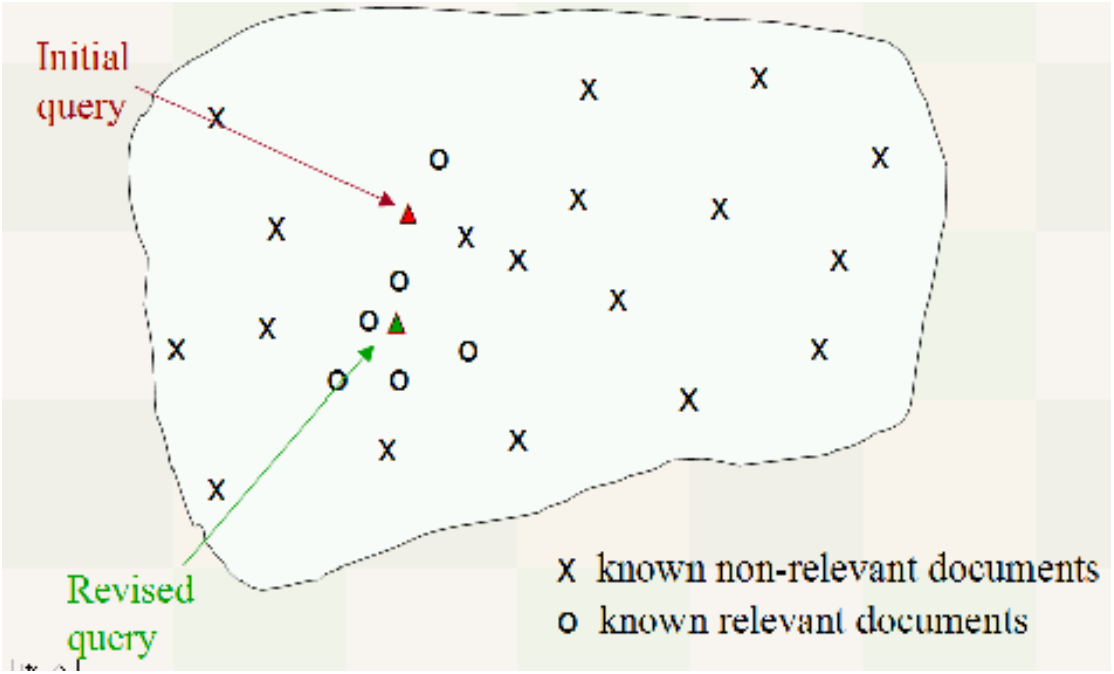


Image Source: https://nlp.stanford.edu/IR-book/

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

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".

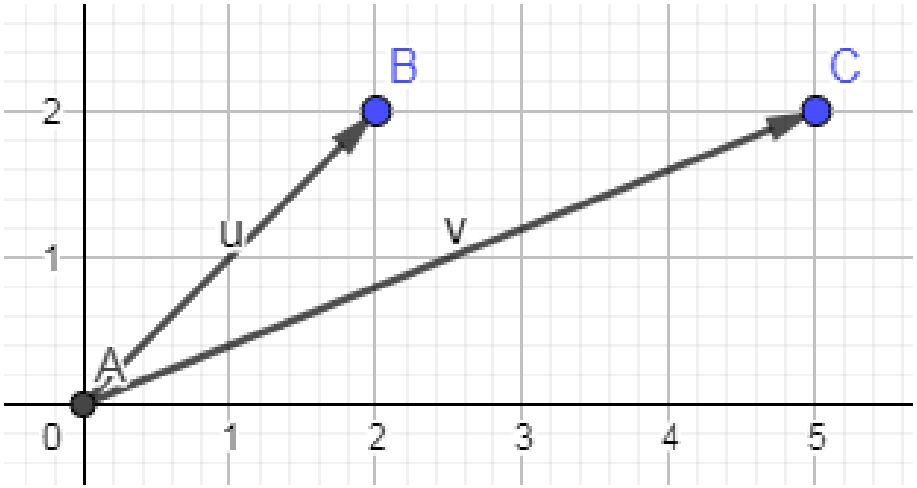
d_1 = "CDs cheap software cheap CDs".

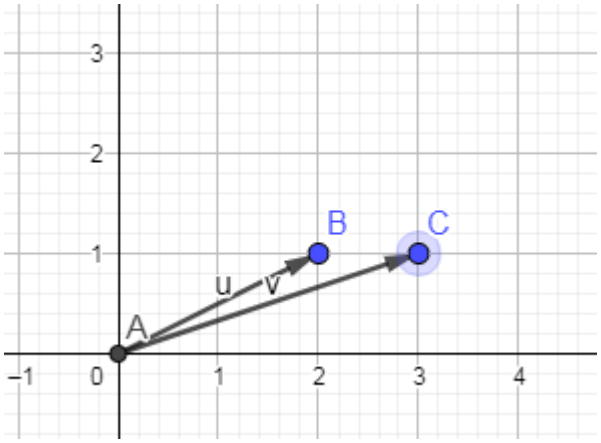
d_2 = "cheap thrills DVDs".

	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.

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