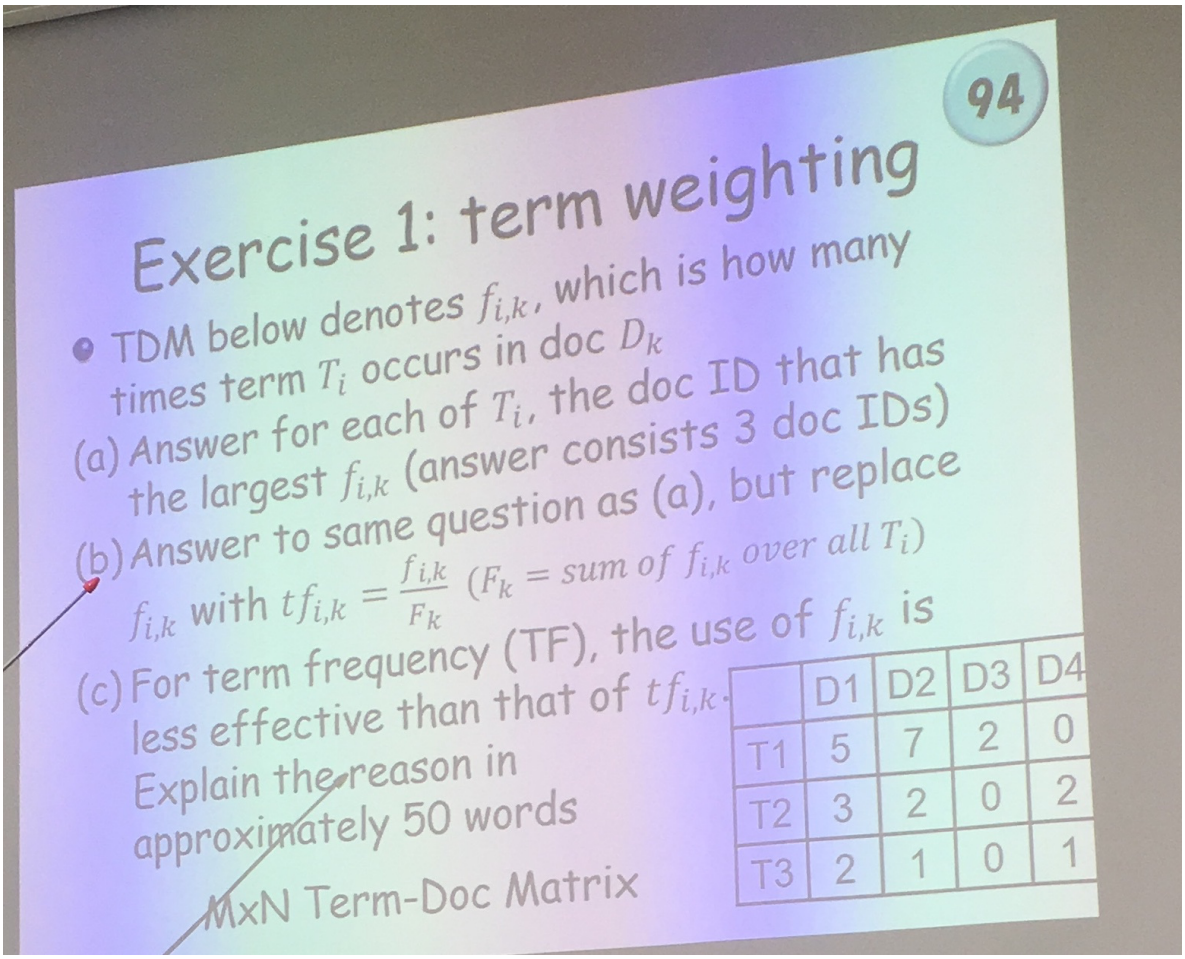
Day 8



(a) T1: D2; T2: D1; T3: D1

(b) tf1,1 = 5/14; tf1,2 = 7/14 = 1/2; tf1,3 = 2/14 = 1/7; tf1,4 = 0
tf2,1 = 3/7; tf2,2 = 2/7 ; tf2,3 = 0; tf2,4 = 2/7
tf3,1 = 2/4 = 1/2; tf3,2 = 1/4; tf3,3 = 0; tf3,4 = 1/4

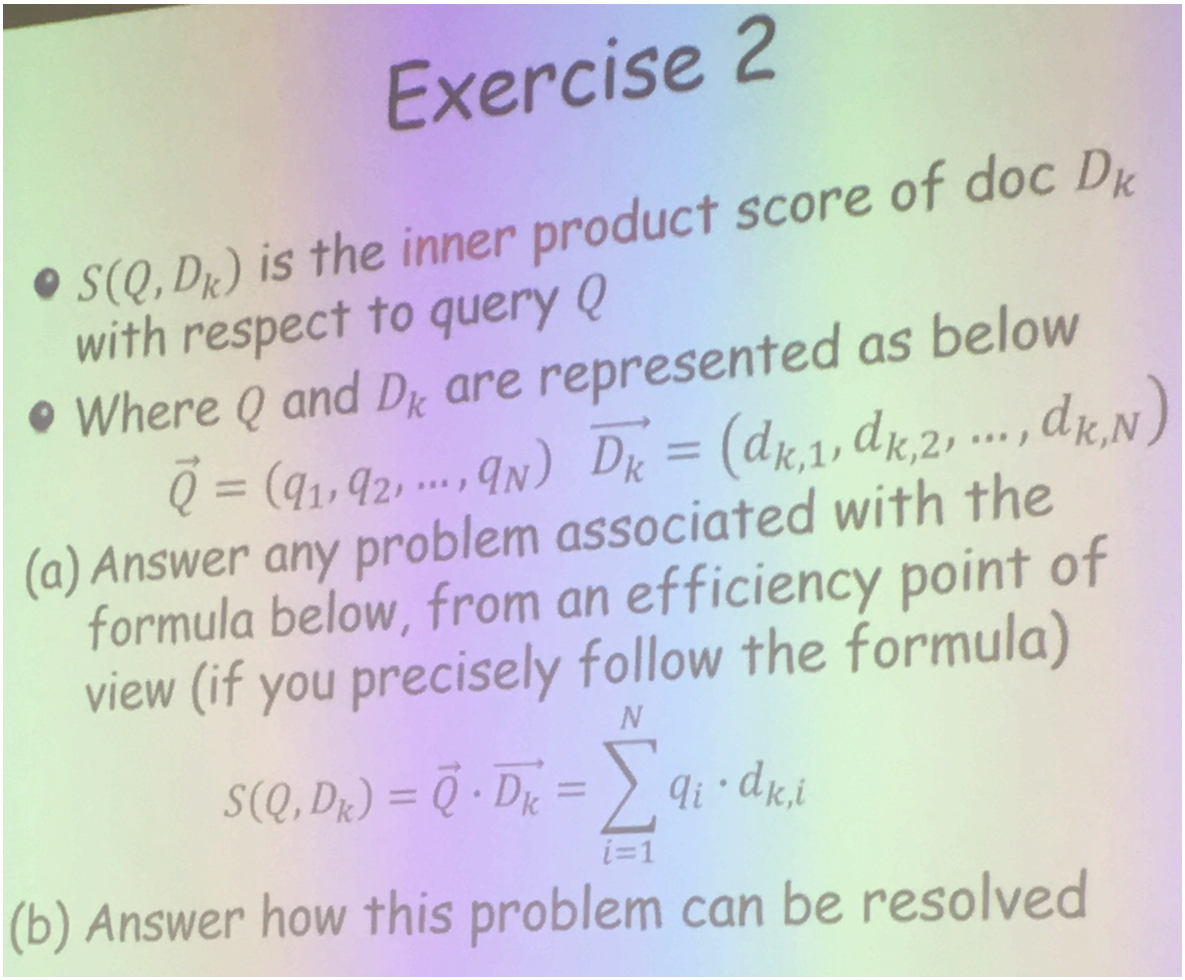
(c)Because the base number (as in the total number of words in the document can be either large or small. It is more reasonable to use fraction to show the frequency of a word in the document.

Answer for Exercise 1

- (a) D_2, D_1, D_1 (b) D_3, D_4, D_4
- (c) A chance to find a term in a document will increase, as the number of terms in that document becomes large, from a probabilistic point of view, and so the absolute frequency of a term in a document is less effective in representing the topic of a document than the relative frequency normalized by the total number of terms

(59 words)

	D1	D2	D3	D4
T1	0.5	0.7	1	0
T2	0.3	0.2	0	0.67
T3	0.2	0.1	0	0.33



(a) The system will be very slow. O(N*K)
(b)

Answer for Exercise 2

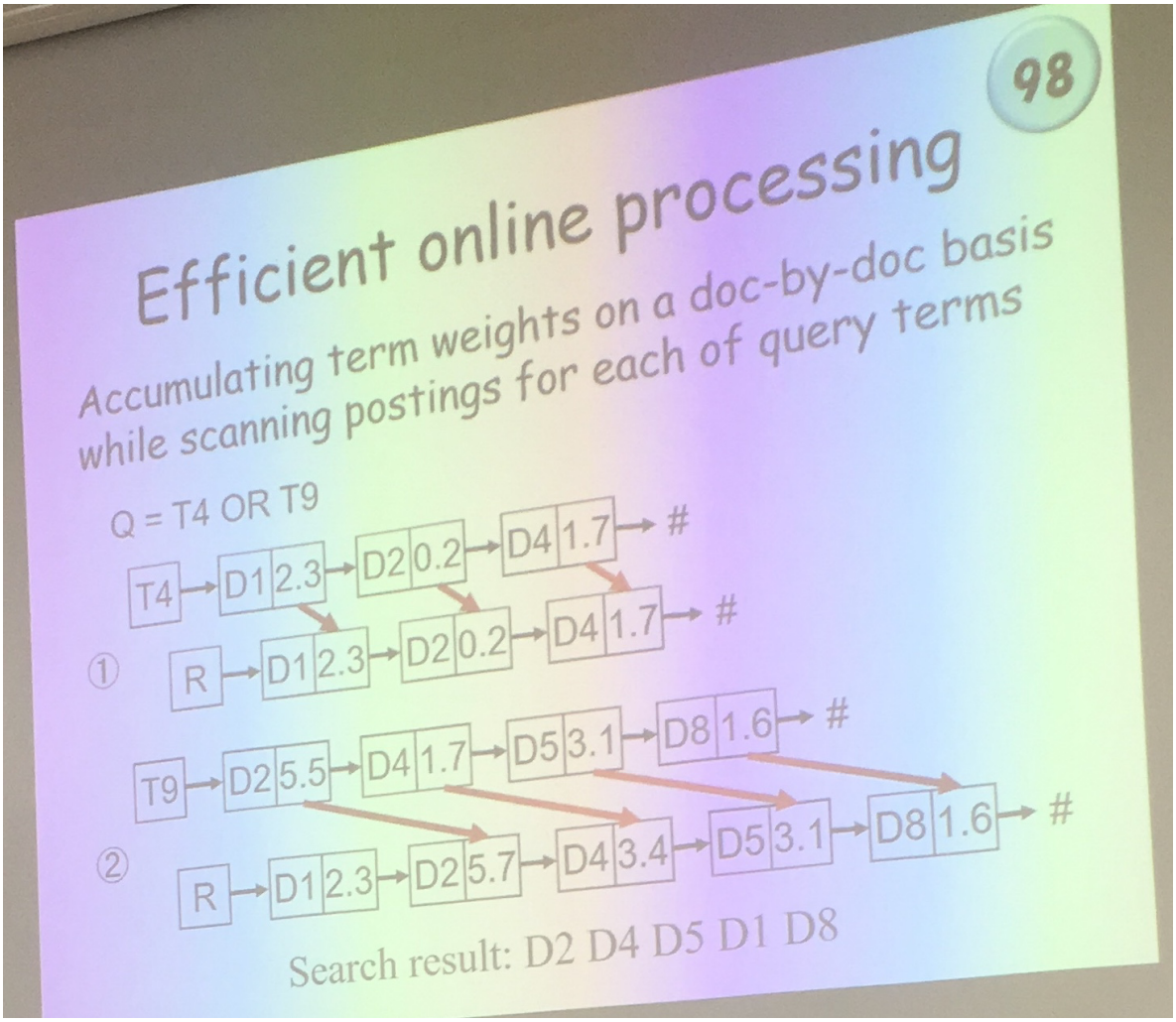
- (a) The formula is correct, but not practical because every product of q_i and $d_{k,i}$ has to be calculated despite a huge number of zero
- (b) Finding soon the documents that include one or more query terms by the Boolean IR
- [optional] If query is short, q_i can be binary (0/1), and so the formula is even simpler

$$S(Q, D_k) = \vec{Q} \cdot \overrightarrow{D_k} = \sum_{i=1}^{N} q_i \cdot d_{k,i}$$

This multiplication always repeats N times

The notation is better, but how to implement needs to be considered

$$\sum_{q_i \neq 0 \wedge d_{k,i} \neq 0} q_i \times d_{k,i} = \sum_{q_i \neq 0 \wedge d_{k,i} \neq 0} d_{k,i}$$



Several remarks

- The cost for computation is different depending on the operation
  - Multiplication is more expensive e than adding
  - Exponential and logarithm are even more
- Perform as many expensive operations as possible during offline
- If your formula happens to be a product, take log to obtain sum (the order of D_k's is kept)

$$\log \prod_i w(T_i, D_k) = \sum_i \log w(T_i, D_k)$$

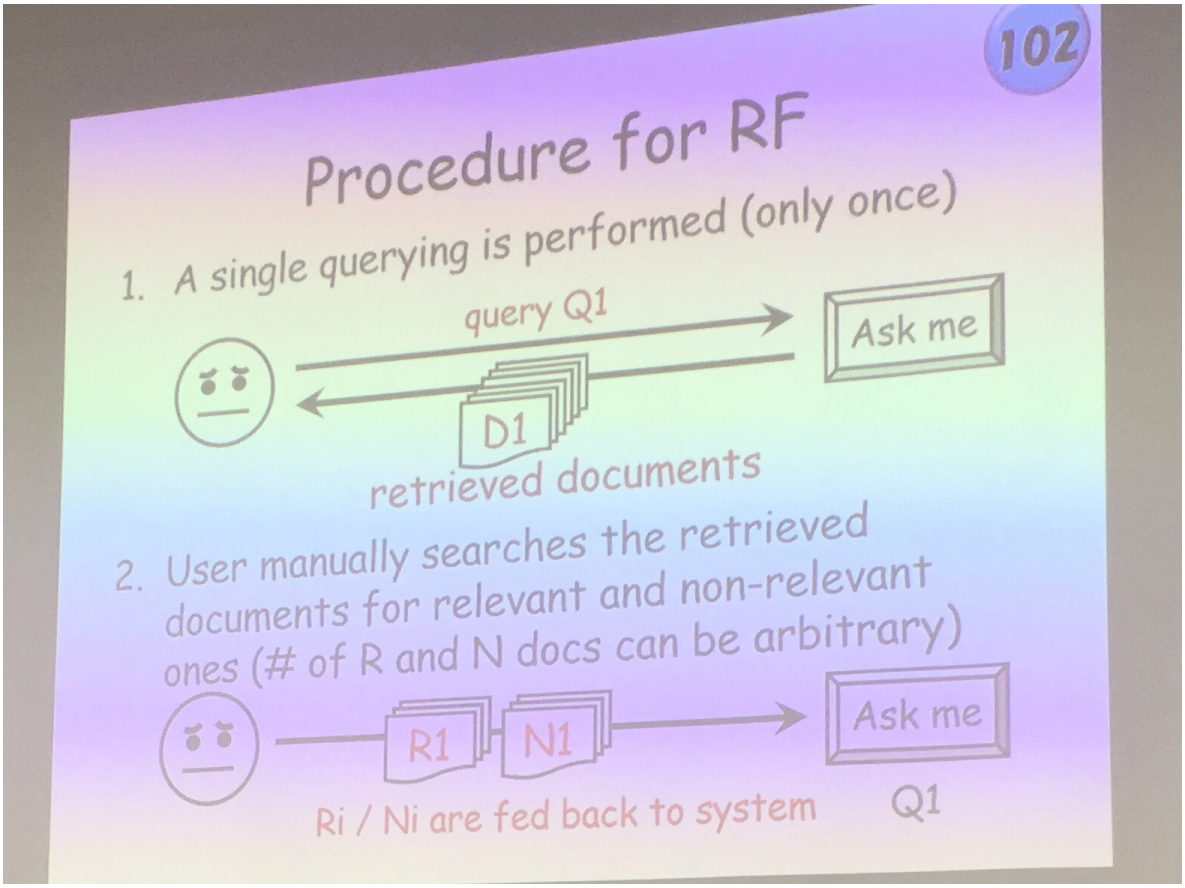
Relevance feedback (RF)

- A user would modify the query and submit it to a system, if not content with results
  - Insufficient awareness of user's need

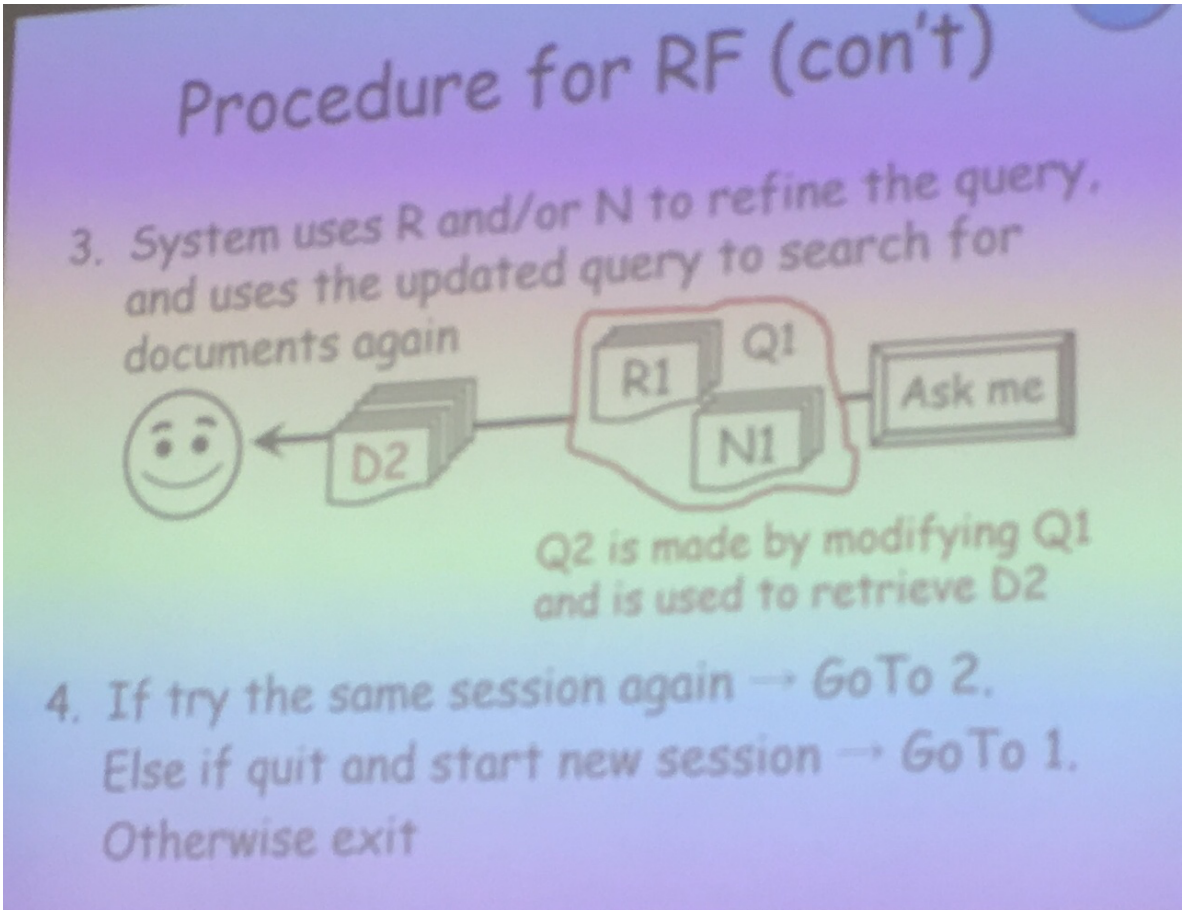
  - Dynamic change of user's need
- RF enables to use relevant (R) and non-R (N) example documents as references for system
  - A user is requested to judge the relevance of at least one document in search result
  - System modifies the query

Procedure for RF

1. A single querying is performed (only once)
2. User manually searches the retrieved documents for relevant and non-relevant ones (# of R and N docs can be arbitrary)



3. system uses R and/or N to refine the query, and uses the updated query to search for documents again
4. If try the same session again -> Go to 2; Else if quit and start new session -> Go to 1; Otherwise exit



How to refine query

- The centroid (centre of clusters) of Ri (arithmetic mean of all doc vectors in Ri) is representative of the reldocs
  - The same idea can be applied to Ni
- Subtraction C(Ri)[centre of correct answers] - C(Ni) [centre of incorrect answers]denotes system's error (**distance** in VS) [C(X) is the centroid of a set of vectors X]
- Rocchio's formula
  - Given Qi, Ri and Ni, Q(I+1) = Qi + Ri -Ni (vector calculation!!!见下一节)
  - Variant: Q(I+1) = Qi + α*Ri - β*Ni because usually Ri is more effective than Ni (α>β)
  - Members in Ni are hard to generalize

Visualization for Rocchio's formula

Query expansion

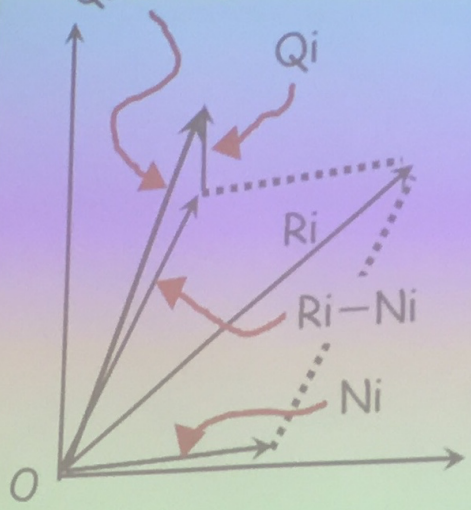
As a user continues RF, vector elements for synonyms of query would be 1

Note that steps 2 to 3. (See procedure for RF) is an application of similar document retrieval, where Q(i+1) is an input for the (i+1)-th iterations

Visualization for Rocchio's formula

105

$Q_{i+1} = Q_i + R_i - N_i$



Query expansion

As a user continue RF, vector elements for synonyms of query would be 1

Note that steps 2. to 3. (see Procedure for RF) is an application of similar document retrieval, where Q_{i+1} is an input for the (i+1)-th iterations

Example of query refinement

		T_1	T_2	T_3	T_4	T_5
R	D_1	9	1	3	0	0
N	D_2	1	0	2	0	6
N	D_3	2	0	1	6	4

$Q_1 = (1 \quad 0 \quad 0 \quad 0 \quad 0)$

Q+R-N $0+3-(2+1)/2 = 1.5$

$Q_2 = (\cancel{7.5}\ 8.5 \quad 1 \quad 1.5 \quad -3 \quad -5)$

Only the 1st quadrant is considered

→ Negative values are changed into 0