I

q t_1 t_2 ... t_n

d d_{w_1} w_2 ... w_n

A d_i ×
   d_j ✓
   d_k ✓

q'

for short term query expection

d_5 · d_1 ✓ d_7

q (java) q' · d_2 · ✓ (java programs)

↑

change weight.

d_3 · d_4

× ×

(coffee)

↓

expand query

move relevant. ↑ weight.

t_4 t_5

q t_1 t_2 t_3

   1 1 1 original q vector

   1.3 0.4 1 0.6 0.5

increase the position to each step.

q' q'' q

query expansion:

In cases of scenario user ~~don't~~ give little feedback.

① . return some samples for each category

② . ~~ques~~ ask more details. to core the core terms of query

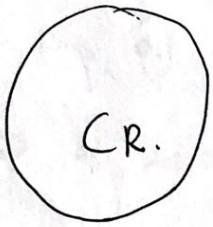
explicit the feedback from user.

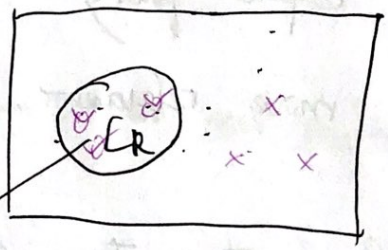
in VSM.

q D_R (maximise) — as close as possible — C_R. order D_R / D_N

D_N (minimise).

Assume for a category in q.





all the doc vectors (with many relevant terms)

d_j → pseudo document.

non-rele infor: negative

relevant info: ~~A - B~~ = + positive

stop words: A - B = 0.

$\frac{1}{|C_R|}$ normalisation.

q original.

terms to be very small

α β γ : how important is the feedback.

high low low IR learns very fast.

β γ.

(usually $\beta > \gamma$).

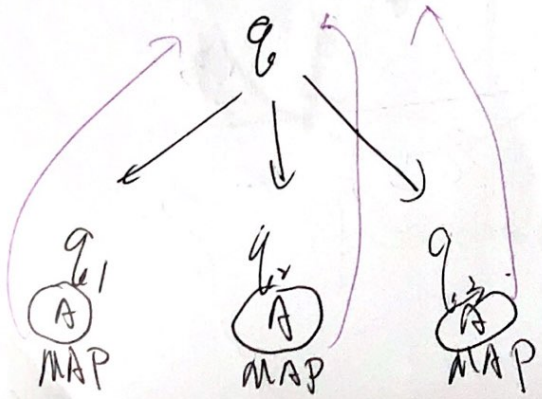
XNR : ~~most~~ non relevant doc. → we mostly want to remove

t_1 t_2 t_3 ..... t_n.

1 1 0.1 0.1 0.1 0.6 ....

Sum up all — only take account in the max non-relevant one (save computation).

α β γ



D_R. D_{IV}.

→ according to user's feedback.

	coffee t_1	java t_2	progr t_3	expr.
d_1				
d_2				
d_3				
d_4				

$d_1 d_2$ .65 .5 ✓

$\begin{bmatrix} d_3 \\ d_6 \end{bmatrix}$ .65 .55 0 ✓

0

java t_2

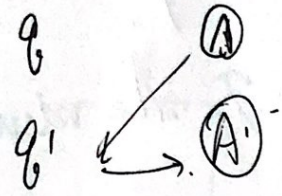
1		1	
−	.65	0.65	.5
		.55	.3

q. −0.65 (irrelev) 0.10 (relevant) 0.5 −.3.

↓ make relevant doc ↑.

make irrelevant doc ↓

Pseudo-Feedback / ~~Blind~~ Blind Feedback.



most of user's query is short.

$q = t_i \quad t_j \quad t_k$.

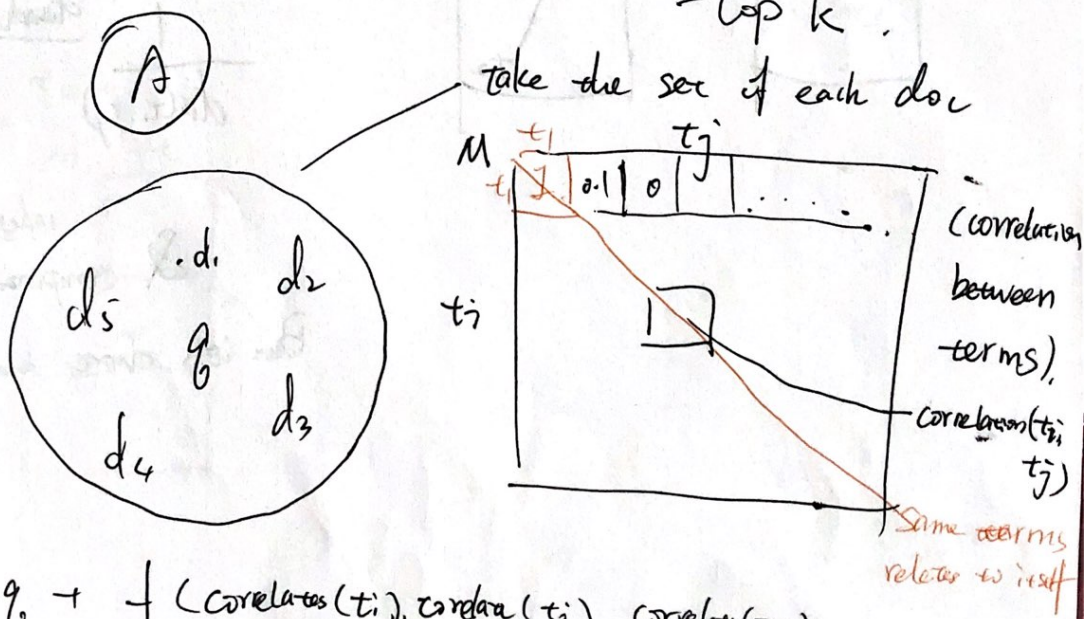
correlates (t_i)

$\ldots \quad (t_j)$ ↘ may overlap.

(t_k).

$q' = q + f(\text{correlates}(t_i), \text{cor}(t_j), \text{cor}(t_k))$

top k.



$q' = q + f(\text{correlates}(t_i), \text{correlates}(t_j), \text{correlates}(t_k))$

q''

freq q_{ij} the number of doc containing i and j.

the more docs in answer set contains it ↑ the relevance.

stop words =

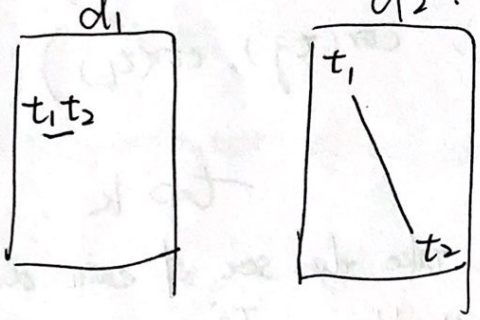
terms never occur together = 0.

always = 1.

sometimes close to 0.

Cheaper to calculate.

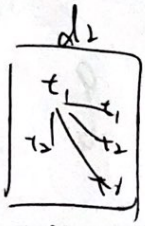
However.



distance measure between them.

show t_1 t_2 more relevant than t_1 t_2

$\frac{1}{dis(t_i, t_j)}$ disadv.



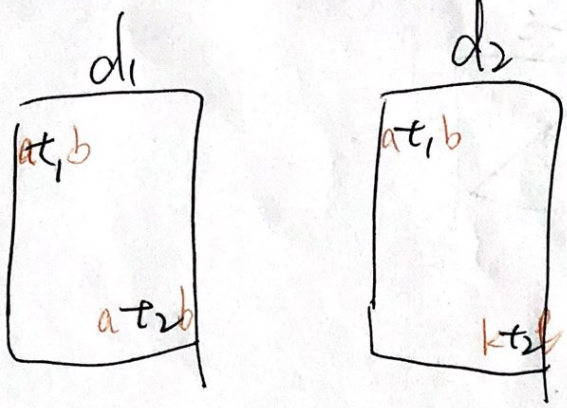
index more complexer

& comparaters

But it's more meaningful.

freque occur togethe -

distance betwn them.



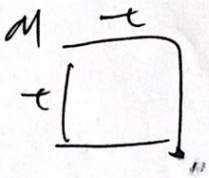
Correlation (t_1, t_2) is $f($ sim $($ neigh (t_1) neigh $(t_2))$).

more expensive.

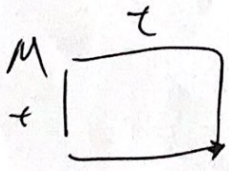
take query stream of the user → create a pseudo - docs.

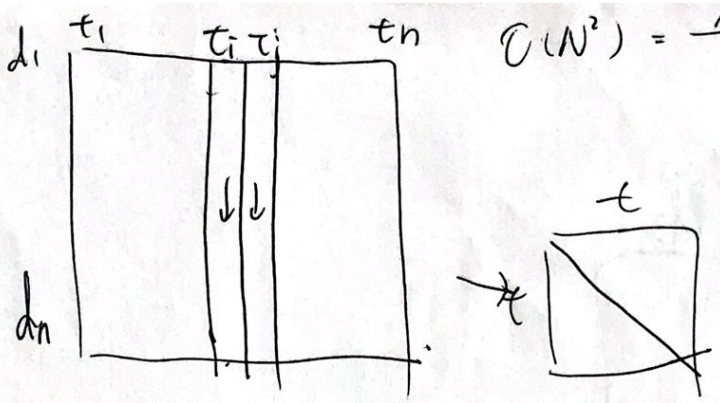
$q = t_1 \quad t_2 \quad t_3$

(A).



q'





$$O(N^2) = \frac{N(N-1)}{2}$$