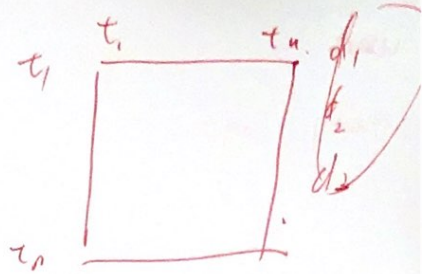
$\vec{q_0}$ = t_1 t_2 t_3

0.7 0.6 0.1



$\vec{q_1}$ t_1 t_2 t_3 ..... correlates.

can have some subtopic → can suggest users

↓

disambiguate query.

if $\vec{q_0} = t_1$. java = C++ programming }

coffee
espresso. }.

info.

① correlated t. select the best several correlated terms.

② diversity t. → help user to make query more specific.

try to find out the trade off between this two.

If I chose k top k correlated terms, how could I say the diversity.

users

query log.

user id	
1	t_1 t_2 t_3
2	t_1 t_2
1	t_2 t_3 t_4
2	t_4 ⟶ & related topic.
⋮	
x	t_3 t_4 ....

IR-2.

Clustering IR.

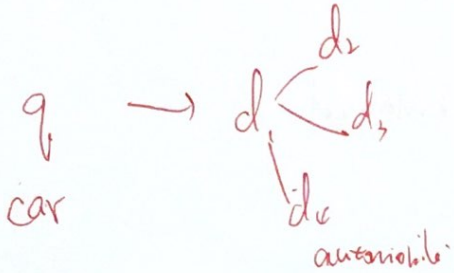
term - term matrix

↓

If a query → cluster query - cluster matrix.

Search result clustering → increase the diversity

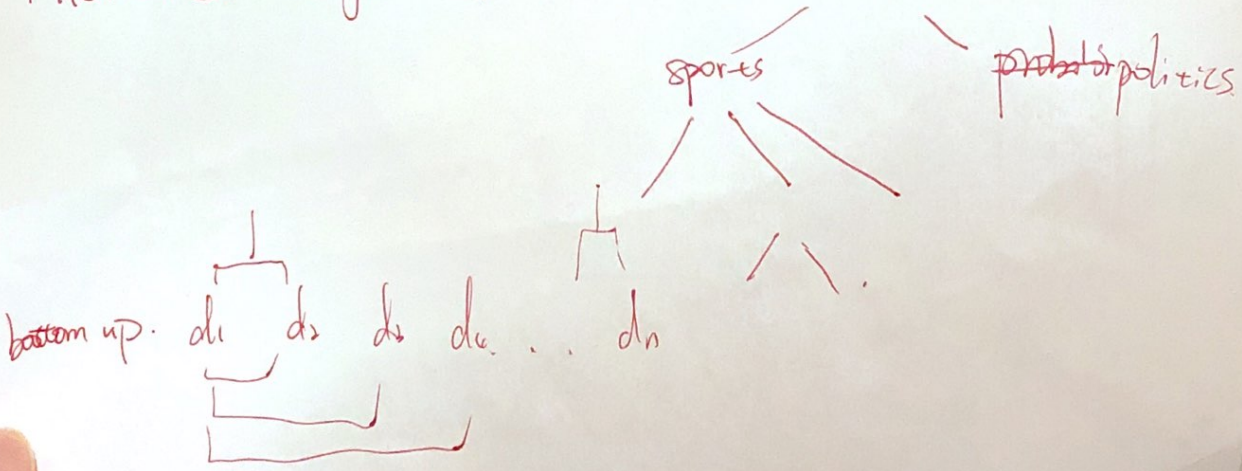
Scatter - clustery → visualise.



Flat algorithm

k - means.

Hierachical algorithm.



N docs and K clusters.

K-means. → use human judgement for evaluation.

K-NN: → classification, use prior knowledge in algorithm.

size of each cluster should be balanced.

① RSS (interior distance)

② K. (make a tight cluster)

③ give panalty to the cluster which size is too small or too large.