WM Homework 1 Report

R01944027 林廷舟

Problem

We now have an enormous corpus set. Given a query, how can we find related documents in the corpus (using the vector space model)?

Model

The final vector space model I construct measures the similarity using ***cosine similarity***. Each term is weighted using TF-IDF. I also tried ***dot product*** similarity but cosine performs better.

Experiment

Model	MAP on query-5.xml
Dot Product	0.463508875332529
Cosine	0.605786303798175
Cosine, Feedback, DR=20 for feedback, a=1.0, b=0.75	0.619657703730465
Cosine, Feedback, DR=10 for feedback, a=1.0, b=0.75	0.63448929814034
Cosine, Feedback, DR=10 for feedback, a=1.0, b=0.25	0.631569485699528

Findings

The most difficult part of this homework I think is how to deal with the huge (compared with those I've deal with) size of the data. Even with inverted index provided, it's still hard to handle. I tried to construct all information I need.

The next interesting stuff is the similarity measure. At the beginning I tried ***dot product*** similarity and got a MAP of 0.46. Then I modified it into cosine similarity and increased the MAP to about 0.60. I think the significant enhance is because that the actual content of the corpus are news article and its length varies a lot, which might cause the documents that has a high cosine similarity to be measured more far away with other documents which has more similar length with the query.

Besides, on models with Rocchio feedback, decreasing the size of *Dr* for feedback seems to improve the result. And the alpha and beta parameter tuning seems not affecting much.