Information Retrieval HW3

1. System model: Vector space model  
   mean average precision of original query: 0.33743028
2. Query expansion method: Rocchio method   
   number of relevant documents used in this round: 3  
   mean average precision after query expansion: 0.58955683
3. Review

This homework was a little difficult for me. At first I chose to implement local analysis, but afterwards I turned to the Rocchio method because I didn’t obtain good results from query expansion with former method. The discussion with teacher was beneficial to me, for I finally found the incorrectness in my original codes. I misunderstood the definition of the word collection and thought it was formed based on the query, but in fact it was all the documents collection which comprises the word collection.

After modifying the erroneous parts, I encountered the problems of endless debugging. Due to messy structure of my code, it was hard to find out the errors by naked eyes. I supposed I would learn more about the organization and management in programming in the future.

In terms of the query expansion techniques, Rocchio method was much easier to carry out than the local analysis, and the result was amazingly good. Because the preliminary mAp was slightly above 0.33, one relevant document for query expansion was enough to improve the ranking result. I picked relevant documents based on the ranking from the first search.

Below is a table of the output results in some of my experiments with different arguments:

|  |  |  |  |
| --- | --- | --- | --- |
| Number of relevant docs picked | Max query terms of each query | mAp after expansion | Difference (original mAp = 0.33743028) |
| 1 | 30 | 0.48086498 | 0.1434347 |
| 1 | 40 | 0.48785867 | 0.15042839 |
| 1 | 50 | 0.486294 | 0.14886372 |
| 2 | 40 | 0.47819191 | 0.14076163 |
| 2 | 50 | 0.48207274 | 0.14464246 |
| 2 | 60 | 0.47928725 | 0.14185697 |
| 3 | 55 | 0.54097309 | 0.20354281 |
| 3 | 70 | 0.55098392 | 0.21355364 |
| 3 | 120 | 0.58634555 | 0.24891527 |
| 3 | 200 | 0.59119546 | 0.25376518 |
| 3 | 300 | 0.59068958 | 0.2532593 |

As you can see, the mAp increased as we picked more documents for Rocchio’s query expansion. The reason for this result was because there were more relevant terms available to add in the original query, which can increase the possibility of retrieving articles that people want. Although we can use all the terms from the relevant articles we picked, yet this may bring about little growth in mAp with much more processing time. Overall, little amount of relevant documents in the Rocchio method is well enough to lead to a better searching result.