Report

Firstly, whole configuration of the program：

The environment of my program is using Python3.7 and using Spyder IDE and the operation system is Windows 8

In the program, we have the main following function:

main() function:

It is defined for running the program. In this function, firstly, it finds the directory where the 'index.py' is located by using the os.getcwd() method. And it then calls the buildIndex() to build index. And then it gives the 3 query to test the query() function, which is used for implementing a cosine similarity measure and implementing the Rocchio algorithm for query refinement. And then it calls the pseudo_query() function, which is used for the pseudo Relevance Feedback. Finally it calls print_dict() and print_doc_list() to export the terms and posting list, the documents and their document id separately. And it exports the time for building the index to the 'Output.txt', ‘Pseudo_Output.txt’ that will be located in the same directory as the 'index.py'.

buildIndex() function:

In this function, it reads documents from collection, tokenize and build the index with tokens. And we remove the word that is located in stp.txt in the ‘time’ file, which may influence of the calculation for cosines similarity and new queries and their weight during the Rocchio algorithm. And we need to put 'time' file in the same directory as the 'index.py'. And implement additional functionality to support relevance. We also use unique document IDs.

print_dict() function:

This function prints the terms and posting list in the index into the 'Output.txt'.

print_doc_list() function:

This function prints the documents and their document id into the 'Output.txt'.

query() function:

This function is used for exact top K retrieval using cosine similarity and returns at the minimum the document names of the top K documents orders in decreasing order of similarity score. And finally it will call rocchio() function for implementing the rocchio algorithm.

rocchio() function:

This function is used to implement rocchio algorithm and return the new query terms and their weights.

pseudo_query() function:

This function is used for the Pseudo Relevance Feedback.

There are some other functions that are auxiliary function for above functions. And the program will generate an 'Output.txt' that is an output for the Rocchio’s algorithm of user feedback. And the program will generate a ‘pseudo_Output.txt’ that is an output for the Rocchio’s algorithm of Pseudo Relevance Feedback.

Secondly Experimental study and performance analysis

We run 3 queries from the test bed to test the system.

The first query ID is 6. The query text is “CEREMONIAL SUICIDES COMMITTED BY SOME BUDDHIST MONKS IN SOUTH VIET NAM AND WHAT THEY ARE SEEKING TO GAIN BY SUCH ACTS”.

The result of the first query:

	Relevant	Nonrelevant	
Retrieved	7	2	9
Not Retrieved	2	412	414
Total	9	414	423

So The Precision P= 7/(2+7)= 0.778

The Recall R= 7/(2+7)=0.778

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	268*	323*	257*	395	304*	326*	308*	288*	171
Precision	1	1	1	0.75	0.8	0.833	0.857	0.875	0.778

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+0.8+0.833+0.857+0.875)/7 = 0.909

Below are positive and negative feedback provided for next iteration of the Rocchio algorithm

Positive feedback: 268 323 257 395

Negative feedback: 304 326 308 288 171

After the first iteration of the Rocchio algorithm

The terms of the new query and their weight are located in "Output.txt".

	Relevant	Nonrelevant	
Retrieved	7	2	9
Not Retrieved	2	412	414
Total	9	414	423

So The Precision P= 7/(2+7)= 0.778

The Recall R= 7/(2+7)=0.778

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	257*	395	323*	268*	308*	326*	304*	334*	370
Precision	1	0.5	0.667	0.75	0.8	0.833	0.857	0.875	0.778

Note: In above table, star means that it is the relevant document.

MAP = (1+0.667+0.75+0.8+0.833+0.857+0.875)/7 = 0.826

Below are positive and negative feedback provided for next iteration of the Rocchio algorithm

Positive feedback: 257 395 323 268 304

Negative feedback: 308 326 334 370

After the second iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Output.txt”.

	Relevant	Nonrelevant	
Retrieved	7	2	9
Not Retrieved	2	412	414
Total	9	414	423

So The Precision P= 7/(2+7)= 0.778

The Recall R= 7/(2+7)=0.778

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	257*	268*	395	323*	304*	308*	326*	334*	370
Precision	1	1	0.667	0.75	0.8	0.833	0.857	0.875	0.778

Note: In above table, star means that it is the relevant document.

MAP = (1+1+0.75+0.8+0.833+0.857+0.875)/7 = 0.874

Below are positive and negative feedback provided for next iteration of the Rocchio algorithm

Positive feedback: 257 268 395 323 304

Negative feedback: 326 334 370

After the third iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Output.txt”.

So for the first query:

MAP = (0.909+0.826+0.874)/3 = 0.870

The second query Id is 39. The query context is “COALITION GOVERNMENT TO BE FORMED IN ITALY BY THE LEFT-WING SOCIALISTS, THE REPUBLICANS, SOCIAL DEMOCRATS, AND CHRISTIAN”

The result of the second query:

	Relevant	Nonrelevant	
Retrieved	7	2	9
Not Retrieved	2	412	414
Total	9	414	423

So The Precision P= 7/(2+7)= 0.778

The Recall R= 7/(2+7)=0.778

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	360*	22*	265*	277*	219*	189*	394	150	173*
Precision	1	1	1	1	1	1	0.857	0.75	0.778

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+1+1+1+0.778)/7 = 0.968

Below are positive and negative feedback provided for next iteration of the Rocchio algorithm

Positive feedback: 360 22 265 277

Negative feedback: 219 189 394 150 173

After the first iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Output.txt”.

	Relevant	Nonrelevant	
Retrieved	8	1	9
Not Retrieved	1	413	414
Total	9	414	423

So The Precision P= 8/(8+1)= 0.889

The Recall R= 8/(8+1)=0.889

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	360*	277*	22*	265*	189*	219*	173*	73*	394
Precision	1	1	1	1	1	1	1	1	0.889

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+1+1+1+1+1)/8 = 1

Below are positive and negative feedback provided for next iteration of the Rocchio algorithm

Positive feedback: 360 277 22 265 173

Negative feedback: 289 219 73 394

After the second iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Output.txt”.

	Relevant	Nonrelevant	
Retrieved	8	1	9
Not Retrieved	1	413	414
Total	9	414	423

So The Precision P= 8/(1+8)= 0.889

The Recall R= 8/(1+8)=0.889

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	360*	277*	22*	265*	173*	189*	219*	73*	394
Precision	1	1	1	1	1	1	1	1	0.889

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+1+1+1+1+1)/8 = 1

Below are positive and negative feedback provided for next iteration of the Rocchio algorithm

Positive feedback: 360 277 22 265 173

Negative feedback: 219 73 394

After the third iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Output.txt”. So for the second query:

MAP = (1+1+0.968)/3 = 0.989

The Third query Id is 68. The query text is “INDIAN FEARS OF ANOTHER COMMUNIST CHINESE INVASION .”

The result of the third query:

	Relevant	Nonrelevant	
Retrieved	2	6	8
Not Retrieved	6	409	415
Total	8	415	423

So The Precision P= 2/(2+6)= 0.25

The Recall R= 2/(2+6)=0.25

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	31*	243	308	46*	290	87	420	68
Precision	1	0.5	0.333	0.5	0.4	0.333	0.286	0.25

Note: In above table, star means that it is the relevant document.

MAP = (1+0.5)/2 = 0.75

Below are positive and negative feedback provided for next iteration of the Rocchio algorithm

Positive feedback: 31 243 308 46

Negative feedback: 290 87 420 68

After the first iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Output.txt”.

	Relevant	Nonrelevant	
Retrieved	2	6	8
Not Retrieved	6	409	415
Total	8	415	423

So The Precision P= 2/(2+6)= 0.25

The Recall R= 2/(2+6)=0.25

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	308	31*	46*	243	257	326	322	304
Precision	0	0.5	0.667	0.5	0.4	0.333	0.286	0.25

Note: In above table, star means that it is the relevant document.

MAP = (0.5+0.667)/2 = 0.584

Below are positive and negative feedback provided for next iteration of the Rocchio algorithm

Positive feedback: 308 46 31 243 322

Negative feedback: 257 326 304

After the second iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Output.txt”.

	Relevant	Nonrelevant	
Retrieved	3	5	8
Not Retrieved	5	410	415
Total	8	415	423

So The Precision P= 3/(3+5)= 0.375

The Recall R= 3/(3+5)=0.375

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	308	243	31*	46*	322	328*	257	326
Precision	0	0	0.333	0.5	0.4	0.5	0.429	0.375

Note: In above table, star means that it is the relevant document.

MAP = (0.333+0.5+0.5)/3 = 0.44

Below are positive and negative feedback provided for next iteration of the Rocchio algorithm

Positive feedback: 308 243 31 46 322

Negative feedback: 257 326

After the third iteration of the Rocchio algorithm

The terms of the new query and their weight are located in "Output.txt". So for the third query:

MAP = (0.75+0.584+0.44)/3 = 0.591

Finally, for the query drift, after using the Rocchio algorithm, it expand the query as shown in the 'Output.txt', which means that the size of the query become larger than original one.

Third: Pseudo Relevance Feedback and performance analysis

We run 3 queries from the test bed that is the same queries from my experimental study. And we assume the top 3 results of the system are considered to be relevant. So others are not relevant

The first query ID is 6. The query text is "CEREMONIAL SUICIDES COMMITTED BY SOME BUDDHIST MONKS IN SOUTH VIET NAM AND WHAT THEY ARE SEEKING TO GAIN BY SUCH ACTS".

The result of the first query:

	Relevant	Nonrelevant	
Retrieved	7	2	9
Not Retrieved	2	412	414

Total	9	414	423

So The Precision P= 7/(2+7)= 0.778

The Recall R= 7/(2+7)=0.778

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	268*	323*	257*	395	304*	326*	308*	288*	171
Precision	1	1	1	0.75	0.8	0.833	0.857	0.875	0.778

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+0.8+0.833+0.857+0.875)/7 = 0.909

After the first iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Pseudo_Output.txt”.

	Relevant	Nonrelevant	
Retrieved	8	8	9
Not Retrieved	1	413	414
Total	9	414	423

So The Precision P= 8/(1+8)= 0.889

The Recall R= 8/(1+8)=0.889

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	257*	268*	323*	308*	326*	304*	334*	288*	370
Precision	1	1	1	1	1	1	1	1	0.889

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+1+1+1+1+1)/8 = 1

After the second iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Pseudo_Output.txt”.

	Relevant	Nonrelevant	
Retrieved	8	1	9
Not Retrieved	1	413	414
Total	9	414	423

So The Precision P= 8/(1+8)= 0.889

The Recall R= 8/(1+8)=0.889

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	257*	268*	323*	308*	326*	304*	334*	288*	370
Precision	1	1	1	1	1	1	1	1	0.889

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+1+1+1+1+1)/8 = 1

After the third iteration of the Rocchio algorithm

The terms of the new query and their weight are located in "Pseudo_Output.txt".

So for the first query:

MAP = (0.909+1+1)/3 = 0.970

The second query Id is 39. The query context is "COALITION GOVERNMENT TO BE FORMED IN ITALY BY THE LEFT-WING SOCIALISTS, THE REPUBLICANS, SOCIAL DEMOCRATS, AND CHRISTIAN"

The result of the second query:

	Relevant	Nonrelevant	
Retrieved	7	2	9
Not Retrieved	2	412	414
Total	9	414	423

So The Precision P= 7/(2+7)= 0.778

The Recall R= 7/(2+7)=0.778

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	360*	22*	265*	277*	219*	189*	394	150	173*
Precision	1	1	1	1	1	1	0.857	0.75	0.778

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+1+1+1+0.778)/7 = 0.968

After the first iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Pseudo_Output.txt”.

	Relevant	Nonrelevant	
Retrieved	8	1	9
Not Retrieved	1	413	414
Total	9	414	423

So The Precision P= 8/(8+1)= 0.889

The Recall R= 8/(8+1)=0.889

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	360*	22*	265*	277*	189*	219*	173*	73*	394
Precision	1	1	1	1	1	1	1	1	0.889

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+1+1+1+1+1)/8 = 1

After the second iteration of the Rocchio algorithm

The terms of the new query and their weight are located in “Pseudo_Output.txt”.

	Relevant	Nonrelevant	
Retrieved	8	1	9
Not Retrieved	1	413	414
Total	9	414	423

So The Precision P= 8/(1+8)= 0.889

The Recall R= 8/(1+8)=0.889

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	360*	22*	265*	277*	189*	219*	173*	73*	394
Precision	1	1	1	1	1	1	1	1	0.889

Note: In above table, star means that it is the relevant document.

MAP = (1+1+1+1+1+1+1+1)/8 = 1

After the third iteration of the Rocchio algorithm

The terms of the new query and their weight are located in "Pseudo_Output.txt".

So for the second query:

MAP = (1+1+0.968)/3 = 0.989

The Third query Id is 68. The query text is "INDIAN FEARS OF ANOTHER COMMUNIST CHINESE INVASION ."

The result of the third query:

	Relevant	Nonrelevant	
Retrieved	2	6	8
Not Retrieved	6	409	415
Total	8	415	423

So The Precision P= 2/(2+6)= 0.25

The Recall R= 2/(2+6)=0.25

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	31*	243	308	46*	290	87	420	68
Precision	1	0.5	0.333	0.5	0.4	0.333	0.286	0.25

Note: In above table, star means that it is the relevant document.

MAP = (1+0.5)/2 = 0.75

After the first iteration of the Rocchio algorithm

The terms of the new query and their weight are located in "Pseudo_Output.txt".

	Relevant	Nonrelevant	
Retrieved	2	6	8
Not Retrieved	6	409	415
Total	8	415	423

So The Precision P= 2/(2+6)= 0.25

The Recall R= 2/(2+6)=0.25

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	308	243	31*	257	326	304	46*	320
Precision	0	0	0.333	0.25	0.2	0.167	0.286	0.25

Note: In above table, star means that it is the relevant document.

MAP = (0.333+0.286)/2 = 0.31

After the second iteration of the Rocchio algorithm

The terms of the new query and their weight are located in "Pseudo_Output.txt".

	Relevant	Nonrelevant	
Retrieved	2	6	8
Not Retrieved	6	409	415
Total	8	415	423

So The Precision P= 2/(2+6)= 0.25

The Recall R= 2/(2+6)=0.25

Table for calculate the MAP: (from left to right the order is the decreasing order of top k documents)

	308	243	31*	257	326	46*	322	304
Precision	0	0	0.333	0.25	0.2	0.333	0.286	0.25

Note: In above table, star means that it is the relevant document.

MAP = (0.333+0.333)/2 = 0.333

After the third iteration of the Rocchio algorithm

The terms of the new query and their weight are located in "Pseudo_Output.txt".

So for the third query:

MAP = (0.75+0.31+0.333)/3 = 0.464