Your Name: Wei Wang

Your Andrew ID: ww5

Homework 1

Collaboration and Originality

Your report must include answers to the following questions:

1. Did you receive help of any kind from anyone in developing your software for this assignment (Yes or No)? It is not necessary to describe discussions with the instructor or TAs.
   If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.

   No

2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?
   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.

   No

3. Are you the author of every line of source code submitted for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.
   If you answered No:
   a. identify the software that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

Yes

4. Are you the author of every word of your report (Yes or No)?
   If you answered No:
   a. identify the text that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

Yes

Your Name: Wei Wang

Your Andrew ID: ww5

Homework 1

Instructions

1 Structured query set

1.1 Summary of query structuring strategies

1. And has high precision, low recall; Or has low precision, high recall. Keep them both in the query to receive relatively balance.
2. Put "AND" among related queries
3. Put "OR" between unrelated or similar queries.
4. Adding "Near" may increase the precision.
5. More important the word is, more probability that it will appear in both title, body and url
6. "Near/n" should be used when words should be a complete phrase

1.2 Structured queries

`718:#AND(Controlling.keywords #NEAR/1(acid.keywords rain.keywords))`

Strategy 3,4,5 has been used
I thought put NEAR can increase precision, while when I remove NEAR from this query, the precision increase
Acid rain is the information I would like to search as a phrase, and I don't want information about knowledge of acid rain, but controlling acid rain instead. Therefore this query can retrieve information I need.

`719: #AND(#OR(Cruise.title ship.title) #NEAR/2(damage sea life))`

Strategy 2,3,4,5has been used
Same as query 718
Cruise and ship are similar words, so #OR will be enough, damage sea life is the action that ship or cruise did, so I put them as NEAR, this two parts should be appear in the document concurrently. More importantly, ship or cruise is the topic I would like to search, therefore put them in the title will be good

`724:  #OR(Iran.url #AND(Iran.keywords Contra.keywords))`

Strategy 2,3,4has been used
No deviation
Iran is the typical information I need, as a result as long as it appear in keywords and url are both fine

`725:#OR(#AND(Low #NEAR/1(white.title blood.title cell.title)) count)`

Strategy 2,3,4,5has been used
No deviation
White blood cell is one phrase that is important, therefore I think they should be structured with NEAR/1, count is not important since Low can also indicate the meaning of "count", therefore I put OR between other words and count

733:#OR(Airline.url #OR(#AND(overbooking.keywords Airline.keywords) #AND(overbooking.title Airline.title)))

Strategy 2,3,5 has been used
No deviation
Overbooking can be bus and other booking systems; Therefore airline is a necessary word show up concurrently. They can be both at title and keywords. Most important word airline can be in the url as well

734:#AND(Recycling successes)

Strategy 2 has been used
No deviation
These are simple information need. I need to search for recycling activities are success.

735:#OR(#AND(Afghan women) condition)

Strategy 2,3 has been used
No deviation
These are simple information need. I need to search for recycling activities are success.

741:#NEAR/1(Artificial Intelligence)

Strategy 6 has been used
No deviation
Artificial Intelligence should be treated as a whole part

744:#AND(Counterfeit ID punishments)

Strategy 2 has been used
No deviation
Three words should show up together.

746:#OR(#AND(Outsource.url #NEAR/1(job.keywords Outsource.keywords)) India)

Strategy 2 has been used
No deviation
India can be an optional part of query.

2 Experimental results

2.1 Unranked Boolean

	BOW #OR	BOW #AND	Structured
P@10	0.0000	0.2000	0.0333
P@20	0.0000	0.2250	0.0500
P@30	0.0033	0.2367	0.0519
MAP	0.0002	0.0489	0.0047
Running Time	01:07	00:12	00:04

2.2 Ranked Boolean

	BOW #OR	BOW #AND	Structured
P@10	0.0400	0.3800	0.0444

P@20	0.0800	0.3650	0.0833
P@30	0.0867	0.3300	0.0963
MAP	0.0079	0.0871	0.0109
Running Time	00:53	00:15	00:05

3 Analysis of results: Query operators and fields

The start of structuring a query is a little bit hard. Since in the real life when input our terms into search engines' search space, I usually assume that the terms are connected by "and", while in this homework, by only implementing AND in the query will result a high precision with low recall. And by finishing this assignment, some of my understanding of the construction of the query are shown below:

By indicating operators, we achieve different methods of retrieval of scores and doc ID.As for field, by indicating the field, we can increase the precision. The field as keyword usually contains the term of none. As result, searching for the keywords can be a efficient way since keywords, title and url are usually at the position which index are small and easier to found.

While to the end of experiments, my result still just a little improvement above the baseline of OR. When I was doing adjustment, I don't want to increase the usage of AND without making sense. Therefore, the result shown above is the most appropriate result that I can provide with.

4 Analysis of results: Queries and ranking algorithms

Operators: the performance of operator is not consistent with what I expect from the beginning. AND and OR operator are the same as common sense, as more AND operator are used, the precision is being increased, contradictory, OR give bigger probability on retrieving the documents that are relevant. This is being achieved since AND will only retrieved the document contains all arguments, and OR will return documents contains at least one argument. As for NEAR/n, I regard it as the extension for AND, and therefore now the algorithm should consider the position of words, which I thought it will increase precision. While it is performing as decreasing the precision. This is because by determine the n, some documents will not be return if the pair of words are away from each other by >n distance.

The ranked model and unranked model don't have much differences in terms of running time, while ranked model perform significantly better than unranked model in terms of the precision. This is caused by the difference in the score. In this assignment, the score is set to be equivalent to term frequency, which is to say, the more the term shows in the document, the higher the rank for the documents.