Your Name: Yiding Ou

Your Andrew ID: yidingo

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.

   No

   If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.

2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?

   No

   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.

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.

   Yes

   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.

4. Are you the author of every word of your report (Yes or No)?

   Yes

   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.

Your Name: Yiding Ou

Your Andrew ID: yidingo

Homework 1

1 Structured query set

1.1 Summary of query structuring strategies

When two terms are actually phases, I will use a NEAR operator to combine them. If two words has similar meaning, I will use a OR operator for them. If both term are important when we trying to retrieve the document, I will use a AND operator for them. Different situation needs different operators. Some term that has a large frequency is more likely to be found in the body field, and some rare term is more likely to appear in the title part.

1.2 Structured queries

718: #OR (#AND (Controlling.body #NEAR/1 (acid.title rain.title)) #AND (Controlling.body #NEAR/1 (acid.url rain.url)))
Since acid and rain is a phase, so I use NEAR operator to combine them. If controlling missed the query will change meaning. So I use AND operator for them. And acid rain is uncommon phase, it is likely to be found in url or title.

719: #AND (#OR (Cruise ship) damage life sea)
Since Cruise and ship have similar meaning, so I use OR operator for them, missing one of them probably will not change the meaning of the query.

724: #OR (#AND (Iran.title Contra.title) #AND (Iran.url Contra.url))
Both terms are needed, otherwise the meaning of query will be changed, and these two term are likely to be found in url or title, since they are uncommon words.

725: #AND (Low #NEAR/1 (white blood cell) count)
while blood cell is a phase, so I use NEAR operator. Low is a important term, since without this term the document we retrieved may contains high white blood cell, and that is irrelevant.

733: #OR (#AND (Airline.url overbooking.url) #AND (Airline.title overbooking.title))
since this is a short query, these two term probably can be found in the url or title. And both of them cannot ignore, otherwise the retrieve result will be affected.

734: #AND(Recycling successes)
Both terms are needed otherwise the retrieve result will lost some relevant documents.

735: #OR (#AND (Afghan.body women.body condition.body) #AND (Afghan.title women.title))
Afghan women are two words that always used in summary and title, so I search the document with all term in body or has Afghan and women in title.

741: #OR (#NEAR/2 (Artificial.body Intelligence.body) #AND (#NEAR/2 (Artificial.title Intelligence.title)))
Artificial intelligence is a phase, and some document may include artificial human intelligence or other phases, so I use NEAR/2 operator here. And these phase is commonly seen in document's title. So I used title field here.

744: #OR (#AND (#NEAR/2 (Counterfeit.body ID.body) punishments.body) #AND (Counterfeit.title ID.title))
I think counterfeit ID is a phase, if we separate those words we probably retrieve bunch of irrelevant doucments. And those are commonly seen title.

746: #OR (#AND (Outsource.body India.body job.body) #AND (Outsource.title India.title))
outsource and india are uncommon terms so they are likely found in title. And both terms are needed to retrieve relevant documents.

2 Experimental results

Present the complete set of experimental results. Include the precision and running time results described above. Present these in a tabular form (see below) so that it is easy to compare the results for each algorithm.

2.1 Unranked Boolean

	BOW #OR	**BOW #AND**	**Structured**
P@10	0.0000	0.2000	0.2500
P@20	0.0000	0.2250	0.2250
P@30	0.0033	0.2367	0.2200
MAP	0.0002	0.0489	0.0443
Running Time	8422.070 ms	1557.406 ms	1625.405 ms

2.2 Ranked Boolean

	BOW #OR	**BOW #AND**	**Structured**
P@10	0.0400	0.3800	0.3200
P@20	0.0800	0.3650	0.2950
P@30	0.0867	0.3300	0.2767

MAP	0.0079	0.0871	0.0513
Running Time	8027.354 ms	1927.018 ms	1981.758 ms

3 Analysis of results: Query operators and fields

Since OR operator has high recall, and AND operator has higher precision. OR operator can give you all the relevant document but those documents are covered with many irrelevant documents, so merely using OR operator is not good enough to construct a good query, in addition, OR operator consumes longer than for the search engine to process the query. As for AND operator, we lost some documents, some documents may contains only partial query terms. And if we merely use AND operator we will lost those documents, on the other hand, the documents AND operator retrieved have a high possibility that it is relevant. However, in our experiment, OR operator does not exhibit high recall feature, the reason probably is we restricting the retrieve document number. The NEAR operator can filter out some irrelevant documents, but it could also filter out some relevant documents. The field operators can speed up the query, search the query in a smaller, but it also has change to delete some relevant documents.

It is hard to mitigate the trade of time and recall level, even fix the operators and using fields.

4 Analysis of results: Queries and ranking algorithms

The advantage of MAP is its stability, it represents the recall level of whole query set. However, it does not showed how good or how bad a single query is, some query may have 1 on its average precision, which is really impressive, but the MAP does not show that. MAP is a Macro average, which means it gives all query equal importance, but in some cases the fact is not this, so the result probably is not precise.

As for p@n, when we conducting the experiment, we do not need have knowledge about the size of retrieve result, and it is really easy to understand. However, comparing to MAP, p@n is less stable. For instance, some simple get 10 relevant document retrieved at top, and we test it against p@5, the outcome will be really good, however, if the have a complex query, we will get a bad result. Another reason is that the result of p@n will not hold for other metrics.