Your Name: Hao Wang

Your Andrew ID: haow2

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: Hao Wang

Your Andrew ID: haow2

Homework 1

1 Structured query set

1.1 Summary of query structuring strategies

I used And, Near/n, and Or operators do do the query and took fields including body, url, keywords, inlink and title into consideration in both Ranked Boolean and UnRanked Boolean. Then analyzed their performance.

1.2 Structured queries

List your structured queries. For each query, provide a brief (1-2 sentences) discussion of:

1. which strategy (from Question 2.1) was used for that query,
2. any important deviations from your default strategies, and
3. your intent, i.e., why you thought that particular structure was a good choice.

69:#AND(#OR(sewing.title instructions.title) #NEAR/10(sewing.body instructions.body))
Get all titles have sewing and instructions, then find sewing and instructions in there body within the range of 10. The result seems to be more accurate in this way, because keywords are in both title and body.

79:#OR(#AND(voyager.keywords voyager.url) #OR(voyager.inlink voyager.url))
Find all voyager in keywords and voyager in url. Then add the result by voyager in link and voyager in url.

84:#AND(#OR(continental plates) #NEAR/5(continental plates))
Find all documents has continental and plates and there range in within 5.

89:#OR(ocd.title ocd.keywords ocd.url)
Find all ocd that are in title, keywords and url

108:#OR(#AND(ralph owen) #NEAR/7(brewster owen))
Find all documents have ralph and owen, and the distance between brewster and owen in within 7.

141:#AND(#NEAR/5(va dmv) #NEAR/5(dmv registration))
Find all documents that va and dmv are in 5 words, and dmv and registration in 5 words. The result will be more relative now.

146:#OR(#AND(sherwood regional) #OR(sherwood regional library))

Find all documents have sherwood and regional, or have sherwood regional library. This is a better query for finding a library in sherwood region.

153:#AND(pocono.title pocono.body pocono.keywords)
Find all documents have pocono in title, body and keywords.

171:#OR(#NEAR/5(ron.body howard.body) #NEAR/5(ron.title howard.title))
Find all documents that ron and howard in body and title and in range of 5.

197:#AND(#OR(idaho state flower) #NEAR/8(idaho flower))
Find all documents have idaho state flower and idaho and flower are in range of 8.

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	0.1100	0.2200
P@20	0.0150	0.1350	0.3100
P@30	0.0200	0.1533	0.3067
MAP	0.0020	0.0665	0.0849
Running Time	00:05.3	00:09.2	00:02.2

2.2 Ranked Boolean

	BOW #OR	BOW #AND	Structured
P@10	0.1700	0.3700	0.2500
P@20	0.2800	0.4450	0.3200
P@30	0.3367	0.4633	03133
MAP	0.1071	0.1882	0.1127
Running Time	00:05.5	00:09.4	00:02.2

3 Analysis of results: Queries and ranking algorithms

Discuss your observations about the differences between the three different approaches to forming queries, and the two different approaches to retrieving documents (i.e., retrieval models) in terms of their retrieval performance and running time.

Hint: Do not just summarize the results from the previous sections; we can see those results above. You are expected to provide your interpretation of the results based on what you learned in the lectures and readings. This is your chance to show what you learned from this homework assignment - take this section very seriously

Hint: Probably this section doesn't need to be longer than ¾ of a page (not counting these instructions).

Generally speaking, OR has the lowest precision than AND and NEAR. NEAR operator is especially good at phrase query. Besides, RankedBoolean is much better than UnRankedBoolean and they have quite similar running time, so I think Rankedboolean is better then UnRankedBoolean.

Take Query 146 as an example. I used the following query:
146:#OR(sherwood regional library)
146:#AND(sherwood regional library)
146:#AND(#NEAR/3(sherwood library) #NEAR/3(regional library))
146:#AND(#NEAR/5(sherwood library) #NEAR/5(regional library))

Here is the result (UnRankedBoolean):

	OR	AND	NEAR/3	NEAR/5
P@10	0.0000	0.1000	0.2000	0.1000
P@20	0.0000	0.0500	0.3500	0.2000
P@30	0.0000	0.1333	0.2667	0.2667
MAP	0.0000	0.1274	0.1568	0.1272

Here is the result (RankedBoolean):

	OR	AND	NEAR/3	NEAR/5
P@10	0.0000	0.1000	0.4000	0.4000
P@20	0.0000	0.0500	0.3500	0.3500
P@30	0.0000	0.1333	0.2667	0.2667
MAP	0.0000	0.1274	0.2428	0.2256

The result in following tables, verified my conclusion.

4 Find all documents have idaho state flower and idaho and flower are in range of 8.Analysis of results: Query operators and fields

Discuss the effectiveness, strengths, and weaknesses of the query operators and fields, and your success and failure at using them in queries. Did they satisfy your expectations?

Hint: Same hints as above.

Because in this part, the most important goal is to test query operators and fields, so I finished all the following experiment in Ranked Boolean.

197:#OR(idaho state flower)
First I used OR operator, and got a very poor result. P@10, P@20, P@30 are all 0, and MAP is 0.0409.

197:#AND(idaho state flower)
Then I changed to AND operator, Now the result is much better than OR, P@10 is 0.4000, P@20 is 0.2500 and P@30 is 0.2000, MAP is 0.0358.

197:#AND(#NEAR/3(idaho state) flower)
Then I changed to NEAR/5. Now the result is much better than OR and AND. P@10 is 0.4000, P@20 is 0.4000, P@30 is 0.6000, MAP is 0.2352. Because, idaho state is more possible to be a phrase.

197:#AND(#NEAR/8(idaho state) flower)
Then I changed the parameter of NEAR operator. This time, the range is larger, so it is more possible to misjudge this time, and the result is not so good as the former one. P@10 is 0.4000, P@20 is 0.4000, P@30 is 0.5333, MAP is 0.2307.

197:#AND(#NEAR/2(idaho state) flower)
This time, I narrowed the range, and the result is not so good as 5. This is because that the phrase can be decorated by other words. P@10 is 0.4000, P@20 is 0.4000, P@30 is 0.6000, MAP is 0.2292.

197:#NEAR/5(#NEAR/5(idaho state) flower)
This time, take flower as a part of the phrase. It seems that the result is not very good. P@10 is 0.4000, P@20 is 0.4000, P@30 is 0.3333, MAP is 0.0472. So it is very possible that flower is not the part of phrase.

197:#AND(#NEAR/3(idaho.title state.title) flower)
This time, I used different field. It seems that idaho and state is less likely to be part of title than body. P@10 is 0.6000, P@20 is 0.3500, P@30 is 0.2333, MAP is 0.0339.

197:#AND(#NEAR/3(idaho.url state.url) flower)
This time, I used different field. It seems that idaho and state is less likely to be part of url than body. P@10 is 0.4000, P@20 is 0.2000, P@30 is 0.1333, MAP is 0.0354.

197:#AND(#NEAR/3(idaho.inlink state.inlink) flower)
This time, I used different field. It seems that idaho and state is less likely to be part of url than body. P@10 is 0.3000, P@20 is 0.1500, P@30 is 0.1000, MAP is 0.0214.

197:#AND(#NEAR/3(idaho.keywords state.keywords) flower)
This time, I used different field. It seems that idaho and state is less likely to be part of keywords than body, but it is much better than other fields. P@10 is 1.000, P@20 is 0.9000, P@30 is 0.7333, MAP is 0.1941.

All in all, it seems that the mix of AND and NEAR operators can have the best outcome. If two words are very likely to be a phrase, then use NEAR to them. Otherwise, use AND. If we just use OR, the performance will be very bad. But it can also have quite good performance if combined with other operators. If we choose a right field, then the result will be very different, they affects very much to the result.