Your Name: Shuang Xu

Your Andrew ID: sxu1

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: Shuang Xu

Your Andrew ID: sxu1

Homework 1

1 Structured query set

1.1 Summary of query structuring strategies

Briefly describe your strategies for creating structured queries. These should be general strategies, i.e., not specific to any particular query.

Consider the terms in the original #OR query.

- If there's a proper name, use #NEAR to constrain the name to improve accuracy.
- If there're terms from different name entity categories, use #AND to connect them to improve accuracy.
- If the term has another expression (or synonym) that is also widely used, use #OR to combine the pair of terms to improve recall.
- If there's too few arguments, use 'title' or 'keywords' field to improve accuracy.

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(sewing.title #OR(instructions.body tutorial.body))

The information need is most likly to find documents about how to sew. Sew is the keyword and would be shown in title. The documents that satisfy our needs should be an instruction or tutorial. Instructions and tutorial means the same thing, so use #OR to improve the recall.

79:#AND(voyager.title)

I'm not sure about the information need of this query. So I simply use #AND operator. And I believe voyager in title can help improve precision.

84:#OR(#NEAR/3(continental plates) #NEAR/3(tectonic plates))

Continental plates is a scientific theory and tectonic plates means the similar stuff. The information need must be the theory of continental plates. Use #NEAR to constrain the theory name and improve the accuracy. Use #OR to find more documents and improve the recall.

89:#OR(#NEAR/3(obsessive compulsive disorder) ocd.keywords)

Ocd is abbreviation for obsessive compulsive disorder and is a medical term. Use #OR to improve the recall.

108:#NEAR/3(ralph owen brewster)

Ralph Owen Brewster (February 22, 1888 – December 25, 1961) was an American politician. The information need is mostly likely to find documents about him. The name in the relevant documents are tend to locate in close position. So #NEAR is supposed to be here to improve precision.

141:#OR(#AND(virginia #NEAR/10(vehicle registration) #AND(dmv registration)))

The information need is to find documents about vehicle registration in Virginia.

146:#NEAR/3(sherwood regional library)

This is the name of library. Use #NEAR to specify.

153:#AND(pocono.title)

Not sure about the information need of this query. So simply use #AND operator. I believe pocono in title can help improve precision.

171:#OR(#NEAR/3(ron howard) #AND(howard actor))

Ron howard is the name for an actor. Howard may refer to Ron howard but there're many Howard in the world, so use actor to constrain.

197:#AND(#NEAR/3(state.body flower.body) idaho)

State flower is a term so use #NEAR. Idaho is a location so use #AND to connect. Both operator will improve precision.

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	0.3400
P@20	0.0150	0.1350	0.3250
P@30	0.0200	0.1533	0.3233
MAP	0.0020	0.0665	0.1528
Running Time	00:07	00:02	00:03

2.2 Ranked Boolean

	BOW #OR	BOW #AND	Structured
P@10	0.1700	0.3700	0.5400
P@20	0.2800	0.4450	0.5250
P@30	0.3367	0.4633	0.4833
MAP	0.1071	0.1882	0.2464
Running Time	00:07	00:02	00:03

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 total running time.
Hint: Do not just summarize the results from the previous sections; we can see those results above. You are expected to provide <u>your interpretation</u> 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 - <u>take this section very seriously</u>
Hint: Probably this section doesn't need to be longer than ¾ of a page (not counting these instructions).

Retrieval performance
BOW #OR produces high recall while BOW #AND produces high precision. Well-designed structured queries can balance recall and precision.

Recall and precision are measures to evaluate how documents satisfy information need. The formula are as follows.
Recall = tp/(tp+fn)
Precision = tp/(tp+fp)

#OR expands the result documents, the more retrieved documents we have, the more likely we can get the relevant documents. #OR brings about larger true positive than #AND, so #OR wins in recall.

#AND reduces false positive, thus reduces the denominator in the precision formula. Although #OR brings larger true positive, it increase false positive as well and the increase in false positive may be greatly larger than the increase in true positive. So the precision tends to decrease in most case.

Ranked Boolean system results in higher precision than unranked Boolean system because the score (term frequency) explains a lot. Important terms tend to appear more often in more relevant documents.

Total running time

#OR performs union operation on inverted lists and #AND performs intersection on the inverted lists. Both #AND and #OR operations will combine score lists, or docid lists in memory. Actually smaller intermediate results can result in shorter running time. #OR produces large intermediate results so BOW #OR approach takes longer time.

Unranked Boolean system and ranked Boolean system do not differ a lot in running time. Although unranked Boolean system access docid list and ranked Boolean system access score list when processing queries, the chance they access inverted lists, which are stored in the disk, are similar. So the running time would not change a lot.

4 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.

#OR is used to expand the search and usually brings higher recall. It's effective for vague information needs. #AND is used to narrow the search and usually brings higher precision. It's effective for specific information needs. #NEAR specifies the location but has strict order. Very effective for specific phrase searching especially proper name. Good for precision but may reduce recall.

For example, there're 73 relevant documents about Ralph Owen Brewster. For ranked system, query #NEAR generates great precision, 100% at P@10, 100% at P@20 and 96% at P@30 compared to #AND operator which only has 80% 65% 73% P@n precision respectively. But at the same time, query #AND returns 59 related documents and #NEAR returns only 47.

The field body can produce high recall results while fields like url, inlink, title or keywords can produce high precision results. It's effective to use keywords or title if there's only one term in query.

For example, there're 122 relevant documents about Pocono. Query #AND(pocono.body) returns 5 related documents but Query #AND(pocono.title) returns 35. And the precision improved about 12.4% after using title field.

#NEAR and title field greatly satisfied my expectations.