Your Name: Shousan Liao

Your Andrew ID: shousanl

Homework 1

1 Introduction

1.1 Collaboration and Originality

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. Did you examine anyone else's software for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.
   No
4. 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.

5. 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.

2 Structured queries

2.1 Query structuring strategies

When structure the query, I would firstly look at the independence and correlation between each term. If the term is synonym such as Carnegie Mellon University and CMU, or operator would be a good one. If different term grouping together could form a better term or interpretation, AND operator is a suitable choice. For example, when user query "Carnegie Mellon University ", it is more reasonable that they want to search the university instead of people's name Carnegie and Mellon.

In terms of near operator, I would apply it when different term has strong correlation but not necessary to be adjacent. For example, in spite that Donald Trump is a very popular keyword, there's also a bunch of resource writing his name with middle name – Donald J. Trump. In this scenario, using near / 1 operator would probably give us more accurate result.

Finally, I would use syn operator when there's only one term for a query while that term is not very specific. The reason behind this decision is that when we are searching for a clear word such the name of a disease or a baseball player, users usually know what they want and expect the result containing that term. On the other hand, query term like chair, table is too general that user may also want to see relevant term like furniture, seat, occur in the document.

Queries

33:#and(elliptical trainer): elliptical trainer is a stationary exercise machine that should be grouped together.

46:#and(#and(alexian brothers) #syn(hospital)): alexian brothers is a specialty care, emergency care institution. Thus, I use #syn(hospital) hoping that some relevant keyword like medical, health could also be considered.

64:#or(moths.title): No strategy applied

71:#near/5(living india): I remove "in" between living and india because it is just an unimportant preposition. Also, since living is not necessarily come after the location (for example: living 5 miles away from India's capital …), I think near is a better operator choice.

75:#and(tornadoes): No strategy applied

110:#and(map brazil): It is a very specific term to search for the map of Brazil, thus using and operator

123:#and(von willebrand disease): It is a very specific disease name, thus using and operator

149:#and(uplift yellowstone national park): It is a very specific query that want to look at the uplift which happens at the Yellowstone national park, thus use and operator.

154:#or(figs.keywords): No strategy applied

155:#and(#and(last supper) #syn(painting)): last supper is the name of painting that should occur together. Painting is what last supper is, but it could also be picture, sketch etc. Therefore, using syn to search synonym of painting.

163:#syn(arkansas.url): Arkansas is a very general query term. Thus, using syn to also include result like the state close to Arkansas.

164:#and(hobby stores): hobby stores supposed to occur together, thus using and operator

171:#and(ron howard): First name and last name should occur together.

177:#and(#syn(best) long care insurance): use #syn(best) because adjective like good, top, great long care insurance could also be included.

200:#near/4(ontario california airport): ontario california airport is equivalent to Ontario international airport Southern California. They might occur together but not necessarily be adjacent, thus using near operator.

3 Experimental Results for Unranked Boolean

	BOW #OR (Exp-3a)	**BOW #AND (Exp-3b)**	**Structured (Exp-3c)**
P@10	0.0000	0.0667	0.0867
P@20	0.0133	0.0900	0.1100
P@30	0.0133	0.0756	0.1178
MAP	0.0040	0.0928	0.1194
Running Time	13:05	2:07	02:02

4 Experimental Results for Ranked Boolean

	BOW #OR (Exp-4a)	**BOW #AND (Exp-4b)**	**Structured (Exp-4c)**
P@10	0.0867	0.2800	0.3000
P@20	0.1400	0.3533	0.3433
P@30	0.1867	0.4022	0.3400
MAP	0.1530	0.3188	0.2859
Running Time	15:09	02:04	02:00

5 Analysis of Ranking Algorithm Behaviors

Ranking algorithms get higher accuracy for both structure and unstructured query. The possible reason behind the low accuracy of unranked algorithm is that it treats all document the same. No term frequency effect is accumulated, once a document is find matching the query, every document shares the same score (1.0). In real world, however, different document supposed to have different importance even if they are discussing the same topic. For example, considering two document A and B, doc A is a medical paper presenting the researching result of a rare disease, doc B is an article talking about top 10 rare disease in 2020s. Even though the name of disease occurs in both documents, it is obvious that doc A would mention the term more frequent than doc B because it is the core idea of the whole paper. In unranked Boolean algorithms, such an effect is ignored and so triggered poor accuracy.

In terms of running time, there're no significant difference because the time complexity for calculating the score is the same. We still need to recursively go through every matched document and return the min and max score.

6 Analysis of the Effects of Query Operators and Fields

Different query operator and fields improve the accuracy which makes structure query achieve higher accuracy than unstructured one.

Query operator

Generally speaking, OR operator generate poor accuracy because it is a very unspecific operator that would introduce too many irrelevant documents. By comparing 3a and 4a with 3b and 4b, the conclusion is quite obvious. AND operator is sometimes too strict, and so introduce NEAR and SYN operator provide us with more flexible searching. Take Ontario California airport for example, its full name is Ontario international airport Southern California. Therefore, I use NEAR/4 to group those term together and the MAP increase from 0.0 to 0.06.

Field

The field url, title, and keywords produce higher accuracy. In experiment 4c, I use title to search moths, used url to search Arkansas, and used keywords to search figs. The reason I choose three of them is that they are too general search terms that might introduce generate too many irrelevant documents. Specifying those search term to appear in more important field significantly improve the accuracy of P10, P20, and P30.