Your Name: Tong Wei

Your Andrew ID: twei1

Homework 1

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)?

   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.

2. Structured query set

2.1. Summary of query structuring strategies

1. For those widely used proper nouns that refers to some particular entities, I use #NEAR/1(a b) to find the exact match.

2. If some word refers to a category, then it may appear in the keyword, url or inlink.

3. If one word is a description of another word, they are close to each other and their order might change, so I build the query with #OR(#NEAR/n(a b) #NEAR/n(b a)) .

4. For a single word, it is more likely to find the relevant documents whose title contains the word.

2.2. Structured queries

(1) 10:#NEAR/6(cheap.body internet.body)

Strategy 3. Cheap is used to describe internet, and there may be other words used to describe internet.

(2) 12:#djs.title

Strategy 4.

(3) 26:#OR(#AND(#NEAR/5(lower #NEAR/2(heart rate))) #AND(#NEAR/5(#NEAR/2(heart rate) lower)))

Strategy 1 and 3. Heart rate is a proper noun and lower is used to describe heart rate. This query is intended to find the documents which contains “lower heart rate” or “heart rate is lower”.

(4) 29:#AND(/#NEAR/2(ps 2) #OR(games.keywords games.url games.inlink))

Strategy 1 and 2. PS2 is a proper noun, PS2 is mostly likely to appear in the articles which are in the games topic. This query is intended to find the documents whose keywords or url contains games or the forwarded from the link that contains games.

(5) 33:#NEAR/1(elliptical trainer)

Strategy 1. This query is intended to find the exact match of elliptical trainer.

(6) 52:#OR(avp.keywords avp.title avp.url)

Strategy 4. Find the documents whose topic is relevant to avp.

(7) 71:#AND(#NEAR/1(living in) india.keywords)

Strategy 1. Living in is a common phrase, this query is intended to find the document related to living in india, so India should be its topic/keyword.

(8) 102:#AND(#NEAR/1(fickle creek) farm)

Strategy 1. Fickle creek farm is the name of a farm, and Fickle creek is also sufficient to describe it.

(9) 149:#AND(uplift at #NEAR/1(yellowstone.body national.body park.body) yellowstone.title)

Strategy 1. Yellostone national park is a proper noun and Yellowstone is very likely to appear in the title.

(10)190:#AND(#NEAR/1(brooks brothers) clearance)

Strategy 1. Brooks Brothers is a famous cloth chain, the two words should close to each other.

3. Experimental results

3.1. Unranked Boolean

	BOW #OR	BOW #AND	Structured
P@10	0.0100	0.0400	0.2000
P@20	0.0050	0.0200	0.1800
P@30	0.0033	0.0433	0.1767
MAP	0.0010	0.0142	0.0640
Running Time	00:57	00:03	00:02

3.2. Ranked Boolean

	BOW #OR	BOW #AND	Structured
P@10	0.1500	0.2500	0.3800
P@20	0.1800	0.2600	0.3200
P@30	0.1667	0.2767	0.2900
MAP	0.0566	0.0980	0.1203
Running Time	00:20	00:02	00:02

4. Analysis of results

4.1 #OR vs #AND vs #NEAR

There are three kinds of query operators: #OR, #AND and #NEAR. #OR is the least effective one, because it combines all the score lists into a new one which increases the size of the score. However, #AND and #NEAR operations intersect multiple score lists. If we sort the list according to their size in increasing order, we only need to check all the documents in the first score list. Therefore, the running time of #OR and #NEAR is shorter.

The precision of #OR operations is the lowest and its recall rate is the highest, because it returns all the documents that contains any terms of the query,. And the precision of #AND, #NEAR is much higher, because these operations have more constrains than #OR.

4.2 Ranked Boolean vs Unranked Boolean

In my implementation, the performance of Ranked Boolean is much better than Unranked Boolean. I uses max heap to find the top 100 entries. The Ranked Boolean and Unranked Boolean algorithms use the same Comparator. First it compares the score of entries, if the scores are the same, then it will compare the document name. For Ranked Boolean, the score of entries are different, so most of the time, we only need to compare the score. However, for Unranked Boolean, as the scores are the same (1), we just compare the document names which is a long string, it costs much more time than comparing scores. Therefor the running time of Ranked Boolean is much shorter than Unranked Boolean.

For all kinds of queries(OR, AND, Structured), the precision of Ranked Boolean is also much better than Unranked Boolean. In Ranked Boolean, documents with higher term frequency is assigned a higher score this is intuitive, because if a term appears many times in a documents, it is more likely that the document is relevant to the term or query. However, in Unranked Boolean, all documents which contains the terms are treated equally.

4.3 Structured Query

Assign proper fields(inlink, title, body, url, keywords) to terms is very useful because it can directly specify the topic, and category of the documents. But in some cases(e.g. it is difficult for us to specify which field term may appears) it may result in the missing of some important documents. In this experiment, the structured queries I build get the highest precision(MAP= 0.1203).