Web Information Management: Homework #1

Due on April 14, 2015

Professor Fang TTh 12:10

Rick Sullivan

Problem 1

Assume that a search engine returns a ranked list of 10 total documents for a given query. According to the ground truth labeling, there are 7 relevant documents for this query, and that the relevant documents in the ranked list are in the 1st, 3rd, 5th, 8th, and 10th positions in the ranked results.

1. Calculate Precision, Recall, F-measure, nDCG, at the 10 retrived documents.
2. Calculate the interpolated precision value for each of the following standard recall levels:

   $$\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$$

   for this individual query.
3. Calculate the average precision.

Part 1
Precision is determined by

$$\frac{\#\ of\ returned\ relevant\ documents}{total\ \#\ of\ returned\ documents}.$$

We have 5 relevant documents returned, and 10 total returned documents, so $P = 0.5 = 50\%$.

Recall is determined by

$$\frac{\#\ of\ returned\ relevant\ documents}{total\ \#\ of\ relevant\ documents}.$$

We have 5 relevant documents returned, and 7 total relevant documents, so $R = 5/7 \approx 0.714 = 71.4\%$.

F-measure is the harmonic mean of recall and precision,

$$\begin{aligned} F &= \frac{2PR}{P+R} \\ &= \frac{2(0.5)(0.714)}{0.5+0.714} \\ &\approx 0.5881 \end{aligned}$$

nDCG

The discounted gain list

$$1, 0, 1/\log_2(3), 0, 1/\log_2(5), 0, 0, 1/\log_2(8), 0, 1/\log_2(10)$$

gives DCG

$$1, 1, 1.63, 1.63, 2.06, 2.06, 2.06, 2.39, 2.39, 2.69$$

The ideal DCG is

$$1, 2, 2.63, 3.13, 3.56, 3.95, 4.30, 4.30, 4.30, 4.30$$

So the nDCG is

$$1, 0.5, 0.62, 0.52, 0.58, 0.52, 0.48, 0.56, 0.56, 0.56$$

Part 2

Problem 1 continued on next page. . .

Precision after n documents:

$$1/1, 1/2, 2/3, 2/4, 3/5, 3/6, 3/7, 4/8, 4/9, 5/10$$

Recall after n documents:

$$1/7, 1/7, 2/7, 2/7, 3/7, 3/7, 3/7, 4/7, 4/7, 5/7$$

So the interpolated precisions are

Recall level	interpolated precision
0.0	1
0.1	1
0.2	0.67
0.3	0.6
0.4	0.6
0.5	0.5
0.6	0.5
0.7	0.5
0.8	0
0.9	0
1.0	0

Part 3

The average precision is the average of the list in part 2, which gives

$$P_{avg} = 56.4\%$$

Problem 2

The *San Jose Mercury News* repository from 2000 to 2005 (i.e., 5 years) contains about 400 million word tokens, with the vocabulary size about 1 million. What would be a good estimation of the vocabulary size one would get in indexing the *San Jose Mercury News* repository from 2000 to 2010 (i.e., 10 years)?

Solution

$$V_R(n) = Kn^\beta$$
$$V_R(400,000,000) = K(400,000,000)^\beta = 1,000,000$$

β is usually between 0.4 and 0.6. Let's choose $\beta = 0.5$. This means

$$K(400,000,000)^{0.5} = 1,000,000$$
$$K = 50$$

Therefore, for double the amount of word tokens, we have

$$V_R(800,000,000) = 50(800,000,000)^0.5 \approx 1,414,213$$

Problem 3

Assuming Zipfs law with a corpus independent constant $A = 0.1$, what is the fewest number of most frequent words that together account for more than 22% of word occurrences (i.e. the minimum value of m such that at least 22% of word occurrences are one of the m most frequent words).

Solution

Given for Zipf's law

$$f * r = 0.1$$

we want to find the smallest n such that

$$\sum_{i=0}^{n} 0.1/r_i \geq 0.22$$

Which is the series $[0.1, 0.15, 0.1833, 0.2083, 0.2283]$. The fewest number of most frequest words that account for 22% of all word occurrences is 5 words.

Problem 4

Assume the following documents comprise your corpus:

Doc 1: banking on banks to raise the interest rate over the previous interest
Doc 2: jogging along the river bank to look at the sailboats
Doc 3: jogging to the bank to look at the interest rate
Doc 4: buzzer-beating shot banked in!
Doc 5: interest of the scenic outlooks on the banks of the Potomac River

Assume that you remove stopwords, lower cases, and do stemming.

1. Unranked Boolean retrieval with the following query:
   (bank OR rate) AND interest
2. Ranked Boolean retrival with the following query:
   (bank OR rate) AND interest

Solution

Stemmed word occurrences

	Documents				
	1	2	3	4	5
bank	2	1	1	1	1
rate	1	0	1	0	0
interest	2	0	1	0	1

Part 1
Documents 1, 3, and 5 would be returned. All documents contain bank and/or rate, so the documents that also have interest match the query.

Problem 4 continued on next page. . .

Part 2

	Documents				
	1	2	3	4	5
bank OR rate	3	1	2	1	1
(bank OR rate) AND interest	2	0	1	0	1

Documents would be returned in the following order: 1, 3, 5 (3 and 5 could be in either order). The OR operation adds occurrence counts, while the AND operation takes the minimum value of either option.

Problem 5

As we discussed in class, stemming can generally help improve retrieval performance in English text, but it may not always be the case. Choose one of the three choices: “stemming” or “no stemming” or “maybe stemming”, for each of the following 4 cases. You need to provide the reasons/explanations for each of your choices.

Solution

1. When the corpus is small and users care more about precision

   No stemming. Stemming reduces resources required to maintain and return relevant information (by reducing vocabulary size), which is not a problem with a small corpus. Also, stemming attempts to simplify the user's query, which can reduce the precision of queries, as multiple words may stem to the same base.

2. When the corpus is small and users care more about recall

   Stemming (or at least probably stemming). Stemming words will make sure that almost *any* relevant document will be returned, even if precision is reduced. Small corpuses (corpi?) will not benefit from stemming in terms of resources, but increased recall may be beneficial.

3. When the corpus is large and users care more about precision

   Maybe stemming. Stemming may give some resource benefits by reducing vocabulary size, but it comes at the cost of precision. Having stemming will increase precision.

4. When the corpus is large and users care more about recall

   Stemming. As mentioned earlier, stemming can both increase recall and improve performance with a large corpus.