CSIT 5930 Search Engines and Applications
Fall 2021 Homework 1

Due: See Canvas

Solution

1. This question is about the case study conducted by Blair and Maron, which has been covered in the lecture. The paper can be downloaded from Canvas Assignment homepage.

   Read this paper to answer the questions.

   a. **[10 points]** In the experiment, the lawyers and paralegals are allowed to revise a query as many times as they want. What might be the reason for this experimental design instead of evaluating the performance for each submitted query?

      The authors of the paper wanted to evaluate the performance of STAIRS, not the capability of the lawyers and paralegals in formulating a STAIRS query. If query formulation is not allowed, then any poor retrieval performance could be attributed to the limitation of the lawyers/paralegals in formulating a good query for STAIRS rather than the retrieval limitation of STAIRS.

   b. **[30 points]** State which of the following statements are true or false according to the paper and quote the text that support your answer.

      (i) The precision and recall obtained by the two lawyers are consistent with each other (i.e., the figures do not lead to contradictory conclusions).

      TRUE (Under Fig. 4 and the associated discussion) "Although there is some difference between the results for each lawyer, the variance is not statistically significant at the .05 level. … we can conclude that at least for this experiment the results were independent of the particular user involved."

      (ii) After the lawyers and paralegals have gone through rounds to formulate an information request into STAIRS' query, the precision and recall of the reformulated information request are significantly improved.

      FALSE (Under Fig. 4, 2nd paragraph) "… the values for Recall and Precision for the substantially revised queries (23.9 percent and 62.1 percent, respectively) did not indicate a statistically significant difference."

      (iii) If the lawyers are to write their own queries directly on STAIRS (not using the paralegals), the experiment found that the lawyers can get significantly better precision and recall than the paralegals.

FALSE (Page 294, 2nd column, last paragraph) "If it were true that STAIRS would give better results when the lawyers themselves worked at the terminal, the values of Recall for the lawyers would have to be significantly higher than the values of Recall when the paralegals did the searching. … the improvement is not statistically significant at the .05 level (z = -0.81)."

c. **[20 points]** Recall is time consuming to evaluate, because in principle the relevance of ALL of the documents needs to be judged against each query. Blair and Maron's experiment did not examine all the documents. Using your own words, describe how they chose subsets of the overall document collection for calculating recall?

Based on P. 291, 2nd column, last paragraph onwards, I write the following short summary:
They identified subsets of the unretrieved database that were believed to contain high concentration of relevant documents. Documents are then sampled from these subsets and presented to the lawyers for judging the sampled documents' relevance. [From the number of relevant documents in the sample and the sample size], the total number of relevant documents in the unretrieved database can then be estimated.

[Note that instead of random sampling directly from the unretrieved database, they used human judgment on where relevant documents are concentrated and focus on the concentration subsets. That is, if this human judgment is wrong, then the recall estimate is wrong!]

d. **[20 points]** In the paper, the authors gave a reason for the low recall of the experiment and some examples drawn from the document collection to illustrate the difficulty of using keywords to retrieve documents. Give a brief summary in, say, 2-3 sentences, to describe the reason and one example given by the authors. Do not copy the whole paragraph(s) from the paper.

The reason is that given a query, there are many relevant documents that do not contain the terms used in the query.

(i) Several examples are given. Any one of the examples is enough. The first example (Page 295, 1st column, last paragraph) given is about the keywords used to describe an accident. In addition to the obvious keyword “accident”, people use different keywords, such as “event”, “incident”, “situation”, “problem”, “difficulty”, “unfortunate situation”, etc., to refer to an accident.

(ii) The second example is about the use of different names to refer to the same thing, from "trap correction" to "wire warp", to "shunt correction system" and finally "roman circle method" and "air truck".

Another example illustrates a general concept encompassing several specialized concepts. For example, instead of "steel", the specific steel product names are used in the documents, e.g., "girders", "beams", "frames" and "bracings".

2. **[20 points]** Suppose there are only 5 unique terms (numbered 1 to 5) in the collection, which contains a total of 10 documents. These five term's term frequencies in a document *D* and their document frequencies are given below:

TF	DF	IDF
$tf_{D,t1} = 2$	$df_{t1} = 1$	$idf_{t1} = log2(10/1) = 3.32$
$tf_{D,t2} = 0$	$df_{t2} = 2$	$idf_{t2} = log2(10/2) = 2.32$
$tf_{D,t3} = 1$	$df_{t3} = 3$	$idf_{t3} = log2(10/3) = 1.74$
$tf_{D,t4} = 5$	$df_{t4} = 2$	$idf_{t4} = log2(10/2) = 2.32$
$tf_{D,t5} = 2$	$df_{t5} = 10$	$idf_{t5} = log2(10/10) = 0$

(a) [5] Write down the document vector when tf/tf_{max} * idf weighting is used.

t_1 = 2/5 * 3.32 = 1.33
t_2 = 0/5 * 2.32 = 0
t_3 = 1/5 * 1.74 = 0.35
t_4 = 5/5 * 2.32 = 2.32
t_5 = 2/5 * 0 = 0

(b) [15] Given the query vector, $Q = \langle 1, 0, 0, 1, 0 \rangle$, compute the cosine similarity values between *Q* and *D* by first writing down the inner product and the magnitudes of Q and D.

inner product = 1.33 + 2.32 = 3.65
|Q| = sqrt(2) = 1.414
$|D| = sqrt(1.33^2 + 0.35^2 + 2.32^2) = sqrt(7.27) = 2.70$
Cosine(Q,D) = 3.65 / (1.414 * 2.70) = 0.956