Dublin City University
School of Computing
CA4009: Search Technologies
Laboratory Session 4 & Laboratory Session 5

November 2019

Module Coordinator: Gareth Jones

Laboratory Tutors: Tianbo Ji, Abhishek Kaushik, Yasufumi Moriya, Procheta Sen

1 Introduction

Laboratories 4 and 5 extend your investigation of the topic of ranked information retrieval (IR) in laboratories 1-3 to examine relevance feedback and query expansion. The laboratories focus on *pseudo* relevance feedback, as introduced in lectures, using the Robertson offer weight ($ow(i)$) for the probabilistic model of information retrieval.

2 Laboratory Reports and Submission

Similar to the previous Laboratories, you should create an electronic report file for this laboratory.

- Include the title and date of the laboratory, and your names at the beginning of your report.
- For each activity described below, enter your answers into your report file, making clear which section your response relates to.
- At the end of the laboratory session you should upload your report to the CA4009 loop page via the link for this laboratory.
- If you do not finish the exercises to your satisfaction, you can complete the assignments in your own time, and submit a revised report via loop. The latest date for submission of the extended report is **one week after the laboratory 5 session**.

3 Query Expansion using Relevance Feedback

Relevance feedback (RF) in information retrieval seeks to improve retrieval effectiveness by using feedback data from the current search run for a query to improve a further search run for the same information need. The first stage in the RF process is to mark documents retrieved in a current run as relevant, and in the next stage of RF, a query expansion method is used to select terms from the documents which have been labelled as relevant. These terms are then added to the query for a subsequent run.

The underlying objective of RF using QE is select expansion terms which are found to be associated with relevant documents with the assumption that adding these terms to the query will make the query a

better statement of the user's information need, and that using this expanded query for a search run should improve the rank of relevant documents and potentially retrieve additional relevant documents since the expanded query is more likely to match with terms contained in relevant documents than the original query.

As described in lectures, document relevance information can come either directly from user feedback (the user directly marks a document as relevant), from observing user behaviour (e.g. which documents they click on), or by assuming that a selected number of the top ranked retrieved documents are relevant (the pseudo relevance feedback approach). This laboratory adopts the pseudo relevance feedback approach, i.e. in your investigations you will assume that a number of top ranked documents from the previous run with the current query are relevant.

3.1 Robertson's Offer Weight $ow(i)$ method

RF using the probabilistic model of IR was introduced in lectures. This section summarises the details of this method.

To perform query expansion with the probabilistic information retrieval (IR) using Robertson's method, you first need to compute the Robertson/Sparck Jones relevance weight $rw(i)$ using the following equation.

$$rw(i) = \log \frac{(r(i) + 0.5)(N - n(i) - R + r(i) + 0.5)}{(n(i) - r(i) + 0.5)(R - r(i) + 0.5)}$$

where: $n(i)$ and N are defined in the previous laboratories, R and $r(i)$ are new variables.

$n(i) =$ the number of documents term $t(i)$ occurs in,
$N =$ the total number of documents in the collection archive.
$r(i) =$ the number of **known relevant** documents term $t(i)$ occurs in,
$R =$ the total number of **known relevant** documents in the collection archive.

In pseudo relevance feedback the value of R is assumed as part of the operation of the algorithm. For example, if you assume that the top 5 ranked retrieved documents are relevant the value of R = 5. The value of $r(i)$ for each of the terms i which occurs in one of these assumed relevant documents can then be computed as the number of these documents it occurs in. It should be obvious to you that $r(i)$ will always be $\leq R$.

The value of $rw(i)$ for each term i can then be used to compute the Robertson offer weight $ow(i)$ for each term i term as follows:

$$ow(i) = r(i) \times rw(i)$$

where $r(i)$ has the same meaning as above.

3.2 Expanding the Query

All the terms occurring on one or more of the top R ranked documents is a potential query expansion term. The purpose of $ow(i)$ is to determine which ones are likely to be the most effective. That is, which ones when added to the query are likely to improve its effectiveness in retrieving relevant documents at a higher rank in a subsequent retrieval run for this query.

To decide which terms to add, the terms occurring in the top ranked documents are ranked in decreasing order of $ow(i)$. The top ranked terms from this ranked list are then selected as the expansion terms to be added to the original query.

A fixed number of the top ranked terms are then added to the query and search is run using this expanded query, and the new ranked list returned. The number of expansion terms to add to the query is determined in a prior training phase to decide the optimal number of terms to add to give the best MAP value for a set of training queries.

The purpose of this laboratory is to explore the impact of adding terms to a query based on the Robertson offer weight.

3.3 Query Expansion based on $ow(i)$

In addition to the name of the term and the term frequency in this document, the search interface used in the earlier laboratories also shows the value of $(N/n(i)$, where as defined above, N is the total number of documents in search collection and $n(i)$ being the total document frequency of a term i in the collection.

To calculate $rw(i)$, you also need to know N and $n(i)$. N is the total number of documents in the collection, which is approximately 500,000 (the actual value of N isn't exactly 500,000, you should consider why using a number of N which is $\approx N$ is OK). To calculate $n(i)$ for each term i of interest, you can use N with the returned value of $N/n(i)$.

As stated above, for pseudo relevance feedback, the value of R is the number of documents which you assume to be relevant.

To determine the $r(i)$ you need to examine the contents of the documents which you have assumed to be relevant.

In the ranked retrieval list you see the most significant terms for this document ordered in decreasing term frequency for each document.

In addition, you can see the complete contents of a single document by entering the following web address into your browser:

`http://136.206.48.37:8084/IRModelGenerator/ DocumentViewer?name=DocName`

where "Docname" is the name of the document whose contents you wish to view.

By counting the number of different assumed relevant documents that a term occurs in, you should now be able to calculate the $ow(i)$ value for a term.

4 Experimenting with Query Expansion in IR

Once you have worked out how to calculated the Robertson $ow(i)$ method for a selected term, in this part of the laboratory, you will investigate query expansion impacts the behaviour of a ranked retrieval list.

4.1 Initial Investigations

- Choose one of the query topics from the TREC topic set used in the previous laboratories.
- Enter the "Title" field of this topic into the interactive search interface using the BM25 ranking function.
- Briefly examine the details of the top ranked retrieved documents.
- Assume that some number of the top ranked documents (say 5) are relevant. This will be your initial value of R.

- Open the full document view of a several of top ranked documents as described above. Look at the most significant terms present in this documents based on their term frequency values in these documents. Select terms which appear in multiple of your assumed relevant documents, and calculate the $ow(i)$ for these terms.
- Based on your calculated $ow(i)$ values, add some of these terms to your original query, and run the search operation again.
- Compare the ranked retrieved list for you expanded query with the original one. How do the ranks of retrieved documents vary between the lists? Have some increased in rank? Have some decreased? Have some disappeared from the top of the list? Have new ones appear?
- Repeat the above steps with different numbers of expansion terms.

4.2 Investigations using qrel data

- Find your selected topics statement in the TREC qrel file used in the previous laboratories.
- Examine the top ranked documents from your retrieval lists to identify relevant documents listed in the qrel file. You should start with the retrieval for the original unexpanded query.

  You should be able to use this information to work our the precision at ranks of 5 and 10 for each ranked list.
- Work out the precision at ranks 5 and 10 for your initial ranked lists and the lists retrieved using the expanded queries.

  How are the precision values affected by query expansion?

Select at least two additional topic statements from the TREC topic sets and repeat the above procedure.

4.3 Extended Investigation: Examining "true" relevance feedback

Using the information in the qrel file, you can repeat the above investigation using feedback only documents which are actualy relevant. For example, instead of assuming that the top 5 ranked documents are relevant, you might find that the documents ranked at positions 2 and 5 are marked as relevant in the qrel file.

$ow(i)$ can recalculated using this true relevance information, and the query expanded based on these new values.

Repeat the above investigation using qrel data to examine the difference between pseudo relevance feedback and true relevance feedback.