INLS 490-154W: Information Retrieval Systems Design and Implementation. Spring 2009.

8. Evaluation-1

Chirag Shah*
School of Information & Library Science (SILS)
UNC Chapel Hill NC 27599
chirag@unc.edu

1 Introduction

While an ideal IR system should get the relevant information for any information request by any user in any situation, there is no system that satisfies all of these. The question is then how well a given system is doing to match its expectations, or better yet, how well it is doing compared to some other system. Measuring the retrieval performance of an IR system has been the one of the biggest challenges for decades. In most situations, it is a hard problem to evaluate an IR system without user judgments.

Here we will look at some of the ways in which we can talk about the "goodness" of an IR system. To keep things simple, we will only consider objective relevance, i.e., if a retrieved document is relevant or not. Also, the only context we will consider is the topic of the information need, and not the situation or other such factors.

2 Recall and precision revisited

To begin our discussion, let us review the notion of recall and precision that we had seen before. In Figure 1, a van diagram is given showing a set of relevant documents (R) for an information need, and a set of retrieved documents (R') by some IR system. Recall is the portion of relevant documents returned, and precision is the portion of the returned document that is relevant. Using Figure 1, this can be formulated as
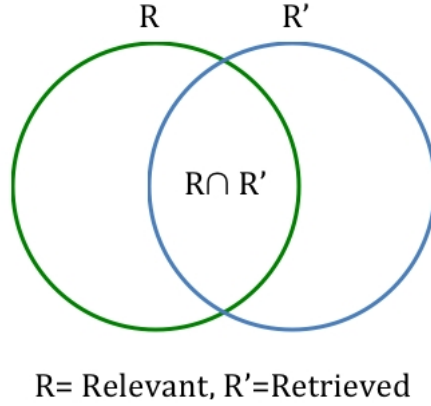
R= Relevant, R'=Retrieved

Figure 1: A model to understand recall and precision in IR

$$\text{Recall} = \frac{R \cap R'}{R} \tag{1}$$

$$\text{Precision} = \frac{R \cap R'}{R'} \tag{2}$$

Now the question is how we can extend these definitions that are based on set notion to rank lists that we usually get at the end of a retrieval process. We can create sets from the rank lists, for which we have several options, and measure recall and precision. These options are given below.

1. At every new document.

2. At every new relevant document.

3. At fixed rank value cutoff such as measuring precision at rank 10.

4. At fixed recall points such as measuring precision at 20% recall.

Let us understand this by an example. Figure 2 shows rankings as given by two different systems or algorithms. For each of these rankings, calculations of recall and precision are shown at every document. Similarly, we can compute these values at other points listed above.

3 Single value measures

By looking at Figure 2, it is not very clear which ranking is better as they both have the same recall and precision values at the end of the list. Of course, this "goodness" depends on the task at hand, but it is often useful to come up with one final number indicating the retrieval effectiveness. One simple way of doing this is averaging precision values. Average precision is calculated by averaging precision when recall increases.

In Figure 2, these points are indicated by circles on the recall values. If we take precision values at those points and average them, we get 62.2% for Ranking #1 and 52.0% for Ranking
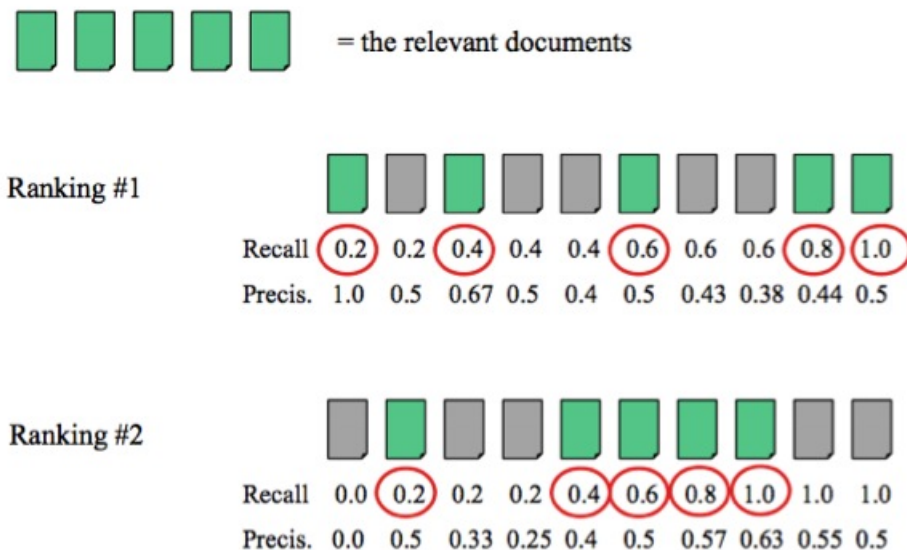
Figure 2: Calculating recall and precision with rank lists (Courtesy: James Allan, UMass Amherst)

#2 as average precision. Thus, using this measure we can immediately say that Ranking #1 is better than #2.

Often we have a number of queries to evaluate for a given system. For each query, we can calculate average precision, and if we take average of those averages for a given system, it gives us Mean Average Precision (MAP), which is a very popular measure to compare two systems.

Another such single value measure is R-precision. It is defined as precision after R documents retrieved, where R is the total number of relevant documents for a given query. Average precision and R-precision are shown to be highly correlated. In Figure 2, since the number of relevant documents (R) is 5, R-precision for both the rankings is 0.4 (value of precision after 5 documents retrieved).

4   Evaluating using trec_eval

Let us now see how we can use a utility developed by NIST, called trec_eval[1] to compute the above measures. One important requirement for computing any of these measures is the availability of relevance judgments. As a part of TREC runs every year, NIST provides such relevance judgments using a large number of documents assessed by human assessors. This file is formatted as

    <topic_id> 0 <doc_id> <relevance>

Where the first column has the topic number, second column is redundant (but still preserved due to historic reasons) carrying value '0', third column has the document ID, and the fourth column has relevance judgment - '0' for non-relevant and '1' for relevant.

---

[1]Available from http://trec.nist.gov/trec_eval/

Topics are also provided by NIST. Each topic has several title, description, and narrative components. These components can be used to automatically creating queries.

Once we have results (in TREC format) obtained from a retrieval run, we can use that result file along with the relevant judgments provided by NIST to evaluate retrieval performance using `trec_eval`. This utility is used as

    `trec_eval <options> <rel_file> <ret_file>`

Some of the useful options are:

-o: Print requested measures in old non-relational format

-q: In addition to summary evaluation, give evaluation for each query

-a: Print all evaluation measures, instead of just official measures

From the output printed with the above utility, one can obtain values of several evaluation measures. These values can then be reported or compared with other systems for further analysis.

5   Summary

1. Recall and precision are the most popular measures of evaluation in IR.

2. Recall increases with every new relevant document introduced in the given set. Precision drops with every new non-relevant document introduced in the given set.

3. On average, as recall increases, precision drops.

4. Average precision, mean average precision (MAP), and R-precision provide us with single numbers to compare retrieval performance.