EVALUATION OF IR SYSTEMS

Reading: Chapter 3, Baeza-Yates.

Computer systems created for information retrieval have an underlying assumption of human language, a specific mathematical model that guides the logical and physical design of the software, as well as the usual document collection, and means for accepting end-user queries. Collectively, these form the *implementation*: each implementation of an IR design differs from every other design. The differences may be subtle. How might we determine whether or not one implementation is "better" than another? Which metric do we choose and how might we defend our choice?

Along with the usual document collection and means for accepting end-user input, IR computer systems designers have certain assumptions about language use and the concept of "meaning", definitions of "information", and logical models, usually expressed mathematically, that form the actual *implementation*. Each IR implementation varies from every other implementation. Evaluation, then, is used to help determine which IR system is "better" than another. The questions of what *should* be measured and what constitutes "better" are open.

In addition, IR systems are often divided into actual systems open to public use and test implementations. The most famous of modern test implementations is TREC (url here). The first large-scale IR test, and the majority of systems today, assume that the best judge of a document's relevancy to an aggregate of end-users is a human judge. Cyril Cleverdon and his staff conducted this first large test, called "Cranfield I", and a later experiment, Cranfield II. In these tests the assumption was that human experts would determine a document's relevancy to several queries. Then the logical model and physical implementation – the actual IR program – would retrieve and rank documents in response to these test queries. The measure of system effectiveness was based on the number of documents the system ranked as relevant compared to those of the human judges. This raises many questions because not all human judges will determine a document's relevancy the same. One measure of this *inter-indexer consistency* is *Cohen's* α (pronounced "Cohen's alpha).

If one had to select a new retrieval system for a library or was asked to write an evaluation paper to select between several competitors, what criteria might one establish to assess each system? Perhaps one would use the common computer science measures of *time* and *space*? Perhaps some measure of *end-user satisfaction*. Is there a particular statistical model you could justify for your decisions? As you can see, evaluation of IR systems is not an easy task.

The best approach is to delineate the goals of the IR system. For example who is the target audience? Are there properties of that audience that should be recognized as influences in evaluation? For instance, would you compare a group of biology researchers to the general public? What about a group of biology researchers compared to all bio-medical information seekers? For our discussion, let's focus on determining some principles for a general audience, such as at a university library system.

Another goal of the IR system is what measure do we want to use: something standard, such as *time* and *space*, the traditional computer science measures, or something novel, or perhaps a combination? In computer science, most systems are measured for their maximum speed (time) at returning a response or completing some task. This measure is called "big o", *O*. Computer systems are then assessed by comparing one system's *O* score to another's. Another measure, Θ, ("big theta") is used to compare an average of the system's performance in worst-case scenarios. While these are valuable measures, we are more interested in how well the IR systems helps humans' desire for information than in how fast it takes a computer system to do some task.

IR assessment is founded on the assumption that there is always a set of documents in the collection that match the user's query, to some degree. Furthermore, it is assumed that all retrieval sets will include documents that should not be included and that some documents are not part of the set, but should be. This leads to four elements used in some way or another in all IR evaluation: documents that are retrieved, those that are not received; those that are relevant, and those that are not relevant. From these we determine *recall* and *precision*. Recall and precision were introduced in 1955 by Kent [Kent et al., 1955; Saracevic, 1975].

Recall and Precision

Let's return to our document collection. A user wants to locate, retrieve, and rank documents from that collection in response to his information need. The belief is that somewhere in the collection, there is at least one relevant document. We, the researchers, would like to evaluate how well our IR system (the framework) matches the user's query to the collection's representation. There is a relationship between the number of relevant documents in the collection and the documents that are actually retrieved. This relationship is "recall."

Recall is approximately the number of *retrieved* (Ret) documents, unioned ∩ with the *number of relevant* (Rel), also called the Answer set (A), divided by the number of relevant. Here are two ways of expressing the idea.

$$\mathrm{Re}\,call \equiv \frac{|\mathrm{Re}\,t \cap \mathrm{Re}\,l|}{|\mathrm{Re}\,l|}$$

Alternatively, we could label the relevant documents R. The number of relevant documents is $|R|$. In response to the query, the system generates an "answer set" *A*; it retrieves a certain number of these documents, $|A|$. The intersection of the relevant documents with the retrieved documents is $|Ra|$. Recall could be expressed as

$$\mathrm{Re}\,call = \frac{|Ra|}{|R|}.$$

[Note that some authors use the equivalency operator to emphasize the approximation of the relationship, others use the equal sign.]

As anyone who has ever used a search engine knows, not all documents within the retrieved set of documents are what we want. The measure for the fraction of those documents in the retrieval (or answer set) that *are* relevant is called *precision.* Using the same terms above, we define *precision* as

$$\mathrm{Pr}\,ecision \equiv \frac{|\mathrm{Re}\,t \cap \mathrm{Re}\,l|}{|\mathrm{Re}\,t|}$$

Using the other notation with *A* for answer set, precision is defined

$$\mathrm{Pr}\,ecision = \frac{|Ra|}{|A|}$$

Note that these are not the only types of recall and precision. These are discussed later. In addition to recall and precision, some authors focus on the *probability* that a document will be relevant, given that it is retrieved: $Pr(Rel|Ret)$. [Read this as "the probability that a document is relevant, given the retrieved documents."]

Some researchers prefer to recognize that some documents are irrelevant in the retrieval set. This measure, called *fallout*, focuses on the percentage of irrelevant documents to the retrieved:

$$\mathit{Fallout} \equiv \frac{\left|\overline{\mathrm{Re}\,t} \cap \mathrm{Re}\,l\right|}{\left|\overline{\mathrm{Re}\,t}\right|}$$

[A bar over a symbol is its *complement*, meaning the opposite. So $\overline{\mathrm{R}}$ is the opposite of R. If R means "retrieved", $\overline{\mathrm{R}}$ means "not retrieved."]

Remember that we might be interested in measuring our IR system for a general audience and so do not want to focus on a specific query, but rather measure our system against a set of similar, generalized queries. [E.g., instead of measuring all users of the query term "cat", we might want to use "cat", "feline", "tabby" and so on.] If so, then we could create another metric, *generality*, *G,* and use it in how we measure precision:

$$\mathrm{Pr}\,\mathit{ecision} = \frac{\mathrm{Re}\,\mathit{call} \bullet G}{\mathrm{Re}\,\mathit{call} \bullet G + \mathit{Fallout} \bullet (1 - G)}$$

This value would also generate a probabilistic measure $\Pr(\mathrm{Re}\,t \mid \overline{\mathrm{Re}\,l})$.

The relationship of recall and precision are demonstrated below. Notice in the second illustration that there can be *high-recall* retrieval that is different from *high-precision* recall. For end-users, high-recall will retrieve many hits; high-precision will retrieve few hits, but they are (theoretically) extremely close to the user's query. "Close" is a numeric measure of *similarity* between query and document. Whether or not the closeness of similarity is real in the end-user's mind is the big challenge! A user may enter the query term "cat" because he *intends* only that term to reflect what he has in his mind as the referent. "Tabby" may not be what he intends, but the IR system may conflate the two terms.

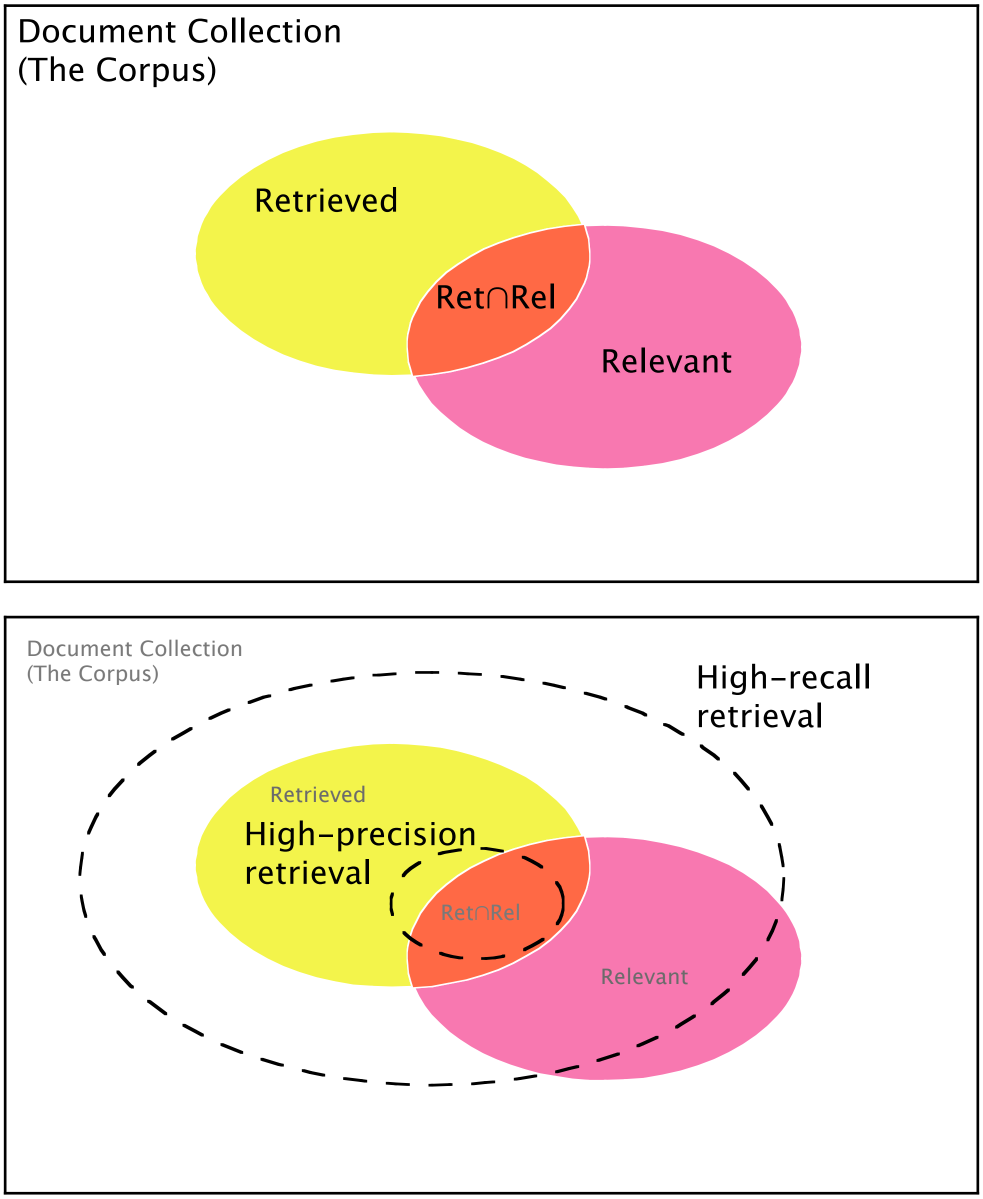

Figure of relevant vs. retrieved.

Ranking Retrieval Results

Let's say our document collection consists of 100 documents. In this collection, let's say we know there are 10% of those documents that are about the topic "fish." When an end-user submits the query "fish" to our collection, we expect that the IR system will retrieve and rank 10 documents. If so, then the *recall* of each document in this system is |retrieved ∩ relevant| ÷ |relevant|, or |100% of the retrieved ∩ 10% of all relevant| ÷ |10%|, 110/10 = 10% recall. In other words, document no. 1, d_1, reflects 10% of all the retrieved documents (which in this example, just happens to be 10 documents out of the total 100). For any document *i* in this collection, the best recall value is 100%, which may be referred to only as 1.

What is the precision for this query? Precision is |retrieved ∩ relevant| ÷ |retrieved|, or |10% ∩ 100%| ÷ |10%|, or 100%. Notice that precision and recall have an *inverse relationship*: as recall increases, precision decreases; when precision increases, recall decreases.

Baeza-Yates and Ribierto-Neto provide a great example (1999, p. 76-77), which is noted here. In response to a query, a retrieval set R_q is created of the following documents, d: $R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$. Of this retrieval set, human judges determined that ten documents are relevant to query q. In addition to retrieving these documents, they are *ranked* in this order:

1. d_{123}•
2. d_{84}
3. d_{56}•
4. d_{6}
5. d_{8}
6. d_{9}
7. d_{511}
8. d_{123}
9. d_{187}
10. d_{25}•
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_{3}•

The • means the documents are relevant to the query.

The first document (d_{123}) is ranked as the most relevant. This document corresponds to 10% of all the relevant documents in the set R_q. Thus, the precision is 100% at 10% recall. [The same as the above example.] But retrieval is rarely perfect! The second document, d_{56}, ranked #3, is the next relevant document. The precision of this document is about 66% because of the first 3 documents (d_{123}, d_{84}, d_{56}) only 2 of the 3 are relevant, or 2/3 of the total 3, which is about .66 or 66%. The set of 3 documents also is viewed as being two relevant documents out of the total 10 in the retrieval set, or 20% recall.

Precision and recall are usually measured at 10% intervals (which yields 11 different numbers: 0%, 10%, … 100%). [Graph example of 11 measures.]

Imagine our theoretical general college IR system is being tested by hundreds of students in the course of year. How would we use the results of their searches to determine the effectiveness of the system? Perhaps this would be too unwieldy so we select a set of students to conducts queries of interest to them over a week. Now we can capture a set of single, related queries and from precision-recall values establish an *average precision at all levels of recall* and we this metric to determine how useful the IR system might be, using this algorithm:

$$\overline{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

$\overline{\text{P}}$(r) is the average precision at each recall level, *r*

N_q is the *n*umber of *q*ueries

$P_i(r)$ is the precision at recall level *r*, for the *i*-th query.

Some researchers prefer to interpolate the values, which we will not discuss (see p. 78). However, now we have a way to compare any number of algorithms for measuring precision and recall; we can use this same approach for comparing two or more different IR systems' performance when processing the same queries.

Problems with precision and recall:

Precision and recall are not perfect measures but they are common in the IR literature. In what ways are they not entirely appropriate? As you may have noticed, there is no way really to know exactly how many documents in a collection actually are relevant. Even if every document were assigned a subject heading, for instance, the user's query may not match the document's "aboutness." Perhaps instead of using two different values (one is precision, the other recall), would it be better to have a single value that somehow uses both, such as Robertson's *E*? Finally, document collections are rarely closed and queries are not submitted in isolation (batch processing). Are there other *concepts* of retrieval effectiveness that could be considered? Perhaps we could study the *difference* in a system's retrieval set or behavior based on the user's point of view.

Evaluation in general seems to fall, then, into two camps. One uses only the variations of mathematical measures of how *similar* the query and document tokens are. The other uses *relevance feedback*: input from the user that may alter how documents are ranked, or alter which terms are used during searching, or the weights (discussed later) are assigned for ranking.

Variations on a theme:

For completeness' sake, here are some variations on evaluation you may want to pursue on your own.

Normalized recall and precision

Harmonic mean

E measure

F measure

Generality

One-parameter criteria

Sliding ratio

Expected search length (ESL)

Operating characteristic curves

Coverage / Novelty

TEST COLLECTIONS

Most retrieval effectiveness studies are performed on test collections, in laboratory settings. Here is a list of some test collections:

- TREC
- RAVE: a relevance assessment vehicle
  - RAVeUnion
  - RAVePlan
  - Interactive Rave
  - RAVeCompile
- CACM (Communications of the ACM)
- CISI (aka ISI) [Small 731]
- SMART