Information Retrieval - Lecture 3 - Metrics

colm.oriordan

October 2020

1 Introduction

Given the range of IR models available, of weighting schemes available and of pre-processing options available, it is pertinent that we have means to allow comparisons between IR systems.

One metric that we could apply is to try and gauge the functionality of the system. How well and how reliably does an IR system satisfy the functional requirements? This approach, of course, can be applied to any type of software system. This typically involves standard testing procedures.

The response time (efficiency) of the system can also be taken as a measure of the system's performance. The system can be empirically tested under average and extreme cases by subjecting the system to series of queries. We can also measure the performance more accurately by analysing the indexing strategies and the comparison algorithms.

In IR systems, we are also interested in evaluating the system on the quality of the performance. We wish to measure how well an IR system's results satisfy a query or a user's information request. Typically, we wish to somehow quantify the quality of the returned list of items.

The performance of such IR systems is usually measured through the use of a *test collection*. The test collections have been created for many tasks within information retrieval including information filtering and adhoc information retrieval. These collections usually comprise a large set of documents, a set of queries representing information needs and a set of human judgements indicating the relevance of each document to each query. By running the system for each of the queries, we can estimate the quality of the IR system by comparing how similarly the system performs to the human judgements, i.e., how close was the answer set of the system to the set of documents deemed relevant.

This is approach has been the dominant approach in empirical information retrieval; however much recent work has questioned this model of evaluation

in dealing with the nuanced complexities of more modern information retrieval systems.

Also, some ambiguity arises with the concept of relevance. In most papers and experiments the concepts relates to how well a document matches to a query that was presented. We may also be interested in how well the document matches the information need. The term *pertinence* had been used to describe this concept. A large difference in pertinence and relevance often indicates that the system is failing to capture, or the user of failing to express, their information need through the system.

As mentioned in an earlier class, a system can have users with one-off adhoc information needs or can involve interactive sessions with querying interspersed with re-formulation and browsing. The former type of session with one-of queries has been the focus of much empirical research in IR. For this approach, the quality of the returned set (or answer set) is of most concern to researchers. For the latter approaches involving more user interaction other metrics are also considered in addition of the quality of the answer; these include user-effort required, user satisfaction and length of searches.

2 Precision and Recall

Precision and recall are the two metrics most widely used to measure the rerieval effectiveness of a system. Given a document collection D and a query q, let R be the set of documents in D relevant to q. Let A be the set of documents actually returned by the system in response to the user query q. Let RA be the set of documents that are relevant to the user and in the answer set returned to the user.

Precision is defined as the percentage of documents returned to the user that are actually relevant to the user:

$$precision = \frac{|RA|}{|A|}$$

Recall is defined as the percentage of relevant documents in the whole collection that are returned to the user.

$$recall = \frac{|RA|}{|R|}$$

We can use these metrics to capture the quality of an answer set.

3 Precision Recall Graphs

In most systems, we do not just return a set of the user. In most systems, for example systems adopting a vector space model with some suitable weighting

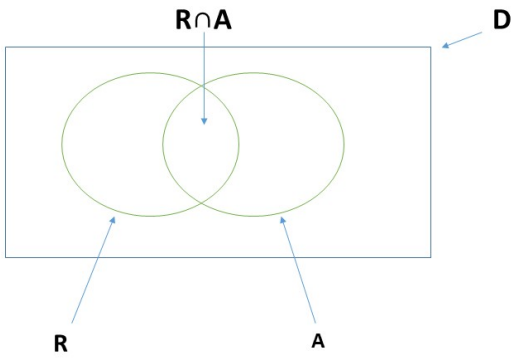

Figure 1: Venn Diagram for Precision and Recall

schemes, we return a ranked list to the user. Documents are typically sorted and presented in decreasing order of estimated relevance.

We can chose a threshold (top k documents, all documents above a certain score) and use that to define a set and then define precision and recall accordingly.

However, we obtain a more accurate representation of the quality of the ranking by plotting precision against recall for a number of points. We can calculate these pairs of values for different points along the ranked list.

An ideal system, for a recall value of 1, the precision would also be 1, i.e. all relevant documents were ranked above all non-relevant documents,

To illustrate the development of a precision-recall graph, consider the following 'toy' scenario. Given a document collection of size 20 ($|D| = 20$) which contains 10 relevant documents for a query ($|R| = 10$) and the top 10 ranked documents in the answer as follows.

$$\mathbf{d_1}, \mathbf{d_2}, d_3, \mathbf{d_4}, \mathbf{d_5}, d_6, d_7, d_8, \mathbf{d_9}, d_{10}$$

Also, assume that given the human judgements that those marked in bold are the relevant ones (i.e d_1, d_2, d_4, d_5, d_9).

Considering the first document only, we can calculate the a precision-recall pair. The precision-recall pair is $(1, 0.1)$. Note if we are considering the first document, hence $|A| = 1$.

Considering the first two documents only, we can again calculate another precision-recall pair - $1, 0.2$. Our precision is still 1 (2/2) and the recall increases to 0.2 (2/10).

For the first three documents we have a precision-recall pair $(0.67, 0.2)$. As we have encountered an non-relevant document, our precision falls and our recall remains the same.

We continue and generate a set of points and the plot precision versus recall.

In reality, generating a precision-recall pair for every document in the ranked list is impractical as we may be iterating through a list of very large list (e.g. several hundred thousand). In these scenarios, we usually the following approach - a limited set of precision-recall points are collected (typically 9 points corresponding to recall values of 10%, 20%, 30% etc.). We descend through the ranked list until the next recall point is reached; we then record precision at this point to generate a precision-recall pair. Note, that this may involve interpolating between two recall points.

Furthermore, a precision-recall graph for one query against a system does not give us a meaningful insight into the true performance of the system. It is more common and more useful to plot the average of a number of query runs against the system. The average precision for the various recall values is calculated (given N queries):

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

4 Single value measures

Oftentimes, people prefer to summarise the performance of a system through a single value. Precision recall graphs are in common usage and capture the behaviour of a system. However, if we are comparing several systems across many collections, the resulting large set of graphs can be unwieldy.

In the literature there have been many approaches used; these include, among others, precision when the first relevant document is found, mean reciprocal rank (inverse of the rank of the first relevant document, averaged across all queries, mean average precision and the E measure.

MAP (or mean average precision) has been widely used. In this case, we summarise the list of precision-recall pairs, by averaging the precision values - this is referred to as average precision. The mean of this across all queries is referred to as the MAP of the system.

The use of the harmonic mean has been proposed as a means to combine both precision and recall into the one score. The harmonic mean at a point is:

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{p(j)}}$$

Note that $F(j)$ is high if and only if both $r(j)$ and $p(j)$ are both high.

A variation on the harmonic mean is the E measure which allows one to vary the relative importance of precision and recall:

$$E(j) = \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{p(j)}}$$

5 Summary of precision-recall related measures

Precision recall measure (and variations) are useful metrics for a number of reasons. They are in widespread use. Precision-recall are definable quantifiable measures and they provide a useful summary of a system's behaviour.

However, there are certain limitations and criticisms that can be levelled at these measures. In many domains we do not have available test collections and even in domains where we have large collections, meaningfully calculating recall is not possible as the collection is too large. Moreover, these metrics are useful in capturing the usefulness of the return set for a given query; they are useful for one off adhoc queries but are inadequate in capturing the complexity of an interactive session.

Another important limitation is the underlying assumption that the set of relevant documents for a given query is the same for every user. We need to distinguish between the notion of a document being 'on-topic' for a query and the notion of a document set being relevant to the user's information need. Indeed, there has been some research showing that the measure of precision and recall do not reflect user satisfaction.

6 User Oriented Measures

In recognition of the final criticism mentioned above, there have been growing attempts to incorporate the notions of the user's experience in the process. This has involved user trials - which are expensive to run and also involve a lot

of more variables (users' previous experience, expectations, knowledge of the domain etc.). Companies with large user bases (e.g. web search companies) can trial variants of algorithms for large sets of their use bases and compare behaviours.

In some experimental settings, researchers have proposed new metrics to capture some aspects of the user's knowledge. Two well known metrics are those of *coverage* and *novelty*.

Let U be a subset of the relevant documents in the collection previously known to the user ($U \subset R$). Let AU be the set of returned documents known to the user.

We can define coverage as:

$$Coverage = \frac{AU}{U}$$

Coverage will be the percentage of the known relevant documents that are returned in the answer set to the user. In many search scenarios, users are searching for information they know to be in the collection. In these cases, high coverage is good. A high coverage score can give a user higher confidence that the system is performing well.

A related concept is that of *novelty*. Let New refer to the set of relevant documents returned to the user that were previously unknown to the user. Clearly, if coverage is high, then novelty will be low and vice versa. We can define *novelty as* as:

$$Novelty = \frac{|New|}{|New| + |AU|}$$

Users, for particular types of queries, are not interested in being told about documents they already know; in these scenarios high novelty scores for the answer set are desirable.

There are many variations on this theme in the literature.

In attempting to capture more fully the user experience researchers have also introduced measures such as recall effort, relative recall, expected search length and many others.

For particular domains with information retrieval, other measures have also been proposed and used; we will cover some more in later classes.

7 Test Collections

For the measures above (e.g. precision and recall), we need to know which documents are considered relevant for which queries. These human judgements are defined as our gold standard and we compare the performance of the system. To this end, there has been a long tradition (starting with the Cranfield test collection) of developing test collections for information retrieval experimentation.

These collections typically comprise a set of documents, a set of queries and a list of relevant documents for each each query. There is a large expense in obtaining these relevance judgements.

Many specialised tasks have been catered for in the literature. These include ad-hoc retrieval, filtering, cross-lingual IR, query-answering among many others. Given the huge diversity of collections, we will forego a detailed discussion in this course. I include some resources:

1. A collection of some test-collections available at:
   `http://ir.dcs.gla.ac.uk/resources/test_collections/`
2. A 2010 review of test collections by Mark Sanderson [1]. This is available from his own website for free access:
   (http://marksanderson.org/publications/my_papers/FnTIR.pdf).

   Note this is a large review but useful if you wish to look up any aspects of information retrieval test collections.

References

[1] Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.