IR Evaluation

April 26, 2015

1 IR Ranking, Search Engine Output

1.1 comparing search engines

Web search engines have their ancestors in the information retrieval (IR) systems developed during the last fifty years. IR methods include (among others) the Boolean search methods, the vector space methods, the probabilistic methods, and the clustering methods [BelCroft87]. All these methods aim at finding the relevant documents for a given query.

One of the primary distinctions made in the evaluation of search engines is between effectiveness and efficiency.Effectiveness, loosely speaking, measures the ability of the search engine to find the right information, and efficiency measures how quickly this is done. For a given query, and a specific definition of relevance, we can more precisely define effectiveness as a measure of how well the ranking produced by the search engine corresponds to a ranking based on user relevance judgments. Efficiency is defined in terms of the time and space requirements for the algorithm that produces the ranking.Carrying out this type of holistic evaluation of effectiveness and efficiency, while important, is very difficult because of the many factors that must be controlled. For this reason, evaluation is more typically done in tightly defined experimental settings and this is the type of evaluation we focus on here.

To measure ad hoc information retrieval effectiveness in the standard way, we need a test collection consisting of three things:

1. A document collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standardly a binary assessment of either relevant or non-relevant for each query-document pair.

Given these ingredients, how is system effectiveness measured? The two most frequent and basic measures for information retrieval effectiveness are precision(the number of relevant retrieved documents divided by the number of retrieved documents) and recall(the number of relevant retrieved documents divided by the number of relevant documents). One main use is in the TREC (Text retrieval conference, http://trec.nist.gov), where many research groups get their system tested against a common database of documents.

2 Set measures

There are several matrices that are used to measure the effectiveness of the IR system. The matrices are True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN).

First, Let us take a look at precision and recall in more detail. As an example, in an information retrieval scenario, the instances are documents and the task is to return a set of relevant documents given a search term; or equivalently, to assign each document to one of two categories, "relevant" and "not relevant". In this case, the "relevant" documents are simply those that belong to the "relevant" category. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. We have the formula for precision and recall as follows:

$$Precision = \frac{num(relevant\ items\ retrieved)}{num(retrieved\ items)} = P(relevant|retrieved) \tag{1}$$

$$Recall = \frac{num((relevant\ items\ retrieved))}{num((relevant\ items))} = P(retrieved|relevant) \tag{2}$$

These notions can be made clear by examining the *confusion matrix.*Given a ranking of documents, we can create a *confusion matrix* that counts the correct and incorrect answers of each type.

	Relevant	Non-Relevant
Retrived	TP	FP
Non Retrived	FN	TN

Table 1: Confusion Matrix

- True Positives(TP) are relevant documents in the ranking
- False Positives(FP) are non-relevant documents in the ranking
- True Negatives(TN) are non-relevant documents missing from the ranking
- False Negatives(FN) are relevant documents missing from the ranking

Now, we can express precison and recall in terms of confusion matrices terms:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

An obvious alternative that may occur to the reader is to judge an information retrieval system by its *accuracy*, that is, the fraction of its classification that are correct. In terms of the contingency table above,

$$accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{5}$$

This seems plausible, since there are two actual classes, relevant and non-relevant, and an information retrieval system can be thought of as a two-class classifier which attempts to label them as such (it retrieves the subset of documents which it believes to be relevant). This is precisely the effectiveness measure often used for evaluating machine learning classification problems. There is a good reason why accuracy is not an appropriate measure for information retrieval problems. In almost all circumstances, the data is extremely skewed: normally over 99.9% of the documents are in the non-relevant category. A system tuned to maximize accuracy can appear to perform well by simply deeming all documents non-relevant to all queries. Even if the system is quite good, trying to label some documents as relevant will almost always lead to a high rate of false positives. However, labeling all documents as non-relevant is completely unsatisfying to an information retrieval system user. Users are always going to want to see some documents, and can be assumed to have a certain tolerance for seeing some false positives providing that they get some useful information. The measures of precision and recall concentrate the evaluation on the return of true positives, asking what percentage of the relevant documents have been found and how many false positives have also been returned.

A single measure that trades off precision versus recall is the *F measure*,which is the weighted harmonic mean of precision and recall:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \ \ where \ \ \beta^2 = \frac{1-\alpha}{\alpha} \tag{6}$$

where $\alpha \in [0,1]$ and thus $\beta^2 \in [0,\infty]$.The default *balanced F measure* equally weights precision and recall, which means making $\alpha = \frac{1}{2}$or $\beta = 1$. It is commonly written as F_1, which is a short of $F_\beta = 1$. When using $\beta = 1$, the formula on the right simplifies to :

$$F_{\beta=1} = \frac{2PR}{P+R} \tag{7}$$

Values of $\beta < 1$ emphasize precision, while values of $\beta > 1$ emphasize recall. For example, a value of $\beta = 3$ or $\beta = 5$ might be used if recall is to be emphasized. Recall,precision, and the F measure are inherently measures between 0 and 1, but they are also very commonly written as percentages, on a scale between 0 and 100.

3 Ranking Measures

Precision, recall, and the F measure are set-based measures. They are computed using unordered sets of documents. We need to extend these measures (or to define new measures) if we are to evaluate the ranked retrieval results that are now standard with search engines. In a ranked retrieval context, appropriate sets of retrieved documents are naturally given by the top k retrieved documents. For each such set, precision and recall values can be plotted to give a precision-recall curve as shown in figure 1.

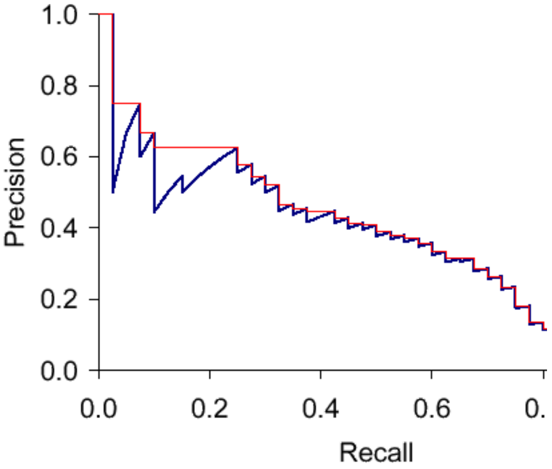

Figure 1: Precision-Recall Curve

The above measures factor in precision at all recall levels. For many prominent applications, particularly web search, this may not be germane to users. What matters is rather how many good results there are on the first page or the first three pages. This leads to measuring precision at fixed low levels of retrieved results, such as 10 or 30 documents. This is referred to as *"Precision at k"*, for example "Precision at 10". It has the advantage of not requiring any estimate of the size of the set of relevant documents but the disadvantage that it is the least stable of the commonly used evaluation measures and that it does not average well, since the total number of relevant documents for a query has a strong influence on precision at k.

Another way to model user behavior is based on the probability that document i is the last document read. This gives an interpretation for *Average Precision*: the expected relevance gained from the user choosing a relevant document i uniformly at random, and reading all documents from 1 to i. Imagine that exactly one of the relevant documents will satisfy the user, but we don't know which one.

$$L_M(i) := \frac{P_M(i) - P_M(i+1)}{P_M(1)} \tag{8}$$

An alternative, which alleviates this problem, is *R-precision.* It requires having a set of known relevant documents Rel, from which we calculate the precision of the top Rel documents returned. (The set Rel may be incomplete, such as when Rel is formed by creating relevance judgments for the pooled top k results of particular systems in a set of experiments.) R-precision adjusts for the size of the set of relevant documents: A perfect system could score 1 on this metric for each query, whereas, even a perfect system could only achieve a precision at 20 of 0.4 if there were only 8 documents in the collection relevant to an information need. If there are |Rel| relevant documents for a query, we examine the top |Rel| results of a system, and find that r are relevant, then by definition, not only is the precision (and hence R-precision) r/|Rel|, but the recall of this result set is also r/|Rel|.

The *Reciprocal Rank* of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank (MRR) is the average of the reciprocal ranks of results for a sample of queries Q:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{9}$$

For example, suppose we have the following three sample queries for a system that tries to translate English words to their plurals. In each case, the system makes three guesses, with the first one being the one it thinks is most likely correct:

Query	Results	Correct response	Rank	Reciprocal Rank
cat	catten,cati,**cats**	cats	3	1/3
torus	torii,**tori**,toruses	tori	2	1/2
virus	**viruses**,virii,viri	viruses	1	1

Given those three samples, we could calculate the mean reciprocal rank as $(1/3 + 1/2 + 1)/3 = 11/18$ or about 0.61.

3.1 Receiver Operating Characteristics curve (ROC)

Another concept sometimes used in evaluation is an ROC curve. An ROC curve plots the true positive rate or sensitivity against the false positive rate or (1 - specificity). Here, sensitivity is just another term for recall. The false positive rate is given by $fp/(fp + tn)$. Figure 2 shows the ROC curve corresponding to the precision-recall curve in Figure 1. An ROC curve always goes from bottom left to the top right of the graph. For a good system, the graph climbs steeply on the left side. Precison-recall curves are sometimes loosely refered to as ROC curve. This is understand, but not accurate.

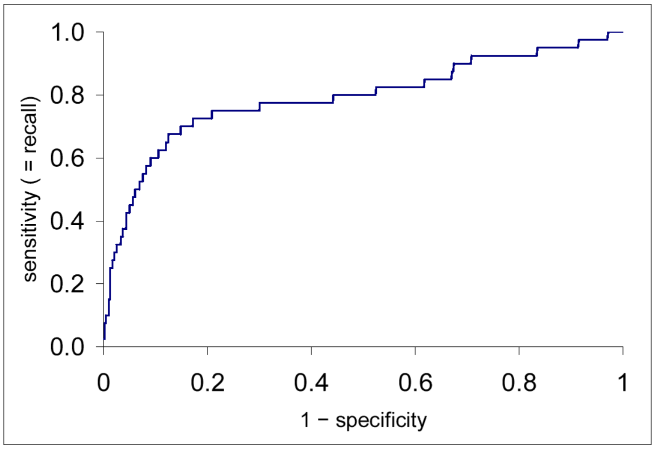

Figure 2: ROC Curve

3.2 nDCG and DCG

A final approach that has seen increasing adoption, especially when employed with machine learning approaches to ranking is measure of cumulative gain and in particular *normalized discounted cumulative gain* (NDCG) . NDCG is designed for situations of non-binary notions of relevance . Like precision at k, it is evaluated over some number k of top search results. For a set of queries Q, let R(j,d) be the relevance scores assessors gave to document d for query j. Then,

$$NDCG(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{log_2(1+m)} \tag{10}$$

where Z_{kj} is a normalization factor calculated to make it so that a perfect rankings NDCG at k for query j is 1. For queries for which $k' < k$ documents are retrieved, the last summation is done up to k'.

Discounted cumulative gain (DCG) is a measure of ranking quality. In information retrieval, it is often used to measure effectiveness of web search engine algorithms or related applications. Using a graded relevance scale of documents in a search engine result set, DCG measures the usefulness, or gain, of a document based on its position in the result list. The gain is accumulated from the top of the result list to the bottom with the gain of each result discounted at lower ranks.

Two assumptions are made in using DCG and its related measures.

1. Highly relevant documents are more useful when appearing earlier in a search engine result list (have higher ranks)

2. Highly relevant documents are more useful than marginally relevant documents, which are in turn more useful than irrelevant documents.

The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result. The discounted CG accumulated at a particular rank position p is defined as:

$$DCG_p = rel_1 + \sum_{i=2}^{|p|} \frac{rel_i}{log_2(i)} \tag{11}$$

where rel_i is the graded relevance of the result at position i.

Let us consider an example for calculating DCG:

Presented with a list of documents in response to a search query, an experiment participant is asked to judge the relevance of each document to the query. Each document is to be judged on a scale of 0-3 with 0 meaning irrelevant, 3 meaning completely relevant, and 1 and 2 meaning "somewhere in between". For the documents ordered by the ranking algorithm as $D_1, D_2, D_3, D_4, D_5, D_6$

the user provides the following relevance scores: 3, 2, 3, 0, 1, 2

That is: document 1 has a relevance of 3, document 2 has a relevance of 2, etc. DCG is used to emphasize highly relevant documents appearing early in the result list. Using the logarithmic scale for reduction, the DCG for each result in order is:

i	rel_i	$log_2 i$	$\frac{rel_i}{log_2 i}$
1	3	0	NA
2	2	1	2
3	3	1.585	1.892
4	0	2.0	0
5	1	2.322	0.431
6	2	2.584	0.774

4 Test Collections

* why we ned them
  * how do we create them
  * QREL files
  * utility of datasets

5 Significance tests

* why we need them
  * popular tests

6 Manual Assessment

* create your own QREL
  * assessment disagreemnts, fatigue
  * experts vs users vs random people

6.1 Crowdsourcing

* cost vs benefit
  * noise
  * quality assurance

7 User Studies

* users vs metrics
  * selecting users
  * IRB
  * types of studies
  * types of measurements