IR Evaluation

April 5, 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:
* accuracy
* precision
* recall
* F1

3 Ranking Measures

* precision, recall @k
* relevant ranks
* Average Precision
* R-precision
* Reciprocal rank

3.1 ROC and Precision-recall curves

3.2 nDCG

* gains - transform grade in usefulness/benefit. What do grades mean? Essentially a benefit model
* discounts - transforms ranks into utility. How much gains still matter as we go down the list? Essentially a user model
* DCG = dot product between gains and discounts
* nDCG = DCG normalized

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