# IR Evaluation

April 3, 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.

* We dont focus on these, but they matter in practice:<br>
* Measure: speed<br>
* Measure: user interface

* we focus on raking/retrieval performance, by analogy to accuracy for Machine Learning<br>
* Document relevant grades

## 2 Set measures

* Confusion matrices: TP, FP, TN, FN
* 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