

# IR Evaluation

March 30, 2015

## 1 IR Ranking, Search Engine Output

- \* comparing search engines
  - \* We don't focus on these, but they matter in practice:
- \* Measure: speed
- \* Measure: user interface
  - \* we focus on ranking/retrieval performance, by analogy to accuracy for Machine Learning
  - \* 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 need 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 disagreements, 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