Comparing the Effectiveness of Different Scoring Functions for Web Search

Marc Najork

Microsoft Research Silicon Valley

14 February 2007

Joint work with Mike Taylor and Hugo Zaragoza

The ranking problem in Information Retrieval

- User issues a query
- IR system (web search engine) consults index to produce result set (“filter set”)
- Problem: Should return results such that most relevant results appear first.
- What determines relevance?
  - Ultimately depends on user’s intent.

IR Performance Measures

- Would like to quantify how closely a ranking algorithm approximates optimal ranking.
- Problem 1: What is optimal?
  - Ground truth established by assembling test set of queries & results labeled by human judges.
  - Tricky issues: How to collect queries? How to label results? How to decide what to label?
- Problem 2: How to measure distance from optimal ranking?
  - Standard distance metrics (Kendall's tau, Spearman footrule) don't correlate to user's satisfaction.

IR Performance Measures

- Issue of distance metrics ("performance measures") has been studied for 40 years
- Good measures should be "rank-sensitive" – give more credit for relevant results on top
- In this talk, we'll use three measures:
  - Mean Reciprocal Rank
  - Mean Average Precision
  - Normalized Discounted Cumulative Gain
- Notion of "document cut-off value"

(Ancient) Measure: Precision

- Given a rank-ordered vector *V* of results $\langle v_1, ..., v_n \rangle$ to query *q*, let $rel(v_i)$ be 1 iff v_i is relevant to *q* and 0 otherwise. The precision of *V* at document cut-off value *k* is the number of relevant documents in the top *k* results:

$$P@k(V) = \frac{1}{k}\sum_{i=1}^{k} rel(v_i)$$

Measure 1: Mean Average Precision (MAP)

- Given a rank-ordered vector V of results $\langle v_1, ..., v_n \rangle$ to query q, the average precision of V at document cut-off value k is the mean of the precisions at every relevant document (or 0 if there are none):

$$AP@k(V) = \underset{v_i : i \le k \wedge rel(v_i)=1}{avg} P@i(V) = \frac{\sum_{i=1}^{k} P@i(V) rel(v_i)}{\sum_{i=1}^{k} rel(v_i)}$$

The mean average precision of the test set is the mean of the AP's of the queries in the test set.

Measure 2: Mean Reciprocal Rank (MRR)

- Given a rank-ordered vector *V* of results $\langle v_1, \ldots, v_n \rangle$ to query *q*, the reciprocal rank of *V* at document cut-off value *k* is:

$$RR@k(V) = \begin{cases} \frac{1}{i} & \text{if } \exists i < k : rel(v_i) = 1 \wedge \forall j < i : rel(v_j) = 0 \\ 0 & \text{otherwise} \end{cases}$$

The mean reciprocal rank of the test set is the mean of the RR's of the queries in the test set.

Measure 3: Normalized Discounted Cumulative Gain (NDCG)

- Given a rank-ordered vector V of results $\langle v_1, ..., v_n \rangle$ to query q, let $label(v_i)$ be the judgment of v_i (0=worst, 5=best). The discounted cumulative gain of V at document cut-off value k is:

$$DCG@k = \sum_{i=1}^{k} \frac{1}{\log_2(1+i)} \left(2^{label(v_i)} - 1\right)$$

The normalized DCG of V is the DCG of V divided by the DCG of the “ideal” (DCG-maximizing) permutation of V (or 1 if the ideal DCG is 0). The NDCG of the test set is the mean of the NDCG’s of the queries in the test set.

Scoring functions

- Ranking algorithms work as follows:
  - Assign a score to each result in the filter set by applying a scoring function to the result
  - Sort the results by decreasing score
- Ideal scoring function can read user's minds (or in the context of evaluation, agrees with the ordering imposed by the judges)
- "Features" to exploit:
  - Words in query & in result documents
  - Structure of result documents
  - Anchor text
  - Hyperlink structure of the web
  - User behavior (e.g. document visitations)
- Scoring functions can be composed – a science upon itself