Introduction to Search and Recommender Systems

Chirag Shah

Evaluation

Reading List:

- Chapter 3 - Baeza-Yates, R. (Ricardo)., Ribeiro-Neto, B. Modern information retrieval. New York: ACM Press.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. Introduction to Information Retrieval Book

Once the documents are indexed, scored and ranked, how do we know whether the retrieval system is good and effective? We need to evaluate the system. Evaluation is the key to build an effective and efficient search and retrieval system.

Why evaluate?

Evaluation is the systematic determination of merit and significance of something using criteria against a set of standards. In general, the goal of the evaluation is the quality of the retrieved results. The specific study designs and effectiveness measures vary by domains and research problems. Depending on the nature of retrieved results, measurement of quality, various techniques of IR evaluation is used.

In general, we evaluate the following:

- There are many retrieval models, algorithms and systems, which one is the best?
- What is the best component/technique for ranking function (e.g., dot-product or cosine), term selection (e.g., stopword removal, stemming), term weighting (e.g., TF, TF-IDF)?
- How far down the ranked list will a user need to look to find some/all relevant documents?

In answering these questions, we mainly concern whether users can make informed decisions based on the retrieved information, whether certain changes in an existing IR system will lead to an improvement in performance.

However, in a broader sense, we evaluate the system not only to measure how many relevant documents it retrieves but also to measure the advantages and disadvantages of the system, its applicability and societal impacts. Studies to gauge these broader impacts are going on but sometimes the results are hard to interpret.

What to evaluate?

We need to evaluate or measure the components that will reflect the ability of the system to satisfy the user.

Cleverdon (1966) listed six main measurable quantities:

1. The coverage of the collection, that is, the extent to which the system includes relevant documents
2. System response time, that is, the average interval between the time the search query is made and the time the search result is given
3. The form of presentation of the output
4. User efforts involved in obtaining answers to their search queries
5. The recall -- the proportion of relevant material actually retrieved in answer to a search query
6. The precision -- the proportion of retrieved material that is actually relevant.

In traditional IR, it is assumed that precision and recall are sufficient for the measurement of effectiveness. However, since the advent of advanced IR technology, and the increased amount of information, there has been much debate in the recent past as to whether precision and recall are in fact the appropriate quantities to use as measures of effectiveness. A popular alternative has been recall and fall-out (the proportion of non-relevant documents retrieved).

How to evaluate?

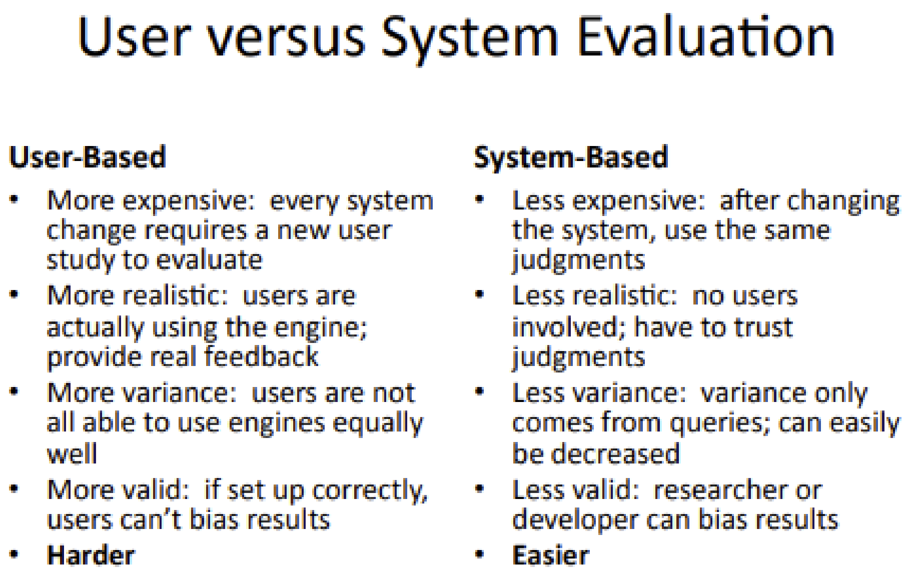
There are two types of evaluation (Kelly, 2009):

1. System-based:
    1. An evaluation based on test-collections (a set of documents, queries and relevance judgments) with no human interaction: Typically, expert judges measure the relevance of documents and then use the judgments to calculate the effectiveness measures. In IR, the test-collection approach can be used to measure the quality and efficiency of retrieved results using various metrics such as Recall, Precision, Mean average precision (MAP), Discounted cumulative gain (DCG), and Normalized DCG.
2. User studies:
    1. User-centered laboratory-based evaluations: Researchers may evaluate the usability and users' satisfaction under controlled conditions in a laboratory setting where they can bring users to interact with systems and measure their responses.
    2. User-centered operational evaluations in natural settings: or in a real-world setting.

Evaluations are generally comparative:

– System A vs. B

– But could be absolute: response time < 1s



User versus System Evaluation

User-Based

- More expensive: every system change requires a new user study to evaluate
- More realistic: users are actually using the engine; provide real feedback
- More variance: users are not all able to use engines equally well
- More valid: if set up correctly, users can't bias results
- **Harder**

System-Based

- Less expensive: after changing the system, use the same judgments
- Less realistic: no users involved; have to trust judgments
- Less variance: variance only comes from queries; can easily be decreased
- Less valid: researcher or developer can bias results
- **Easier**

Image credit: Ben Carteret

Other Methods

Online testing: Test using live traffic on a search system
Query Logs: Users' search history, click-through data

Relevance Judgment

Judgments are how we quantify the quality of search results. Judgments can come from human experts, user interactions or from quantitative measures, but regardless of the source, they measure how relevant a result item is to a given search query.

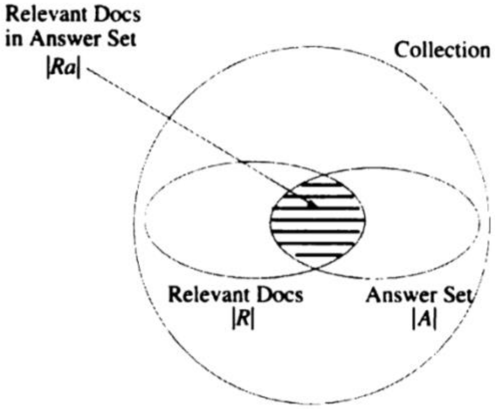
There are two major types of scale for relevance judgment: binary (relevant/not-relevant) and graded (degrees of relevance).

Metrics for Binary Relevance Evaluation

Precision and Recall

irrelevant	retrieved & irrelevant	not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant
	retrieved	not retrieved

Let R be a set of relevant documents for a reference collection based on an information request. Assume the IR system we want to evaluate is used for this test collection and generates an answer set A. Let R_a be the set of documents in the intersection of set A and R. We can then evaluate the performance by recall and precision scores.



Recall is the proportion of retrieved relevant documents in the relevance document set R.

$$Recall = \frac{|Ra|}{|R|}$$

Precision is the proportion of retrieved relevant documents in the answer set A.

$$Precision = \frac{|Ra|}{|A|}$$

F-score incorporates recall and precision by taking their harmonic mean.

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

Precision and recall can be plotted into a graph for a more intuitive understanding: the k+1 th document is nonrelevant when the recall is the same for the first k documents and precision drops; and the document is relevant when both recall and precision increase.

E Measure (parameterized F Measure)
A variant of F measure that allows weighting

$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R} + \frac{1}{P}}$$

emphasis on precision over recall:

• Value of β controls trade-off:

– β = 1: Equally weight precision and recall (E=F).

– $\beta > 1$: Weight recall more.

– $\beta < 1$: Weight precision more.

Average Precision (AP)

For a ranked sequence of relevant documents, we can use AP to evaluate whether this sequence is ranked higher or not.

$$\text{AveP} = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{\text{number of relevant documents}}$$

where k is the rank in the sequence, P(k) is the precision value at k, and rel(k) is an indicator function which is 1 when the kth document is relevant.

Mean Average Precision (MAP)

MAP is commonly used among the TREC community to evaluate the performance of an IR system with multiple information requests. It basically just averages the AP for each query with the equation

$$\text{MAP} = \frac{\sum_{q=1}^{Q}\text{AveP(q)}}{Q}$$

where Q is the total number of queries.

Metrics for Non-binary relevance

Why?

Documents are rarely entirely relevant or non-relevant to a query

Many sources of graded relevance judgments

– Relevance judgments on a 5-point scale

– Multiple judges

– Click distribution and deviation from expected levels (but click-through != relevance judgments)

Cumulative Gain

With graded relevance judgments, we can compute the gain at each rank.

Cumulative Gain at rank n:

$$CG_n = \sum_{i=1}^{n} rel_i$$

Where rel_i is the graded relevance of the document at position i.

Discounting Based on Position

Users care more about high-ranked documents, so we discount results by $1/\log_2$ (rank)

Discounted Cumulative Gain:

$$DCG_n = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{log_2 i}$$

Normalized Discounted Cumulative Gain (nDCG)

To compare DCGs, normalize values so that an ideal ranking would have a normalized DCG of 1.0 normalize by DCG of the ideal ranking:

$nDCG_n = DCG_n / IDCG_n$

nDCG ≤ 1 at all ranks
nDCG is comparable across different Queries