Evaluation of IR systems

Introduction

Evaluation of IR systems

functional requirements

standard testing techniques

performance

- response time
- space requirements
- measure by empirical analysis, efficiency of algorithms and data structures for compression, indexing ..

retrieval performance

- How useful is the system? Not really an issue in data retrieval systems where perfect matching is possible (as there exists a correct answer).
- long history of evaluation; IR is a highly empirical discipline

Test Collections

Evaluation of IR systems is usually based on a test reference collection involving human evaluations.

- Test collection usually comprises:
- a collection of documents (D)
- a set of information needs that can be represented as queries
- a list of relevance judgements for each query-document pair

Issues

- Can be very costly to obtain relevance judgements
- Crowd sourcing
- Pooling approaches
- Relevance judgements don't have to be just binary
- Agreement among judges?

Test Collections

Test Collections/Tasks

- Long history of empirical evaluation in IR; many test collections created to tackle various IR problems.
- TREC provides a means to empirically test the performance of systems in different domains.
  - ad-hoc retrieval: Classic IR task of retrieving relevant documents for a query. Different tracks have been proposed including Web track (retrieval on web corpora), Million Query Track (large number of queries).
  - Interactive Track: (users interact with the system for relevance feedback)
  - Contextual Search: multiple queries over time
  - Entity Retrieval: the task is to retrieve entities (people, places, organizations)
  - Spam Filtering: Identifying and filtering out non-relevant or harmful content such as email spam
  - Question Answering (QA): The goal is to retrieve precise answers to user questions rather than returning entire documents.

Test Collections/Tasks

- Cross language retrieval: The goal is to retrieve relevant documents in a different language from the query. Requires machine translation.
- Other formats: blogs, social platforms, microblog, video
- Conversational IR: retrieving information in conversational IR systems
- Sentiment Retrieval: emphasis is on identifying opinions, sentiments
- Fact checking: misinformation track
- Domain specific retrieval - example genomic data
- Summarisation tasks

- Relevance is assessed for the information need and not the query
- Tuning and optimisation can occur for many IR systems. It is considered good practice to tune on one collection and then test on another.

Interaction with the system may be:

- one-off query
- interactive session

For the former, “quality" of the returned set is the important metric.

For interactive systems, other issues have to be considered—duration of session, user-effort required etc. These issues make evaluation of interactive sessions more difficult.

Precision and Recall

Evaluation of Unranked Sets

Unranked Sets

- The most commonly used metrics are: precision and recall
- Given a set D and a query Q:
- Let R be the set of documents relevant to Q. Let A be the set actually returned by the system.
- Precision is defined as $\frac{|R \cap A|}{|A|}$
- Recall is defined as $\frac{|R \cap A|}{|R|}$

Having two separate measures (precision, recall) is useful as different IR systems may have different user requirements. As examples:

- Web search: precision is of importance
- Legal domain, research: recall is of importance

There is a trade-off between the two measures. For example, by returning everything, recall is maximised, but precision will be poor.

Recall is non-decreasing as the number of documents returned increases.
Precision usually decreases as the number of documents returned increases.

	Relevant	Non Relevant
Relevant	True Positive (TP)	False Negative (FN)
Non-Relevant	False Positive (FP)	True Negative (TN)

Table: Confusion Matrix of True/False Positives and Negatives

- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)

Accuracy

The accuracy of a system: the fraction of these classifications that are correct
(tp + tn) / (tp + fp + fn + tn)
Accuracy is a commonly used evaluation measure in machine learning classification work.
Why is this not a very useful evaluation measure in IR?

Many single value measures exist that combine precision and recall into the one value:

- F-measure
- Balanced F-measure

Evaluation of Ranked results

Precision-Recall plots

- Returned documents are usually ranked.
- Typically plot precision against recall.
- In an ideal system, for a recall value of 1, we would have a precision value of 1. i.e., all relevant documents have been returned and no irrelevant documents have been returned.

Example

Given |D| = 20 and |R| = 10 and a ranked list of length 10.
Let the returned ranked list be:

$$\mathbf{d_1}, \mathbf{d_2}, d_3, \mathbf{d_4}, d_5, d_6, \mathbf{d_7}, d_8, d_9, d_{10}$$

where those in bold font are those that are relevant.

- Considering the list as far as the first document: Precision = 1, Recall = 0.1
- As far as the first 2 documents: Precision = 1, Recall 0.2
- As far as the first 3 documents: Precision = 0.67, Recall 0.2

Usually plot for recall values = 10% ... 90%.

Typically calculate precision for these recall values over a set of queries to get a truer measure of a system's performance.

$$P(r) = \frac{1}{N} \sum_{i=1}^{N} P_i(r)$$

Single value measures

1. Evaluate precision when every new relevant document retrieved. Average precision values.
2. Evaluate precision when first relevant document retrieved.
3. R-precision: Calculate precision when the final relevant document has been retrieved.
4. **Precision at k (P@k)**
5. **Mean Average Precision (MAP)**

Precision Histograms

- Used to compare 2 algorithms over a set of queries.
- Calculate the R-Precision (or possibly another single summary statistic) of two systems over all queries.
- The difference between the 2 are plotted for each of the queries.

Precision-Recall

Advantages

- widespread use
- give definable measure
- summarise behaviour of IR system.

Disadvantages

- Not always possible to calculate recall measure effective of queries in batch mode
- Precision and recall graphs can only be generated when we have ranking
- Not necessarily of interest to user.

User-Oriented Measures

- Let D be the document set
- Let R be the set of relevant documents
- Let A be the answer set returned to the users
- Let U be the set of relevant documents previously known to the user

- Let *AU* be the set of returned documents previously known to the user.

$$Coverage = \frac{|AU|}{|U|}$$

Let *New* refer to the set of relevant documents returned to the user that were previously unknown to the user. We can define *novelty as* as:

$$Novelty = \frac{|New|}{|New| + |AU|}$$

Related Issues

- The issues surrounding interactive sessions are much more difficult to assess.
- Much of the work in measuring user satisfaction comes from the field of HCI.
- The usability of these systems is usually measured by monitoring user behavior or via surveys of user's experience.
- Another closely related area is that of information visualisation—how best to represent the retrieved data for a user etc (later lecture)