Evaluation in information retrieval

Stephen Robertson
Microsoft Research Ltd., Cambridge, U.K.
and City University, London, U.K.

Summary

- The traditional IR evaluation experiment
  - up to and including TREC
  - and a range of problems and issues arising
- Interactive retrieval
- Okapi experiments
- TREC tasks: Routing/filtering and HARD

Note

This deck of slides ranges over a variety of topics in information retrieval evaluation – certainly more than I shall be able to cover in a 1.5 hour session.

My talk will therefore use *a selection only* of these slides.

The traditional IR experiment

To start with you need:

- An IR system (or two)
- A collection of documents
- A collection of requests

Then you run your experiment:

- Input the documents
- Put each request to the system
- Collect the output

The traditional IR experiment

Then you need to:

- Evaluate the output, document by document
- Discover (??) the good documents your system has missed
- Analyse the results

What is a document?

Traditionally: a package of information structured by an author

The traditional IR experiment

What is a request?

Traditionally, a description of a topic of interest

More properly, a partial representation of an underlying information need or problem (ASK)

What is a system?

Traditionally, a device which accepts a request and delivers or identifies documents

(Note: “device” may be an organisation, may involve people)

The traditional IR experiment

Possibly bad assumptions about systems:

System is pure input-output device (put in the request, get out the answer set)

- most real searches involve interaction

System is program

- this implies that the user is outside the system – more on this later
- there are certainly other humans involved (e.g. authors, indexers)

The traditional IR experiment

Why do we need a complete system?

Many tests are really about components

But we do not in general know how to evaluate components

What is a good (relevant) document?

Traditionally, one judged (by an expert) to be on the topic

More properly, one judged by the user to be helpful in resolving her/his problem

The traditional IR experiment

Possibly bad assumptions about relevance:

Relevance is binary

- users are often uncomfortable with yes/no relevance

Relevance of a single document can be judged independently of context

- users may respond differently to a document depending (e.g.) on what they have seen before

Topical relevance = utility

- there may be many other factors involved in utility

The traditional IR experiment

More questions about relevance:

- Relevant to what exactly?
- Is it subjective or objective?
- Who makes the judgement?
- When and with what context?
- On the basis of what data?
- Are there different types of relevance?

The traditional IR experiment

Studies of relevance have shown (*inter alia*):

- Even when queries/needs are very carefully defined, judges disagree
- Mostly, these differences are at the edges
- Mostly, systems show the same relative performance with different sets of judgements
- Multi-level judgements may reveal greater differences between systems

Measurement of performance

Assuming binary relevance and an input-output system, the function of the system is:

- To retrieve relevant documents
- Not to retrieve non-relevant documents

Potentially, for any request there may be any number of relevant documents in the collection

Measurement of performance

Measure for (1):

$$Recall = \frac{\text{No. of relevant docs retrieved}}{\text{Total relevant in the collection}}$$

Measure for (2):

$$Precision = \frac{\text{No. of relevant docs retrieved}}{\text{Total retrieved}}$$

As defined, these relate to set output only

Measurement of performance

Ranked output:

- Plot recall against precision
  - Precision/recall at different score thresholds
  - Precision at different recall levels (10%, 20%…)
  - Precision at different document cutoffs (5, 10, 20…)
- Calculate average precision at different recall levels (various methods)
- Calculate precision=recall at the document cutoff where total retrieved=total relevant

Measurement of performance

Various other measures

Various problems (interpolation/extrapolation; averaging over requests)

trec_eval: program by Chris Buckley used for TREC (more on TREC later)

Measures like recall and precision are somewhat crude as diagnostic tools

Design of IR experiments

Traditionally, run different systems on same set of requests and documents (and relevance judgements)

Good for comparisons of mechanisms embedded within systems

Wonderful for combinatorial experiments with system variables

Not so good for many user experiments

Portable test collections

Collections of documents, requests and relevance judgements are valuable tools

(saves you having to make your own!)

Several such collections exist now

The most extensive are those generated for TREC

TREC
The Text REtrieval Conference

Competition/collaboration between IR research groups worldwide

Run by NIST, just outside Washington DC

Common tasks, common test materials, common measures, common evaluation procedures

Now various similar exercises (CLEF, NCTIR etc.)

Some evaluation issues

Powerful tradition of laboratory experiments…

… very good for addressing some research questions…

… but not so good for others

Some major problem areas: users, interaction and task context

Need to balance requirement for laboratory controls with realism and external validity

Some user issues

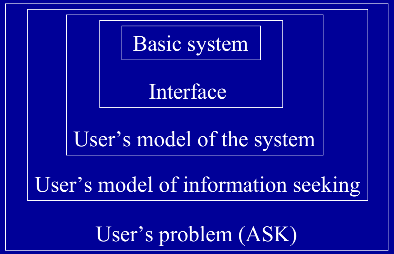
- Interaction
  - Users interact with systems (within sessions and between sessions).
- Relevance
  - Stated requests are not the same as information needs;
  - Relevance should be judged in relation to needs not requests.

Some user issues

- The cognitive view
  - An information need arises from an anomalous state of knowledge (ASK);
  - The process of resolving an ASK is a cognitive process on the part of the user;
  - Information seeking is part of that process;
  - Users’ models of information seeking are strongly influenced by systems.

Some user issues

So: what is the system and where is the user?


Basic system
Interface
User’s model of the system
User’s model of information seeking
User’s problem (ASK)

Some user issues

Adapting laboratory methods to user-centred research questions is hard!

Okapi experiments

(City University 1989—98)

Experimental environment

Mechanism	Interface	User
Functionality	Interaction	Behaviour
Functions to support behaviour	User perception of functions related to task	User perception of task
←--------------------- Evaluation --------------------→		

Okapi systems

Design principles:

- Natural language queries
- Stemming
- Weighting and ranking based on probabilistic model
- Relevance feedback with query expansion

Okapi systems

Versions:

- Character-based interactive system (VT100 system)
- Basic Search System (retrieval engine - supports weighting functions)
  - Boolean and proximity searches, passage retrieval
- Query layer (supports development and maintenance of query, including relevance assessments)
- Various interfaces:
  - a casual user GUI
  - an expert-user interface
- Scripts for running test collection queries

Some results

… from experiments and studies on the Okapi system over several years.

- Careful specification of the weighting and ranking algorithms is critical…
- … the Okapi BM25 algorithm, devised for TRECs 2 and 3, has been very successful.
- Relevance feedback can be a very powerful device.
- In a live-use context, relevance feedback is used moderately frequently…
- … and to reasonable effect.

Some results

- Users commonly repeat searches, either with minor variations or identically.
- They would like to use relevance judgements experimentally/constructively.
- But giving the user more control is not always effective.

Some conflicts

- In a lab test, we try to control variables, i.e. separate the different factors...
  - ...but in interactive searching, the user has access to a range of interactive mechanisms.
- In a lab test, we try to keep user outside the system...
  - ...but in interactive searching, the user/searcher is inside (part of) the system
- In a lab test, we can repeat an experiment, with variations, any number of times...
  - ...but in interactive searching, repetition is difficult and expensive and unlikely to produce identical results.

Routing/filtering experiments at TREC

Basic TREC methods

- Accumulating collections of documents
- Accumulating collections of requests or 'topics'
- Relevance judgements on pooled output from participants, made by the 'users'
- Old topics/documents may have relevance judgements from previous rounds
- Variety of tasks and evaluation measures

Routing/filtering experiments at TREC

The task

- Incoming stream of documents
- Persistent user profile
- Task: send appropriate incoming documents to the user
- Learn from user relevance feedback
- Simulation is not perfect

Routing/filtering experiments at TREC

Batch routing:

- Take a fixed time point, with a 'history' and a 'future'
- Optimise query in relation to history
- Evaluate against future

  in particular, evaluate by ranking the test set
- Results: excellent performance, but some danger of overfitting

Routing/filtering experiments at TREC

Adaptive filtering:

- Start from scratch
  - text query
  - possibly one or two examples of relevant documents
- Binary decision by system
- Feedback only on those items 'sent' to the user
- For scoring systems, thresholding is critical
- Evaluation measures are more difficult

Some results

- For routing (substantial training set, evaluation by ranking of test set), iterative query optimisation is very good indeed
- Threshold setting and adaptation is critical to filtering
- Full adaptive filtering is computationally heavy

The TREC HARD Track

The task: improve performance by making use of:

- Background information about the user and their need;
- Information from one limited interaction with the user

  (System has one chance to ask the user questions – may be more than one question, but only one screenful, and only limited time)

Conclusions

- There is a well-established tradition of laboratory evaluation in IR, including methods and measures
- This tradition is extremely useful, but also has extreme limitations
- If you want to evaluate your system, think very carefully!