Richer theories, richer experiments

Stephen Robertson

Microsoft Research Cambridge
and City University

`ser@microsoft.com`

Themes of the talk

- Search as a science
- The role of experiment and other empirical data gathering in IR
- The (partial) standoff between the Cranfield tradition and user-oriented work
- The role of theory in IR
  – the relation of theories and models to empirical data
- Abstraction

A caricature

On the one hand we have the Cranfield / TREC tradition of experimental evaluation in IR

- a powerful paradigm for laboratory experimentation, but of limited scope

On the other hand, we have observational studies with real users

- realistic but of limited scale

[please do not take this dichotomy too literally!]

Experiment in IR

The Cranfield method was initially only about "which system is best"

system in this case meaning complete package

- language
- indexing rules and methods
- actual indexing
- searching rules and methods
- actual searching

... etc.

It was not seen as being about theories or models...

Implicit models

Cranfield 1: *effectiveness is a consequence of the general approach*

  e.g. 'Faceted classification', 'Uniterms'

Cranfield 2: *effectiveness is a consequence of the combination of low-level components*

  e.g. whether synonyms / morphological variants / generic & specific terms are conflated

Relevance is just a way of measuring effectiveness

Theory and experiment in IR

'Theories and models in IR' (J Doc, 1977):

Cranfield has given us an *experimental* view of what we are trying to do

- that is, something measurable

We are now developing models which address this issue directly

- this measurement is an explicit component of the models

We have pursued this course ever since...

Some models

- Traditional probabilistic models:
  - explicit primary hidden variable which is relevance
  - prediction (estimation) of this variable
  - strong connection of PRP to IR evaluation metrics
- Logical models
  - relevance embedded in 'd implies q'
- Language models:
  - simple model: relevance embedded in 'd implies q'
  - extension: relevance itself as a language model
- Others
  - divergence from randomness: again embedded notion

Hypothesis testing

Focus of *all* these models is predicting relevance

    (or at least what the model takes to be the basis for relevance)

    – with a view to good IR effectiveness

No other hypotheses/predictions sought

    ... nor other tests made

This is a very limited view of the roles of theory and experiment

The scientific method
(simple-minded outline!)

Choose a range of phenomena

Collect empirical data
- by observation and/or experiment

Formulate hypotheses/models/theories

Derive testable predictions
- about events which may be studied empirically

Conduct further observation/experiment
- designed to test predictions and therefore validate theories

Refine/reject models/theories
- and reiterate

Traditional science

The traditional image of science involves experiments in laboratories
- but actually this is misleading

Some sciences thrive in the laboratory
- e.g. chemistry, small-scale physics

Others have made a transition
- e.g. the biochemical end of biology

Others still are almost completely resistant
- e.g. astrophysics, geology

(not to mention such non-traditional sciences such as economics)

Abstraction

Laboratory experiments involve abstraction
- choice of variables included/excluded
- control on variables
- restrictions on values/ranges of variables

Models and theories also involve abstraction
- choice of variables included/excluded
- choice of phenomena to be explained

[but usually different abstractions, for different reasons]

Actually, a model *is* an abstraction

Abstraction in experiments

Why?

- First, to make them possible

Why else?

- study simple cases
- clarify relationships
- reduce noise
- ensure repeatability
- validate abstract theories

(Abstraction in theories is *one* of the reasons for abstraction in experiments)

Newton's laws

Newton's laws have many uses...
    ... including predicting the motion of planets
    ... as well as pendulums and projectiles
There are many ways to test them
    by deriving experimentally testable hypotheses
... and they suggest other experimental measurements
    e.g. of m, the mass of an object
    or g, the acceleration due to gravity
    or G, the gravitational constant

Abstraction in Newton's laws

Abstraction allows the unification of astronomy and local physics

... and also the separation of *use*, *testing*, and *measurement*

(testable hypotheses do not need to be useful)

Testing Newton's laws

Information retrieval phenomena

- people writing documents
- users needing information
  - to solve some problem or accomplish some task
- these users undertaking information-seeking tasks
- various mechanisms to help them
  - by identifying documents
  - or perhaps by extracting information from documents
- a notion of (degrees of) success or failure
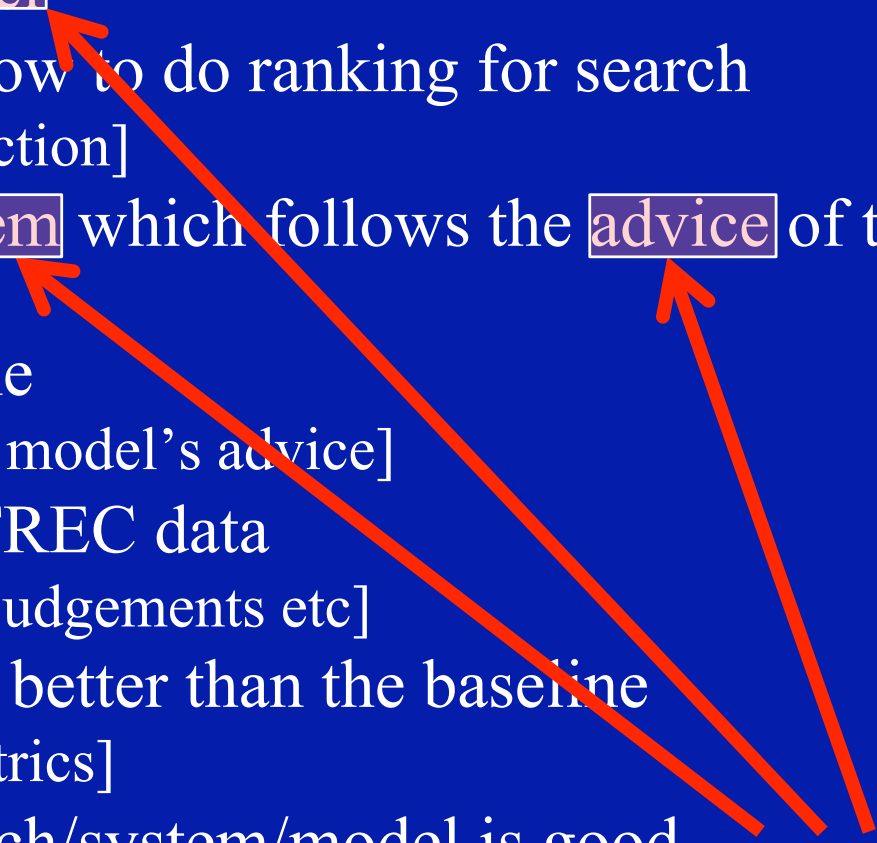
Science and engineering

As IR engineers, we concentrate on

– constructing the mechanisms

– measuring success or failure

This is entirely right and proper, but...

... as IR scientists, maybe we should look a little further

A typical SIGIR paper

1. Construct a model
2. Ask the model how to do ranking for search
   [or other search function]
3. Construct a system which follows the advice of the model
4. Choose a baseline
   [system without the model's advice]
5. Evaluate using TREC data
   [queries, relevance judgements etc]
6. Oh, look, it does better than the baseline
   [on the usual IR metrics]
7. Ergo, the approach/system/model is good

Traditional IR Evaluation

Primarily concerned with:

evaluating *systems* rather than models or theories

... but has *de facto* become the usual way to evaluate models or theories

evaluating in terms of *useful* outcomes

(despite the above)

There seem to be several disconnects here...

User-oriented research

A lot of observational work

    ... but also, increasingly, laboratory experiments

    ... within and outwith the Cranfield/TREC tradition

Emphasis of the models and theories:

    understanding user behaviour

Some points of contact

- The interaction of mechanisms and user behaviour

- Understanding the abstraction that is relevance

- Understanding easily observable behaviours
  - clicks
  - terminations
  - reformulations

- Models and theories which address these issues

Theories and models

So…

We are all interested in improving our understanding

… of both mechanisms and users

One way to better understanding is better models

The purpose of models is to make predictions

But what do we want to predict?

useful applications / to inform us about the model

Predictions in IR

1. What predictions would be *useful*?
   relevance, yes, of course...
   ... but also other things
   - redundancy/novelty/diversity
   - optimal thresholds
   - satisfaction
     - ... and other kinds of quality judgement
   - clicks
   - search termination
   - query modification
     - ... and other aspects of user behaviour
   - satisfactory termination
   - abandonment/unsatisfactory termination
     - ... and other combinations

Predictions in IR

2. What predictions would *inform us about models*?

more difficult: depends on the models

many models insufficiently ambitious

in general, observables/testables

calibrated probabilities of relevance

hard queries

clicks, termination

patterns of click behaviour

query modification

Richer models, richer experiments

Why develop richer models?

- because we want richer understanding of the phenomena
- as well as other useful predictions

Why design richer experiments?

- because we want to believe in our models
- and to enrich them further

A rich theory should have something to say *both* to lab experiments in the Cranfield/TREC tradition, *and* to observational studies

Conclusions

- The Cranfield/TREC tradition has proved *immensely* valuable and useful over the last half-century
- It provides one source (but only one) of empirical data about IR phenomena
- The challenge to theory is to provide a view of the field that addresses these phenomena broadly
- ... and the challenge to empirical work is to inform the development of theory
- Good theory is a much better route to good systems than pure empiricism