Metrics, Statistics, Tests

Stephen Robertson
Microsoft Research Cambridge
stephenerobertson-at-hotmail-dot-co-dot-uk
Why metrics?

Need to measure!
- as a vehicle for understanding
- as a basis for decision-making
- not an end in itself

All observation is measurement
- but we tend to use “measurement” and “metrics” to refer to quantitative observation
  • probably summarised

All experiments involve observation
- of outputs/outcomes, conditions, intermediate effects ...

January 2012 PROMISE Winter School, Zinal
Types of variable

Classification of types

– Nominal
– Ordinal
– Interval
– Ratio

Ordinal property: values are ordered
Interval property: differences are comparable
Ratio property: ratios are comparable
Outline

• The IR evaluation tradition
  – metrics based on relevance
  – work on commercial web search engines
  – questions of statistical significance

• Some current challenges to this tradition
  – the further study of web search
Traditional IR evaluation

System function:
- to separate relevant from non-relevant documents
- to rank relevant above non-relevant documents
- to rank highly relevant above less relevant documents

Purposes of evaluation:
- to decide how well any system performs the above function
- to choose the best systems/components/algorithms

Therefore first metric to talk about is relevance
Assumptions about relevance

Start with a user with an information need

– a query or user prompt to system
– a system outcome consisting of retrieved items
  • (ranking or set)

Assume the user can make judgements

– on each document separately
– on some scale
– about the value of this document
Assumptions about relevance

Note: we do not assume only topical relevance. Judgements may include any number of factors:

- language, comprehensibility, level, authority, currency, even aesthetics

... but we do usually assume that judgements can be made on individual items more-or-less independently:

- which is clearly an over-simplification
- and may cause specific problems with some factors
  - e.g. currency
Relevance scales

Most common in academic evaluation:
  either binary relevance
  or a very short scale
    e.g. highly relevant / partially relevant / not relevant

Most common in commercial system evaluation:
  multi-point scale
    e.g. perfect/excellent/good/fair/bad
    or even two kinds of bad
Why the difference?

Usual assumption of academic evaluation:
start with a specification of an information need,
covering all criteria
like TREC topics

Usual assumption of commercial system evaluation:
start with sampled queries from a log
any query may represent multiple information needs
Metrics based on binary relevance

- Set retrieval / classification:

\[
\text{Recall} = \frac{\text{Relevant retrieved}}{\text{Total relevant}}
\]

\[
\text{Precision} = \frac{\text{Relevant retrieved}}{\text{Total retrieved}}
\]

\[
F_1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

+ others – e.g. utility
Ranked retrieval

Simple user assumption:

– user steps down the ranked list
– and stops at some point
  (but we don’t necessarily know where)
– therefore the system should rank the good stuff as high as possible

The ranking may go on for ever

– but we don’t really care about stuff a long way down

(more about user models later)
Metrics based on binary relevance

- Recall-Precision graph (for ranked retrieval)
Metrics based on binary relevance

• Ranked output

  $\text{MAP} : \ (\text{Mean}) \ \text{Average Precision}$
  $\text{MRR} : \ (\text{Mean}) \ \text{Reciprocal Rank}$
  $P@5 : \ \text{Precision at rank} \ 5$
  $RPrec : \ \text{Precision at rank} \ R \ (R = \text{total rel})$

  etc. (many others)

Note: “Mean” means arithmetic average or mean over queries/topics (even if it is not part of the name, this is what is usually done – more below).
Graded relevance

Some documents are *more relevant* than others
  – but maybe also some documents are relevant *in different ways*

• **Set-based retrieval**
  – can generalise recall and precision
    • calculate separately for each relevance threshold
    • or weights for grades of relevance

• **Ranked retrieval**
  – main metric is (n)DCG
    • gain function for grades of relevance
    • discount for rank
    • maybe cut off at some rank
    • (n) normalise by maximum attainable at rank
Other qualities

Need different metrics for other qualities
  – Known items / navigation
  – Document parts / sections
  – Facets
  – Diversity / intents
  – Novelty
  – …

There are many new metrics suggested and used
Software

• trec_eval – program used for much TREC evaluation
  – input: ranked output of system plus qrels file of relevance judgements
  – output: a whole range of different metrics

• Other programs used by other campaigns
The statistics of traditional evaluation

Inference – about generalisation

Basic ideas
  – sample from some population
  – make measurements on the sample
  – draw inferences concerning the population

Many complications
  – we seldom have real samples!
  – populations may be ill-defined
  – may be simultaneously sampling from multiple populations (more below)
The usual approach

Assume that:

- the object of an experiment is to compare systems for effectiveness
- the critical issue is the number of queries/topics

Now:

- treat topics as if sampled from a population
  ... and therefore each set of per-topic measurements as a sample from a population
- use standard statistical significance tests
Basics of statistical significance

• A null hypothesis
e.g. “there is no difference in effectiveness between systems A and B”

• A test statistic (function of the data)
e.g. $t$ or $\chi^2$

• A distribution of this test statistic under the null hypothesis and appropriate assumptions
e.g. Student’s $t$ distribution

• An unlikely result according to this distribution
e.g. $P(|t| > 3.17) = 1\%$ for 10 d.f.
Paired \( t \)-test

For comparing two systems on (e.g.) \( m = \) average precision:

<table>
<thead>
<tr>
<th>Topic</th>
<th>System A</th>
<th>System B</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( m_A1 )</td>
<td>( m_B1 )</td>
<td>( m_A1 - m_B1 )</td>
</tr>
<tr>
<td>2</td>
<td>( m_A2 )</td>
<td>( m_B2 )</td>
<td>( m_A2 - m_B2 )</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>( MAP_A )</td>
<td>( MAP_B )</td>
<td>( MAP_A - MAP_B )</td>
</tr>
</tbody>
</table>

Null hypothesis: \( MAP_A - MAP_B = 0 \)

Statistic uses variance of \( m_{Ai} - m_{Bi} \)
There are several other tests in common use. The *t*-test is based on particular distributional assumptions
  – not generally satisfied
  – but seems to be fairly robust
Tests such as Wilcoxon make less assumptions
Why a paired test?

We know that variation between queries/topics is huge

- actually much bigger than variation between systems

Pairing helps to reduce the influence of topics

- if your experiment produces unpaired data, need larger samples for similar significance
The problem of hard topics

Some topics are hard
  – meaning that most systems perform poorly on them

Many metrics (including AP) pay little attention to hard topics
  – an improvement for one topic from 0.02 to 0.05 is swamped by another topic going from 0.7 to 0.5

One solution: GMAP (geometric mean average precision)
  – equivalent to taking the mean of the log of AP
    (log 0.05 – log 0.02 is greater than log 0.7 – log 0.5)
Stability and sensitivity of metrics

Work based on significance tests:

- some metrics are more sensitive than others
  - detect significant differences where others do not

Work based on learning to rank:

- some metrics are more stable and reliable than others
  - even if what you really want to optimise is precision@5, it is better to use average precision in learning to rank
The other sampling problem

Basic test data: a set of topics and a collection of documents
  – the documents too might be considered a sample
  – requiring another kind of generalization
There is interaction between the two
  – each per-topic measurement is based on this sample of documents
  – document collection looks different from the point of view of each topic
  – making another source of variation between topics
This issue has scarcely been studied
The other sampling problem

<table>
<thead>
<tr>
<th>Topic</th>
<th>System A</th>
<th>System B</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(m_{A1})</td>
<td>(m_{B1})</td>
<td>(m_{A1} - m_{B1})</td>
</tr>
<tr>
<td>2</td>
<td>(m_{A2})</td>
<td>(m_{B2})</td>
<td>(m_{A2} - m_{B2})</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>(MAP_A)</td>
<td>(MAP_B)</td>
<td>(MAP_A - MAP_B)</td>
</tr>
</tbody>
</table>

We have some sense of the significance of the difference \(MAP_A - MAP_B\), but none of the difference \(m_{A1} - m_{B1}\)
Metrics and user models

User interaction with ranking is quite complex

Even if we do assume “step-down-and-stop”...

- where does the user stop?
- ... and why?

  • satisfaction of need
  • frustration and abandonment
  • frustration and query modification
  • intention to return later
Metrics and user models

In any case, stopping point likely depends on

- interactions with earlier items
- relevance of last item

Much recent work on metrics based on specific user models

- e.g. stochastic models of this interaction
Incomplete judgements

Typical traditional assumption

– we know all the relevant items
– ... or at least most of them
– ... so that we can assume unjudged items are not relevant

But with larger collections this becomes untenable

much current work on alternatives
based on sampling
Beyond Cranfield

Experiments with real users

Think of interaction
  session rather than single query

Think of the wider task
  information rather than documents
  task that engenders information need
  onions and outcomes

Think of clues
  indirect indication rather than direct measurement
Observation in web search

What can we usefully observe about user activity?

- clicks!
- dwell time, return from clicked page
- termination
- reformulation / new search
- eye-movement

(Good / bad not necessarily obvious)
Web search engines: some recent work

• Predicting relevance judgements from logged user data
  (papers by Agichtein and others)
• Comparing rankers by interleaving
  – apply two rankers
  – interleave the results
  – observe clicks
  (papers by Joachims, Radlinski and others)