Metrics, Statistics, Tests

Tetsuya Sakai
Microsoft Research Asia, P. R. China
@tetsuyasakai

February 6, 2013@PROMISE  Winter School 2013 in Bressanone, Italy
Why measure?

• IR researchers’ goal: build systems that satisfy the user’s information needs.
• We cannot ask users all the time, so we need metrics as surrogates of user satisfaction/performance.
• “If you cannot measure it, you cannot improve it.”
http://zapatopi.net/kelvin/quotes/

An interesting read on IR evaluation: [Armstrong+CIKM09]
Improvements that don't add up: ad-hoc retrieval results since 1998
1. Traditional IR metrics
   - Set retrieval metrics
   - Ranked retrieval metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary
Do you recall **recall** and **precision** from Dr. Ian Soboroff’s lecture?

A: Relevant docs  
B: retrieved docs

**E-measure** = \( \frac{|A \cup B| - |A \cap B|}{|A| + |B|} \)

= \( 1 - \frac{1}{0.5*(1/\text{Prec}) + 0.5*(1/\text{Rec})} \)

where \( \text{Prec} = \frac{|A \cap B|}{|B|}, \ \text{Rec} = \frac{|A \cap B|}{|A|} \).

A generalised form

= \( 1 - \frac{1}{\alpha*(1/\text{Prec}) + (1-\alpha)*(1/\text{Rec})} \)

= \( 1 - (\beta^2 + 1)*\text{Prec}*\text{Rec}/(\beta^2 * \text{Prec}+\text{Rec}) \)

where \( \alpha = 1/(\beta^2 + 1) \). See [vanRijsbergen79].
F-measure [Chinchor MUC92]

- Used at the 4th Message Understanding Conference; much more widely used than E
- F-measure = 1 – E-measure
  \[ F \text{-measure} = 1/(\alpha \times (1/\text{Prec}) + (1-\alpha) \times (1/\text{Rec})) \]
  \[ = (\beta^2 + 1) \times \text{Prec} \times \text{Rec} / (\beta^2 \times \text{Prec} + \text{Rec}) \]
  where \( \alpha = 1/(\beta^2 + 1) \).
- F with \( \beta = b \) is often expressed as \( F_b \).
- \( F_1 = 2 \times \text{Prec} \times \text{Rec} / (\text{Prec} + \text{Rec}) \)
  i.e. harmonic mean of Prec and Rec

User attaches \( \beta \) times as much importance to Rec as Prec
\( dE/d\text{Rec} = dE/d\text{Prec} \) when \( \text{Prec}/\text{Rec} = \beta \)
[vanRijsbergen79]
LECTURE OUTLINE

1. Traditional IR metrics
   - Set retrieval metrics
   - Ranked retrieval metrics

2. Advanced IR metrics

3. Agreement and Correlation

4. Significance testing

5. Testing IR metrics

6. Lecture summary
Normalised Discounted Cumulative Gain
[Jarvelin+TOIS02]

- Introduced at SIGIR2000, a variant of Pollack’s sliding ratio [Pollack AD68; Korfhage97]
- Popular “Microsoft” version [Burges+ICML05]:
  \[ nDCG@l= \]
  \[ \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \]

  - \( l \): document cutoff (e.g. 10)
  - \( r \): document rank
  - \( g(r) \): gain value at rank \( r \)
    - e.g. 1 if doc is partially relevant
    - 3 if doc is highly relevant
  - \( g^*(r) \) gain value at rank \( r \) of an ideal ranked list

Original Jarvelin/Kekalainen definition not recommended: a system that returns a relevant document at rank 1 and one that returns a relevant document at rank \( b \) are treated as equally effective, where \( b \) is the logarithm base (patience parameter). \( b \)’s cancel out in the Burges definition.
nDCG: an example

Evaluating a ranked list at $l=5$ for a topic with 1 highly relevant and 2 partially relevant documents

<table>
<thead>
<tr>
<th>System output</th>
<th>Discounted $g(r)$</th>
<th>Ideal list (relevant docs sorted by relevance levels)</th>
<th>Discounted $g^*(r)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonrelevant</td>
<td></td>
<td>Highly rel</td>
<td>3/log2(1+1)</td>
</tr>
<tr>
<td>Highly rel</td>
<td>3/log2(2+1)</td>
<td>Partially rel</td>
<td>1/log2(2+1)</td>
</tr>
<tr>
<td>Nonrelevant</td>
<td></td>
<td>Partially rel</td>
<td>1/log2(3+1)</td>
</tr>
<tr>
<td>Partially rel</td>
<td>1/log2(4+1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonrelevant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partially rel</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cutoff $l=5$

$$nDCG@5 = \frac{2.3235}{4.1309} = 0.5625$$
Average Precision

- Introduced at TREC (1992～), implemented in trec_eval by Buckley
- Like Prec and Rec, cannot handle graded relevance

AP = \frac{1}{R} \sum_r I(r) Prec(r)

where Prec(r) = \frac{rel(r)}{r}

11-point average precision (average over interpolated precision at recall=0, 0.1, ..,1) not recommended for precision oriented tasks, as it lacks the top heaviness of AP. A top heavy metric emphasises the top ranked documents.
User model for AP [Robertson SIGIR08]

- Different users stop scanning the ranked list at different ranks. They only stop at a relevant document.
- The user distribution is uniform across all (R) relevant documents.
- At each stopping point, compute utility (Prec).
- Hence AP is the expected utility for the user population.

Non-uniform stopping distributions have been investigated in [Sakai+EVIA08].

<table>
<thead>
<tr>
<th>Ranked list for a topic with R=5 relevant documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonrel</td>
</tr>
<tr>
<td>20% of users</td>
</tr>
</tbody>
</table>
Q-measure
[Sakai IPM07; Sakai+EVIA08]

• A graded relevance version of AP (see also Graded AP [Robertson+SIGIR10; Sakai+SIGIR11]).

• Same user model as AP, but the utility is computed using the blended ratio $BR(r)$ instead of $Prec(r)$.

$Q = (1/R) \sum_r I(r)BR(r)$

where $BR(r)$

$= \left( \text{rel}(r) + \beta \sum_{k=1}^{r} g(k) \right) / \left( r + \beta \sum_{k=1}^{r} g^*(k) \right)$

$\beta$: patience parameter

(when $\beta=0$, $BR=Prec$, hence $Q=AP$; when $\beta$ is large, $Q$ is tolerant to rel docs retrieved at low ranks)

Combines Precision and normalised cumulative gain (nCG) [Jarvelin+TOIS02]
Value of the first relevant document at rank $r$ according to $BR(r)$ (binary relevance, $R=5$)

For $r \leq R$:
$$BR(r) = \frac{1+\beta}{(r+\beta r)} = \frac{1}{r} = P(r)$$

For $r > R$:
$$BR(r) = \frac{1+\beta}{(r+\beta R)}$$

User patience
$P^+$ [Sakai AIRS06; Sakai WWW12]

- Most IR metrics are for **informational** search intents (user wants as many relevant docs as possible), but $P^+$ is suitable for **navigational** intents (user wants just one very good doc).

- Same as $Q$, except that the user distribution is uniform across rel docs above the **preferred rank** $r_p$, not all rel docs.

$$P^+ = \left(\frac{1}{\text{rel}(r_p)}\right) \sum_{r=1}^{r_p} l(r) \cdot BR(r)$$

**Preferred rank:** rank of the most relevant doc in the list that is closest to the top.

In this example, $r_p=4$. 
Expected Reciprocal Rank
[Chapelle+CIKM09; Chapelle+IRJ11]

Also quite suitable for navigational intents, as it has the **diminishing return** property, i.e. whenever a relevant doc is found, the value of a new relevant doc is discounted.

\[
\text{ERR} = \sum_r dsat(r-1) \Pr(r) \frac{1}{r}
\]

where

\[
dsat(r) = \prod_{k=1}^r (1-\Pr(k))
\]

Pr(r): probability that doc at rank r is relevant

\(\div\) prob that the user is finally satisfied at r

Pr(r) could be set based on gain values
e.g. 1/4 for partially relevant; 3/4 for highly relevant

dsat(r): prob that the user is dissatisfied with docs [1,r]
Rank-Biased Precision [Moffat+TOIS08]

• Moffat and Zobel argue that recall shouldn’t be used: RBP is precision that considers ranks.
• RBP does not range fully between [0,1]
  e.g. When R=10 and p=.95, the RBP for a best possible ranked list is only .4013 [Sakai+IRJ08].
• User model: after examining doc at rank r, will examine next doc with probability p or stop with probability 1-p. Unlike ERR, disregards doc relevance.

\[ \text{RBP} = (1-p) \sum_r p^{r-1} \frac{g(r)}{\text{gain}(H)} \]

\text{gain}(H): \text{gain for the highest relevance level } H \text{ (e.g. 3 for highly relevant)}
Time-Biased Gain [Smucker SIGIR12]

• Instead of document ranks, TBG uses time to reach rank $r$ for discounting the information value.
• TBG has the diminishing return property.

TBG in [Smucker SIGIR12] is binary-relevance-based, with parameters estimated from a user study and a query log:

$$TBG = \sum_r I(r) \times 0.4928 \times \exp(-T(r) \ln 2/224)$$

where $T(r)$ is the estimated time to reach $r$

$$= \sum_{m=1}^{r-1} 4.4 + (0.018 \times \text{lm} + 7.8) \times \text{Pclick}(m)$$

(Pclick=.64 if relevant, .39 otherwise)
### Traditional ranked retrieval metrics summary

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>nDCG</th>
<th>Q</th>
<th>P+</th>
<th>ERR</th>
<th>RBP</th>
<th>TBG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Graded relevance</strong></td>
<td>🙁</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Intent type</strong></td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Nav</td>
<td>Nav</td>
<td>Inf</td>
<td>Inf</td>
</tr>
<tr>
<td><strong>Normalised</strong></td>
<td>YES</td>
<td>YES (nDCG)</td>
<td>YES</td>
<td>YES</td>
<td>NO (ERR)</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td><strong>User model</strong></td>
<td>😊</td>
<td>😞</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Diminishing return</strong></td>
<td>😊</td>
<td>😞</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Document length</strong></td>
<td>😊</td>
<td>😞</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Discriminative power</strong></td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
</tbody>
</table>

*Discriminative power will be explained later*
Normalisation and averaging

• Usually an arithmetic mean over a topic set is used to compare systems e.g. AP->Mean AP (MAP)
• Normalising a metric before averaging implies that every topic is of equal importance, no matter how $R$ varies
• Not normalising implies that every user effort (e.g. finding one relevant document) is of equal importance – but topics with large $R$ will dominate the mean, and different topics will have different upperbounds
• Alternatives: median, geometric mean (equivalent to taking the log of the metric and then averaging) to emphasise the lower end of the metric scale e.g. GMAP [Robertson CIKM06]
Condensed-list metrics
[Sakai SIGIR07; Sakai CIKM08; Sakai+IRJ08]

Modern test collections rely on pooling: we have many unjudged docs, not just judged nonrelevant docs i.e. relevance assessments are incomplete.

Standard evaluation: assume unjudged docs are nonrelevant
System output
Condensed-list evaluation: assume unjudged docs are nonexistent

But condensed-list metrics overestimate systems that did not contribute to the pool, while standard metrics underestimate them [Sakai CIKM08; Sakai+AIRS12a]
“Binary Preference” was probably the first condensed-list metric in the literature but...

• [Buckley+SIGIR04] proposed bpref, which is in fact a variant of condensed-list Average Precision. It lacks the top heaviness of AP and is less robust to incompleteness. See [Sakai SIGIR07; Sakai +IRJ08].

• [Buttcher+SIGIR07] used Ahlgren/Gronqvist RankEff but this metric is in fact a known variant of bpref called bpref\_N (bpref\_allnonrel in trec\_eval). See [Sakai CIKM08].

• Hence bpref and bpref\_N are not recommended.

More on handling incomplete and biased relevance assessments: [Yilmaz+CIKM06] [Aslam+CIKM07] [Carterette SIGIR07] [Webber+SIGIR09]..
[Sakai+IRJ08]
Condensed-list versions of AP, Q, nDCG (AP’, Q’, nDCG’) are relatively robust to incompleteness.

Discriminative power (number of significant differences obtained)

Relevance data downsampling

Condensed-list AP (AP’) is also known as Induced AP [Yilmaz+CIKM06]

Rank correlation with system ranking based on full relevance data

Relevance data downsampling
LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
   - Diversified search metrics
   - Session, summarisation and QA metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary
Diversified search

• Given an ambiguous/underspecified query, produce a single Search Engine Result Page that satisfies different user intents!

• Challenge: balancing relevance and diversity
Diversified search test collections

Traditional IR test collection

Topic

- Relevance assessments

- Relevance assessments

- Relevance assessments

Diversified IR test collection

Topic

- Sub-topic books
  - Relevance assessments

- Sub-topic films
  - Relevance assessments

- Sub-topic character
  - Relevance assessments

- Sub-topic pottermore website
  - Relevance assessments

- Sub-topic office
  - Relevance assessments

- Sub-topic workplace
  - Relevance assessments

- Sub-topic microsoft software
  - Relevance assessments

Topics may be tagged with **ambiguous** (i.e. multi-sense) or **faceted** (i.e. multi-aspect)
Subtopics may be tagged with **informational** or **navigational**
α-nDCG
[Clarke+SIGIR08; Clarke+WSDM11]

- Replaces the gain of nDCG by
  novelty-biased gain

\[ ng(r) = \sum_{i=1}^{m} l_i(r) (1-\alpha)^{\text{reli}(r-1)} \]

- Graded relevance of a doc = number of nuggets covered by doc (Cannot handle graded relevance assessments)
- Discounts gain based on relevant information already seen (diminishing return) e.g. \( \alpha = 0.5 \)
  - If doc at \( r=1 \) is nonrelevant to \( i \), discount factor for \( r=2 \) is \((1-0.5)^0=1\).
  - If doc at \( r=1 \) is relevant to \( i \), it’s \((1-0.5)^1=0.5\).
- But probability that user misses an existing nugget in doc is 0...

Used at the TREC web track diversity task
Intent-Aware metrics

[Agrawal+WSDM09; Chapelle+IRJ11]

M-IA = P(i | q)Mi + P(j | q)Mj

where P(· | q) is the intent probability (popularity)
D-measures

[Sakai+SIGIR11; Sakai+IRJ13]

System output

0
0
2.1
0

Ideal list based on Global Gains

Partially rel:1
Perfect:7
Highly rel:3
Nonrel:0
Partially rel:1
Partially rel:1

Metric M computed based on Global Gains (D-M)

0.7*1+0.3*7=2.8
0.7*3+0.3*0=2.1
0.7*1+0.3*1=1.0

Balancing relevance and diversity:
D#-M = 0.5*intentrecall + 0.5*D-M

Only Intent 1 is covered:
Intent recall (a.k.a. subtopic recall) = 1/2
[Zhai+SIGIR03]

D(#)-nDCG: used at the NTCIR INTENT task
D#-nDCG at work

Example from the NTCIR-10 INTENT-2 task
(to be concluded at the NTCIR-10 conference in June 2013)
DIN-nDCG and P+Q [Sakai WWW12]

Unlike $\alpha$-nDCG, IA metrics and D-measures, considers whether each intent is informational or navigational (do not reward redundant information for nav intents).

System output

\[
\begin{array}{cc}
1 & 1 \\
1 & 0 \\
1 & 0 \\
3 & 0 \\
\end{array}
\]

\[
\begin{array}{cc}
1 & 1 \\
1 & 0 \\
1 & 7 \\
3 & 0 \\
\end{array}
\]

\[
\begin{array}{cc}
1 & 1 \\
1 & 0 \\
1 & 7 \\
3 & 0 \\
\end{array}
\]

DIN-nDCG

D-nDCG

Combine just like IA metrics

P+Q

Preferred rank

Ignore redundant information for navigational intents

Compute nDCG based on the modified Global Gain

Q for i

P+ for j
## Diversity metrics summary

[Sakai+SIGIR11; Sakai WWW12; Sakai+IRJ13]

<table>
<thead>
<tr>
<th></th>
<th>(\alpha)-nDCG</th>
<th>IA metrics</th>
<th>D#</th>
<th>DIN#</th>
<th>P+Q#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graded relevance</td>
<td></td>
<td></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
</tr>
<tr>
<td>Computational complexity</td>
<td></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
</tr>
<tr>
<td>Maximum value is 1</td>
<td></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
</tr>
<tr>
<td>Intent popularity</td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
</tr>
<tr>
<td>[Clarke+WSDM11]</td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
</tr>
<tr>
<td>Informational/navigational</td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
</tr>
<tr>
<td>Discriminative power</td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
</tr>
<tr>
<td>Concordance test</td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
</tr>
</tbody>
</table>

**Note:**

*Discriminative power and concordance test will be explained later*
LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
   - Diversified search metrics
   - Session, summarisation and QA metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary
Session DCG

[Jarvelin+ECIR08; Kanoulas+ SIGIR11]

Extending DCG to multiple ranked lists: **concatenate**
top l docs of m ranked lists in a session and compute

\[ sDCG = \sum_{r=1}^{m^*l} \frac{g(r)}{\log_4(qnum(r)+3)\log_2(r+1)} \]

The original session DCG [Jarvelin+ECIR08] has a problem: documents in earlier lists may be discounted more than those in later lists. [Kanoulas+SIGIR11] also describes an evaluation method for sessions based on multiple possible browsing paths over multiple ranked lists.
ROUGE, POURPRE

• Traditional IR evaluates a (ranked) list of documents, but text summarisation and question answering evaluate textual outputs.
• Instead of documents, nuggets and N-grams are used as the basic unit of evaluation.
• ROUGE [Lin ACL04ws] for summarisation is a recall/F-measure of automatically extracted word N-grams etc., based on gold standard summaries.
• POURPRE [Lin+IRJ06] for QA is an F-measure of answer nuggets, where nugget matching is done automatically using word N-grams.
S-measure, T-measure
[Sakai+CIKM11; Sakai+AIRS12b]

- Evaluating direct textual responses, not ranked lists of web pages
- Evaluate based on information units, not relevant documents
- Present important information first; minimise the user’s reading effort

Unlike nugget precision/recall, S-measure (position-aware weighted recall) says (a)<(b). T-measure (a kind of precision) says (b)>(c). S# combines S and T.
LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary
Measuring agreement

- **Cohen’s kappa**
  
  For two raters who classify N items into C nominal categories

  Observed

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
</tr>
<tr>
<td>#Concordant=60</td>
<td>#Concordant=56</td>
</tr>
</tbody>
</table>

  Chance expected

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
</tr>
</tbody>
</table>

  Cohen’s kappa
  
  Excess of observed concordant
  
  =

  Chance expected nonconcordant
  
  = (60-56)/(100-56)=0.09

  range: [-1, 1]
  
  1: complete agreement
  
  0: completely due to chance

- **Cohen’s weighted kappa**
  
  For two raters who assign items into C ordinal categories e.g. relevance levels 1, 2 and 3 (|C|=3).

  Considers relative concordances as well as absolute ones

- **Fleiss’ kappa**
  
  For three or more raters who classify items into C nominal categories
Pearson’s correlation
(Pearson product moment correlation)

• Degree of linear relationship between two variables \((X, Y)\). Range: \([-1, 1]\)

• \(\text{covariance}(X, Y)\)
  
  \[ \text{stddev}(X) \times \text{stddev}(Y) \]

• For a sample, compute
  
  \[ N \sum XY - \sum X \sum Y \]

  \[ \sqrt{(N \sum X^2 - (\sum X)^2)(N \sum Y^2 - (\sum Y)^2)} \]

Shows that the values of the proposed metric correlate highly with sDCG
Kendall’s τ rank correlation

- Similarity of the orderings of the data by X and Y (not absolute values)
- $\tau = (\text{conc} – \text{disc})/\text{all}$

**all**: all pairs of observations=$N(N-1)/2$

$\text{conc}$: concordant pairs
$(x_i>y_j \text{ and } y_i>y_j \text{ or } x_i<x_j \text{ and } y_i<y_j)$

$\text{disc}$: discordant pairs
$(x_i>x_j \text{ and } y_i<y_j \text{ or } x_i<x_j \text{ and } y_i>y_j)$

Alternatives to Kendall’s τ:
[Yilmaz+SIGIR08; Carterette SIGIR09; Webber+TOIS10]

Range: $[-1, 1]$
LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation

4. Significance testing
   - Standard significance tests
   - Computer-based significance tests

5. Testing IR metrics

6. Lecture summary
Why do significance tests?

• Useful for discussing whether the difference in effectiveness between Systems A and B is substantial or due to chance.
• Null hypothesis $H_0$: all systems are equivalent
• $p$-value: $\text{Pr}(\text{observed or more extreme data}|H_0)$
• Difference is statistically significant if $p$-value is less than the significance level $\alpha$ ($\alpha$ is just a threshold so report $p$-values)

<table>
<thead>
<tr>
<th></th>
<th>Accept $H_0$</th>
<th>Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>correct</td>
<td>Type I error ($\alpha$)</td>
</tr>
<tr>
<td>(equivalent)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>Type II error ($\beta$)</td>
<td>correct</td>
</tr>
<tr>
<td>(different)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Statistical significance does not imply practical significance
• Statistical insignificance does not imply practical insignificance
(Student’s) t-test

- **Paired test**: one topic set, two systems X and Y (typical setting in IR experiments)
- **Observed diffs**: \( z=(z_1,...,z_N)=(x_1-y_1,...,x_N-y_N) \)
- **Assumption**: errors are normally distributed
  (Even if not, central limit theorem says the distribution approaches normal as N grows large)
- **\( H_0 \)**: \( \mu=0 \) (population mean of differences is zero)
- **\( H_1 \)(alternative hypothesis)**: \( \mu \neq 0 \) (**two-tailed**)
- **Under \( H_0 \)**, \( t(z)=\bar{z}/(\bar{\sigma}/\sqrt{N}) \) where \( \bar{\sigma}=\sqrt{\sum_i (z_i-\bar{z})^2/(N-1)} \)
  follows Student’s t distribution with N-1 degrees of freedom
Paired nonparametric tests
(fewer assumptions, less statistical power)

- Wilcoxon signed-rank test
  Assumption: errors come from a continuous distribution symmetric about 0
  - Rank zi’s by magnitude;
  Test statistic \( W = |\sum \text{sign}(zi) \times \text{rank}(zi)| \)

- Sign test
  Assumption: errors come from a continuous distribution
  - Only the sign of zi matters (ordinal scale)
  Test statistic \( |n^+ - n^-| / \sqrt{n^+ + n^-} \) follows standard normal distribution

Friedman test can be used for more than two systems

Remove topics where Zi=0 (Reduce N)
On significance testing in the 20\textsuperscript{th}-century IR literature

• [vanRijsbergen79] “parametric tests are inappropriate because we do not know the form of the underlying distribution. [...] One obvious failure is that the observations are not drawn from normally distributed populations.”

“[...] the sign test [...] can be used conservatively.”

• [Hull SIGIR93] “While the errors may not be normal, the t-test is relatively robust to many violations of normality. Only heavy skewness [...] or large outliers [...] will seriously compromise its validity.”
LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation

4. Significance testing
   - Standard significance tests
   - Computer-based significance tests

5. Testing IR metrics

6. Lecture summary
Why use computational power for significance testing?

• Standard significance tests were developed before the high-performance computer age. They rely on several assumptions (e.g. normality) on the underlying distributions, which often do not hold.

• Instead of making many assumptions, use the observed data and computational power to estimate the distributions!

• “The use of the bootstrap either relieves the analyst from having to do complex mathematical derivations, or in some instances provides an answer where no analytical answer can be obtained.” [Efron+93, p.394]
Bootstrap test for two systems

[Savoy IPM97; Sakai SIGIR06]

\[ z = (z_1, \ldots, z_N) \text{ where } z_i = x_i - y_i; \]
\[ t(z) = \frac{z}{\sigma/\sqrt{N}} \text{ where } \bar{z} \text{ and } \sigma \text{ are mean and standard deviation of } z; \]
\[ w = (z_1 - \bar{z}, \ldots, z_N - \bar{z}); \]
\[ \text{count} = 0; \]
for \( b = 1 \) to \( B \) do {
    \[ w^*_b = \text{bootstrap sample of size } N \]
    obtained by sampling with replacement from \( w; \]
    \[ t(w^*_b) = \frac{\bar{w}^*_b}{\sigma^*_b/\sqrt{N}} \text{ where } \bar{w}^*_b \text{ and } \sigma^*_b \text{ are} \]
    mean and standard deviation of \( w^*_b; \]
    \[ \text{if } |t(w^*_b)| \geq |t(z)| \text{ } \text{count }++; \]
}\]
\[ ASL = \text{count} / B; \]

\[ \text{i.e. p-value: how rare is this observation under } H_0? \]

See [Smucker+CIKM07] for randomisation test for two systems and comparison with classical and bootstrap tests

Histogram of \( t(w^*_b) \) for the difference in Mean Average Precision
Randomised version of Tukey’s Honestly Significantly Different (HSD) test for three or more systems [Carterette TOIS12]

If you have three or more systems but you are using pairwise tests, you may be jumping to wrong conclusions! **Family-wise error rate=\(1-(1-\alpha)^{\text{nsystempairs}}\)**

foreach pair of runs \((r_1, r_2)\) do \(\text{count}(r_1, r_2) = 0;\)

for \(b = 1\) to \(B\) do {
    create matrix \(X^*\) whose row \(t\) is a permutation of row \(t\) of \(X\) for every \(t \in T;\)
    \(\text{max}^* = \max_i \overline{x}_i^*; \text{min}^* = \min_i \overline{x}_i^*\) where \(\overline{x}_i^*\) is the mean of \(i\)-th column vector of \(X^*;\)
    foreach pair of runs \((r_1, r_2)\)
    if\((\text{max}^* - \text{min}^* > |\overline{x}(r_1) - \overline{x}(r_2)|\) where \(\overline{x}(r_i)\) is the mean of the column vector for run \(r_i\) in \(X\) )
    \(\text{count}(r_1, r_2) += ;\)
}

foreach pair of runs \((r_1, r_2)\) do \(\text{ASL}(r_1, r_2) = \text{count}(r_1, r_2)/B;\)
Is significance testing useless? (from outside IR literature)

- [Johnson99] The insignificance of statistical significance testing
  - [...] determining which outcomes of an experiment or survey are more extreme than the observed one, so a P-value can be calculated, requires knowledge of the intentions of the investigator.
  - If the null hypothesis truly is false (as most of those tested really are), then P can be made as small as one wishes, by getting a large enough sample.
  - The famed quality guru W. Edwards Deming (1975) commented that the reason students have problems understanding hypothesis tests is that they may be trying to think.

- [Ioannidis05] Why most published research findings are false
  - [...] most research questions are addressed by many teams, and it is misleading to emphasize the statistically significant findings of any single team. What matters is the totality of the evidence.
  - [...] instead of chasing statistical significance, we should improve our understanding of the range of R values — the pre-study odds — where research efforts operate.
  - Despite a large statistical literature for multiple testing corrections, usually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding.
LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary
Discriminative power
[Sakai SIGIR06; Sakai SIGIR07]

A method for comparing the robustness to topic variance: given a test collection, how many significantly different system pairs can be obtained?

Example from [Sakai+SIGIR11]

20 runs: 20*19/2 = 190 run pairs sorted by p-value

Discriminative power results are consistent with the swap method [Voorhees+SIGIR02] results but the latter needs to split the topic set in half. Discriminative power is now more widely used e.g. [Robertson+SIGIR10; Clarke+WSDM11; Smucker SIGIR12]
Comments on discriminative power

[Sakai WWW12]

• Metrics with low discriminative power are not useful because they can’t give you conclusive results.

• It does not tell you whether the metric is measuring what you want to measure or not.

• Q: If a metric *knows* one list from Google and the other is from Bing, and says Bing is better no matter what the query is, isn’t discriminative power 100% and useless? [Sanderson FnTIR10]

• A: No, that’s cheating. A metric is a function of (a) the system output and (b) the gold standard. It doesn’t know which one is Google!
Side-by-side test

Microsoft’s campaign in 2012: blind comparison of Google’s and Bing’s ranked lists

San Francisco! Bing is better than Google!

Which is better? Left or right?
Predictive power [Sanderson+SIGIR10]

Is a metric “right?” Let’s ask people!

I am N-DCG, human-cyborg relations. RED is obviously better.

BLUE is better
RED is better
RED is better
RED is better
RED is better

• Difficult to apply directly to diversified search metrics (each diversified list is intended for a population of users having different intents)

• Mechanical Turkers are not real users; need screening
Concordance test (a.k.a. intuitiveness test)  
[Sakai WWW12; Sakai+IRJ13]

Is a diversity metric “right?” Let’s ask simpler metrics!

I am $\alpha$-nDCG, human-cyborg relations. RED is obviously better.

I am Precision. I only care about relevance. BLUE is better.

I am Intent recall. I only care about diversity. RED is better.

Agree/disagree
Leave-One-Out Test [Zobel SIGIR98]

Used for testing whether new systems can be evaluated fairly with a pooling-based test collection and an evaluation metric.

Original relevance assessments = Union of contributions from Teams A, B, C and D

“Leave Team A Out” relevance assessments

Remove Team A’s unique contributions

Evaluate Team A using this LOO set. Can this “new” team evaluated fairly?
LECTURE OUTLINE

1. Traditional IR metrics
2. Advanced IR metrics
3. Agreement and Correlation
4. Significance testing
5. Testing IR metrics
6. Lecture summary
Summary: using metrics correctly

• Understand and use the right metrics to evaluate your task.
• Several methods exist for discussing which metrics are “good.”
• Do significance testing with proper baselines.
• But statistical significance does not imply practical significance; statistical insignificance does not imply practical insignificance.
• Use multiple metrics/test collections and look for consistency.

“If you cannot measure it, you cannot improve it.”
Further reading 1/2

- [Carterette TOIS12] Carterette: Multiple testing in statistical systems-based information retrieval experiments, ACM TOIS, 2012.
- [Ioannidis05] Ioannidis: Why most published research findings are false, PLoS Med, 2005.
Further reading 2/2

- [Sakai+EVIA08] Sakai and Robertson: Modelling A User Population for Designing Information Retrieval Metrics, EVIA 2008.g
- [Sakai+AIRS12b] Sakai and Kato: One click one revisited: enhancing evaluation based on information units, AIRS 2012.
- [Zobel SIGIR98] Zobel: How reliable are the results of large-scale information retrieval experiments? SIGIR 1998.