Lecture 16: IR Evaluation (2)
Recap: Top-k Pooling

Purpose

• To select a subset of documents for relevance judgments
• Why only a subset? Because there are too many documents!
• Ideally, the subset should include a lot of relevant results …
  • Such that we can easily assume unjudged results are non-relevant

Top-k Pooling (TREC standard)

• Assume we have multiple different retrieval systems …
  • Not a big issue in the context of TREC evaluation 😊
• Using each system to retrieve top k results
• Use the union set of top k results from different systems as the pool
• Judge all the results in the pool; unjudged results are non-relevant
Zobel (1998)

Can fit the curve using an exponential function

\[ n = C p^g - 1 \]

Pooling efficiency reduces when depth \((k)\) increases: if we go down a ranked list, it is more and more difficult to find new relevant results.

Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In SIGIR 98: 307-314.
Recap: DCG & nDCG

System output
• \([D_1, D_2, D_3, D_4, D_5]\]

\[
DCG@5 = \frac{1}{\log_2 2} + \frac{3}{\log_2 3} + \frac{0}{\log_2 4} + \frac{0}{\log_2 5} + \frac{3}{\log_2 6}
\]

Ideal ranked list
• \([D_2, D_5, D_1, D_6, X]\]

\[
IDCG@5 = \frac{3}{\log_2 2} + \frac{3}{\log_2 3} + \frac{1}{\log_2 4} + \frac{1}{\log_2 5} + \frac{0}{\log_2 6}
\]

\[
nDCG@5 = \frac{DCG@5}{IDCG@5}
\]

Relevance Judgments

<table>
<thead>
<tr>
<th>Document</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>D_1</td>
<td>1</td>
</tr>
<tr>
<td>D_2</td>
<td>2</td>
</tr>
<tr>
<td>D_3</td>
<td>0</td>
</tr>
<tr>
<td>D_4</td>
<td>0</td>
</tr>
<tr>
<td>D_5</td>
<td>2</td>
</tr>
<tr>
<td>D_6</td>
<td>1</td>
</tr>
<tr>
<td>D_7</td>
<td>0</td>
</tr>
</tbody>
</table>
Evaluation Measures: A Summary

Multi-level relevance judgments
• Highly relevant results are more important
• Many measures use an exponential function $2^{\text{relevance}}$

Discounting function
• Penalize the importance of results at lower ranks
• Recent measures all explicitly model how users view search results

Position based model (DCG/nDCG, RBP)
• Discount only depends on the rank of the result

Cascade model (ERR, TBG)
• Discount depends on previous results
• ERR (prev. results’ relevance); TBG (prev. results’ relevance & length)
Evaluation Measures: A Summary

The most popular two measures today

• Mean average precision (despite that it only looks into binary relevance)
• nDCG@$k$ ($k$ usually set to 10, 5, or 20)

Other popular ones

• P@$k$: kind of old … (but still popular outside the IR community)
• RR: quite common even today
• ERR: mainly for web search (TREC made it popular)
• RBP and TBG are less popular
  • But have important theoretical implications
• How to build reliable IR test collections?
  • How many queries, pooling, relevance judgments criteria, etc.
• What are popular search evaluation metrics measuring?
  • P@k, RR, AP, DCG/nDCG, RBP, ERR, TBG, …
• How to draw meaningful conclusions from “scores”?
  • Significance tests
How to draw conclusions?

How to quantify the differences between two systems?

• Voorhees & Buckley (2002): by the differences in MAP

• How much difference is enough?
How to draw conclusions?

A few things to consider

• We compare “population” means based on “sample” means
  • “Population”: MAP of system A and B
  • “Sample”: MAP of A and B on some particular set of queries
  • “Observed sample”: MAP on the queries in the test collection

• The magnitude of the difference
  • 0.49 vs 0.48: seems insignificant and random
  • 0.49 vs. 0.30: seems a “real” difference

• Sample size
  • Law of large numbers: as the sample size increases, sample mean approximates population mean.
  • A large sample often gives a more accurate estimate
Significance Test in IR

Significance Test in IR

• To quantify how likely the observed difference between two systems is by chance

• **Null hypothesis**: the two systems have no difference (in terms of their performance by a measure such as MAP or P@10)

• $p$ value: $P(\text{difference } \geq \text{observed difference} \mid \text{null hypothesis})$

• A greater $p$ value suggests:
  • The observed difference is more likely just by chance.
  • There is less likely a real difference between the two systems.

• Intend to help researchers draw conclusions in a more rigorous way

• Is a common practice in almost all scientific disciplines …

• But sometimes (in fact, in may cases) misused and overused …
Significance Test: A Simple Example

Example: tossing a coin

• We suspect P(head) ≠ 0.5; “null” hypothesis P(head) = 0.5

Conduct experiments to verify the null hypothesis

• Tossed the coin \(n\) times; observed heads \(m\) times

• If the null hypothesis is true, \(m\) should follow Binomial(\(n, 0.5\))
Significance Test: A Simple Example

Conduct Experiments

• Let $n = 40$, the expectation of $m$: $E(m) = 0.5$ $n = 20$
• Examine to which extent the observed $m$ differ from $E(m)$
• $P(\text{difference} \geq \text{observed difference} \mid B(n, 0.5))$

Example: $m = 15$

• $P( m \leq 15 \text{ or } m \geq 25 \mid B(40, 0.5)) = 0.1539$ (two-tail test)
• $P( m \leq 15 \mid B(40, 0.5)) = 0.0769$ (one-tail test; only if you wish to test the coin is less likely to get heads than normal)
Significance Test: A Simple Example

Determine a threshold of the $p$ value to draw conclusions

- At most 0.05; 0.05, 0.01, 0.001 are very common
- No gold standard; depends your research community (reviewers 😊)
- If $p < \text{threshold}$, we reject the null hypothesis and claim the two systems have a significant difference

One-tail or two-tail?

- Depends on your research hypothesis
- Should almost always use two-tail tests in IR
  - Because when you build a new system and compare with a baseline, the new system can either outperform or underperform …
- But using one-tail test is pretty common in papers …
An Example for Comparing IR Systems

• The example is often called a binomial test or sign test
• Example in IR:
  • Purpose: comparing System A and B by P@10
  • Null hypothesis: the two systems are equally good
  • Alternative hypotheses, e.g., two systems are not equally good
  • Run experiments on the n queries
  • If the null hypothesis is correct, after running n queries, the number of queries where A > B and A < B should be roughly the same
  • Get the observed number of queries A > B and A < B from experiments
  • Calculate p value and compare with the threshold, e.g., 0.05
  • Draw conclusions based on the p value and the threshold

• But practically sign test is rarely used in IR
  • Only considers A > B or A < B; ignore the magnitude of difference …
Paired t-test

Purpose

• Compare means of two related samples
• Two systems’ performance on the same set of queries

An example

• Usually for examine “before-after” effect
• Purpose: is a medication helpful for reducing blood pressure?
• Ask a group of $N$ patients to take the medication
• Two related samples:
  • Blood pressure level prior to the medication
  • Blood pressure level after the medication
  • The two samples are for the same group of patients
• Null hypothesis: no difference (the medication does not work)
Paired t-test

Two sample: $X_1$ and $X_2$
- e.g., $X_1$ and $X_2$ can be P@10 of system A and system B on a query
- $H_0$ (null hypothesis): $X_1 - X_2 = 0$
- $H1$ (two-tail): $X_1 - X_2 \neq 0$

Assumption for applying paired t-test
- Each observations are independent
  - e.g., the P@10 of each query is independent of others
  - This is not completely true, but often considered as acceptable
- $X_1 - X_2$ is normally distributed
- The normal distribution assumption can be ignore for a “large” sample, e.g., $N \geq 50$
  - In general, it is okay to ignore the normal distribution hypothesis in IR because we often evaluate using at least 50 queries
Paired t-test

Computing test statistics based on observed $X_1$ and $X_2$

- Let $X_D = X_1 - X_2$
- $\bar{X}_D$: the mean of $X_D$ over the sample; $S_D$: sample variance

$$t = \frac{\bar{X}_D}{\frac{S_D}{\sqrt{N}}}$$

How it works

- Calculate the test statistics $t$ based on the observed samples
- $t$ should follow a $t$-distribution if the null hypothesis is true
- Determine the $p$ value based on the probability of observing a $t$ greater than or lower than the observed value
- Compare $p$ value with the threshold (e.g., 0.05, 0.01, 0.001)
- Draw conclusions based on the $p$ value
Paired t-test

Student's t distribution

• Parameter $\nu$: degree of freedom ($N - 1$ in the case of paired t-test)
• Get the $p$ value based on $t$ and the corresponding t-distribution
Paired t-test: an example

Compare QL and RM3

- Test collection: Robust04 and 249 queries (N = 249)
- Mean P@10: QL: 0.440; RM3: 0.459
- H₀: QL and RM3 yield the same P@10
- H₁: QL and RM3 yield different P@10 (two-tail)
- \( X_D = X_1 - X_2 = 0.019 \)
- Base on the data, \( S_D = 0.115 \)
- Test statistics \( t = \frac{\bar{X}_D}{\frac{S_D}{\sqrt{N}}} = 2.642 \)
- Degree of freedom = \( N - 1 = 248 \)
- \( P( |t| = 2.642 \mid DF = 248 ) = 0.009 < 0.01 \)
  
<table>
<thead>
<tr>
<th>QL3</th>
<th>RM3</th>
<th>Sig. (by two-tail paired t-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.440</td>
<td>0.459</td>
<td>**</td>
</tr>
</tbody>
</table>

- Reject H₀ & adopt H₁
- Usually use * for \( p < 0.05 \), ** for \( p < 0.01 \), and *** for \( p < 0.001 \)
- Reporting exact \( p \) values is often unnecessary
Other tests

Student’s t-test for independent samples

• When: compare system A’s performance on query set X with system B’s performance on another query set Y
• This rarely happens in IR research …
• Normally we compare two systems on the same set of queries
• This test assumes equal variance of the two samples …

Welch t-test

• Do not require the equal variance assumption of the independent sample t-test
• When: when the equal variance assumption is a concern

Wilcoxon signed-rank test

• When: the normal distribution assumption of paired t-test is not true (less likely a concern for IR research because we often have $N \geq 50$)
Significance Test

Misuse
• Do the data assumptions hold? Are you using the wrong test?
• Play the game of $p$ value …
  • Partly because of the community (especially reviewers) …
• e.g., use one-tail when two-tail is appropriate
  • $p = 0.09 \rightarrow p = 0.045$; Yeah! $p < 0.05$
  • You should almost always consider using a two-tail test, unless the two side events is unlikely to happen

Misinterpretation
• $p$ value $\neq$ the likelihood that the null hypothesis is wrong
  • $p$ value: $P(\text{difference} \geq \text{observed difference} | \text{null hypothesis})$
• Statistical significant $\neq$ practical significance
  • A large sample size can easily make it statistically significant
  • But do we really care the 0.001 difference in P@10? (practical significance)
IR Evaluation: Online vs. Offline

Offline evaluation (the Cranfield style)
- Test collection + metric + significance test
- Pros: automatic & repeatable; no additional cost
- Cons: cost of building a test collection; limited correlation with UX

Online evaluation
- Put systems into real user traffic
- Determining the goodness of a system based on user feedback
- Implicit feedback is usually preferred over explicit ones
- Two popular types: interleaved experiment vs. quality prediction
Hassan et al. (2010)

Purpose

• To predict whether or not a search goal is successfully fulfilled.
  • Input: only user behavior sequence; no relevance judgments
• Search goal: defined as an atomic information need resulting a search session involve one or multiple queries.

Data Annotation

• 2,712 sessions with atomic goals from Yahoo! search logs
• Who: external assessors (not the real searchers)
• Criteria: 5 levels (definitely successful, probably successful, unsure, probably unsuccessful, and definitely unsuccessful).
• 5 levels => 0/1 labels (definitely & probably successful => 1)

Hassan et al. (2010)

Characterizing a search session as a sequence of actions

- START & END
- Q: submitting a query
- Click: normal search result (SR), ads (AD), related search (RL), spelling suggestion (SP), shortcut click (SC), other (OTHER)
- Time interval between each action.
- Examples:

```
Goal 1: Q 4s RL 1s SR 53s SR 118s END
Goal 2: Q 3s Q 5s SR 10s AD 44s END
```
Hassan et al. (2010)

A Hidden Markov model (HMM) of action sequence

- Each action only depends on its previous action
- An example HMM estimated from two sessions using MLE

- Estimate HMM for successful and unsuccessful sessions separately.
- Classify using $P(\text{session|success}) / P(\text{session|unsuccessful})$
### Hassan et al. (2010)

**Action transition probability ratio (successful/unsuccessful)**

- $> 1$: transition is more frequent in successful sessions

<table>
<thead>
<tr>
<th>Action following query</th>
<th>Odds-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>2.0</td>
</tr>
<tr>
<td>SR</td>
<td>1.8</td>
</tr>
<tr>
<td>RL</td>
<td>1.2</td>
</tr>
<tr>
<td>SP</td>
<td>0.9</td>
</tr>
<tr>
<td>Q</td>
<td>0.5</td>
</tr>
<tr>
<td>OTH</td>
<td>0.3</td>
</tr>
<tr>
<td>END</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query -&gt; *</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Action leading to end</td>
<td>Odds-ratio</td>
</tr>
<tr>
<td>SR</td>
<td>1.5</td>
</tr>
<tr>
<td>SC</td>
<td>1.2</td>
</tr>
<tr>
<td>OTH</td>
<td>1.0</td>
</tr>
<tr>
<td>RL</td>
<td>0.7</td>
</tr>
<tr>
<td>Q</td>
<td>0.1</td>
</tr>
</tbody>
</table>

* * -> END
Hassan et al. (2010)

Highly probable successful/unsuccessful sessions

<table>
<thead>
<tr>
<th>Highly probable successful paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q SR END</td>
</tr>
<tr>
<td>Q SR SR END</td>
</tr>
<tr>
<td>Q SR SR SR END</td>
</tr>
<tr>
<td>Q SR SR SR SR END</td>
</tr>
<tr>
<td>Q AD END</td>
</tr>
<tr>
<td>Q SC END</td>
</tr>
<tr>
<td>Q SR Q SR SR END</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Highly probable unsuccessful paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q END</td>
</tr>
<tr>
<td>Q Q END</td>
</tr>
<tr>
<td>Q OTH END</td>
</tr>
<tr>
<td>Q SR Q END</td>
</tr>
<tr>
<td>Q Q Q END</td>
</tr>
<tr>
<td>Q RL END</td>
</tr>
<tr>
<td>Q Q SR Q SR Q END</td>
</tr>
</tbody>
</table>
Hassan et al. (2010)

Modeling time using a gamma distribution

- Click time in successful/unsuccessful sessions are different

![Graph showing time distributions]

**Figure 3:** Time distributions of $SR \rightarrow Q$ transitions for successful and unsuccessful search goals.

- Can also compute $P(\text{time|successful}) / P(\text{time|unsuccessful})$
Hassan et al. (2010)

Does it work?

- 1. Online method (Markov Model) is significantly better than using offline evaluation metrics (Relevance and DCG)
- 2. Combining online and offline methods cannot improve much

<table>
<thead>
<tr>
<th>Offline metrics (Cranfield style)</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance (Eqn. 3)</td>
<td>84.2</td>
<td>93.3</td>
<td>88.4</td>
<td>80.2</td>
</tr>
<tr>
<td>DCG</td>
<td>84.7</td>
<td>91.0</td>
<td>87.6</td>
<td>79.1</td>
</tr>
<tr>
<td>Markov Model</td>
<td>89.8</td>
<td>92.3</td>
<td>91.1</td>
<td>85.2</td>
</tr>
<tr>
<td>Markov Model + DCG</td>
<td>88.7</td>
<td>94.1</td>
<td>91.3</td>
<td>85.4</td>
</tr>
</tbody>
</table>

- Here Relevance is just the weighted average relevance score of the top three results

\[
Relevance = \frac{R_{pos1} + \frac{R_{pos2}}{2} + \frac{R_{pos3}}{3}}{1 + \frac{1}{2} + \frac{1}{3}}
\]
Hassan et al. (2010): Summary

Main Findings

• Goal success can be predicted accurately at session level based on user interaction alone without any relevance judgments.
• User action bigrams are useful implicit feedback signals.
• Combining online & offline outperforms online only a little bit.

How to apply to evaluate search engines?

• Select a representative subset of real search logs
• Split search logs into separate goals (Jones & Klinkner, CIKM ’08)
• Data annotation & train prediction models
• Put systems to be evaluated into real user traffic
• Evaluate and compare systems by predicted success rates
• Can apply to UX measures other than success
Summary

Offline evaluation (the Cranfield style)
- Test collection + metric + significance test
- Pros: automatic & repeatable; no additional cost
- Cons: cost of building a test collection; limited correlation with UX

Online evaluation
- Put systems into real user traffic
- Determining the goodness of a system based on user feedback
- Implicit feedback is usually preferred over explicit ones