Test Collections

Temporal Information Retrieval

Jaspreeet Singh
Outline

• Why do we need evaluation?
• What is a test collection?
  – Defining the task
  – Defining metrics
  – Choosing the right document collection
  – Defining topics and gathering judgements

Test collections for Temporal IR
Why do we need evaluation?
What is a test collection

• Documents + topics + judgements = test collection

• Example of a test collection:
  – TREC diversity web track test collection
    • Web crawl
    • Input queries
    • User relevance judgements

How did they build this collection?
1. Define the task

- Think of a real world scenario and abstract it.

Task abstraction

<table>
<thead>
<tr>
<th>In the real world...</th>
<th>In the abstract world...</th>
</tr>
</thead>
<tbody>
<tr>
<td>user has context</td>
<td>user has no context</td>
</tr>
<tr>
<td>searches to accomplish a larger goal</td>
<td>searches occur in isolation</td>
</tr>
<tr>
<td>searches many times</td>
<td>searches once</td>
</tr>
<tr>
<td>reads a few documents, jumps around</td>
<td>reads linearly through the ranked list</td>
</tr>
<tr>
<td>consumes information in a variety of ways</td>
<td>reading counts for relevance</td>
</tr>
<tr>
<td>goals change over time</td>
<td>goal is abstract</td>
</tr>
</tbody>
</table>

- Operationalize the task – Break it down into simple steps
- Drives the rest of the test collection
The 100m sprint

- Intention and task? Who is the world's fastest man?
- Measure? Time taken
- Documents? The standard 100m track
- Competitors? The contestants
- Judgements? At the Olympics using a watch and then entered in the record books
In Information Retrieval ...

- TREC Diversity Track

  „...a diversity task whose goal was to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list“

Which algorithm produces the best ranking?
2. Determine the metric

- How do you measure the performance of a retrieval model?
- Precision—What fraction of retrieved results for a query are relevant?
  \[
  \text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}
  \]
- Recall—What fraction of the relevant results for a query are returned?
  \[
  \text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}
  \]
- set-based measures
“The justification is that almost anyone would be prepared to look at a few more documents if it would increase the percentage of the viewed set that were relevant (that is, if the precision of the larger set is higher)”
User centric metrics

- nDCG – Normalised Discounted Cumulative Gain

\[ CG_p = \sum_{i=1}^{p} rel_i \quad DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)} \quad nDCG_p = \frac{DCG_p}{IDCG_p} \]

\[ G[k] = \sum_{i=1}^{m} J(d_k, i)(1 - \alpha)^{r_i,k-1}. \]

Alpha – nDCG (takes diversity and novelty into account)

Captures the users seeking behavior

- ERR – estimated reciprocal rank

- [http://plg.uwaterloo.ca/~gvtermac/novelty.pdf](http://plg.uwaterloo.ca/~gvtermac/novelty.pdf)
Diversity track measures

• Intent Aware metrics for diversity
  – IA-Precision @ k
  – Subtopic Recall @ k
  – MAP (Mean Avg. Precision)

\[
\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{q} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})
\]

3. Document Collection

• Tasks sometimes imply the collection.

• Tweet filtering? – Twitter dataset
• Diversity? – which dataset?

• The documents affect how the systems will search. Changing the documents will change the performance of systems.

• Collections can opportunistic (email conversations), constructed (tweets for a hashtag) or naturalistic (large web crawl)
4. Find Topics

- Try to put yourself in the user’s shoes
- Find the intent and not just the query

- How do „you“ build these topics?
  - Manual
  - Query log driven
  - Explore the collection
  - Observe real users

Topics are hard to find! More topics you have the lesser the variability in your results.
I want to know the history of Rudolph Giuliani the American politician between 1987-2007.

- **Giuliani the litigator. Life as a lawyer in New York.**
- **Mayoral Campaigns - 1989 (losing to Dinkins) 1993 (improving police protection, beating Dinkins) 1997 (first Republican to win a second term in New York City)**
- **Mayoralty - mayor of New York City from 1994 through 2001. The major obstacles he had to overcome during his time as mayor. (Law enforcement, New York City budget, etc.)**
- **2000 U.S. Senate campaign. His main opponent being Hilary Clinton.**
- **September 11 terrorist attacks. Giuliani's work for helping New York recover.**
- **Post-mayoralty - what did Giuliani do in the political scene after leaving his post as mayor (after 2001) (running for president for 2000, became CIA Director; head of the NYC Anti-Terrorism Commission)**
- **His personal life - knighthood, time person of the year, cancer, affair, divorce**
Diversity

- Jaguar
  - Car
  - Guitar
  - Animal
5. Judging relevance

• In Ad-hoc retrieval, relevance is defined minimally as:
  "A document is relevant if any part of the document is relevant, even a single sentence"

• And independently:
  "A document is relevant independent of all other documents the user has already seen."
Eliciting relevance

• Having this simple definition of relevance makes it easier for the assesor

• It is very important to have clear (and minimal) instructions for the assessors otherwise the users have to take decisions „is it relevant enough?“

• Leads to inconsistencies -> bad test collection

• How do you judge large scale collections for a ranking task? – Pooling
  • All competitors are trying to solve the same problem. Each competitor submits „runs“ which are pooled together and evaluated.
Other things to consider ...

• Is the test collection stable?
  – Standard stability tests: Buckley and Voorhess (SIGIR 2000)

• Do you need expert assessors?
  – TREC employs ex-CIA officials to judge relevance (may not be true but they do not use students certainly)

• Do the assessors agree with each other?
  – Inter rater agreement
Summary

• Define the task first. It influences everything else.
• Select the metric to evaluate performance in the task
• Choose the document collection.
• Choose topics which abstract the user’s real world need.
• Judge documents for a given topic independently and with a minimum requirement.
Borrowed heavily from:


• More on the latest TREC web track : http://www-personal.umich.edu/~kevynct/trec-web-2014/
Historical Search

• Task: I want to know the history of ..... 
• Collection: News archive 
• Measure: Time aware Subtopic Recall 
• Topics and judgements?

• http://pharos.l3s.uni-hannover.de:7080/ArchiveSearch/starterkit/relevance.html
IR experiments look a bit like ths....

<table>
<thead>
<tr>
<th>Method</th>
<th>T-Sbr</th>
<th>TIA-NDCG</th>
<th>TIA-PREC.</th>
<th>TIA-MAP</th>
<th>TIA-ERR</th>
<th>TIA-SBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>0.302</td>
<td>0.209</td>
<td>0.01</td>
<td>0.01</td>
<td>0.023</td>
<td>0.453</td>
</tr>
<tr>
<td>TIA-SELECT</td>
<td>0.325</td>
<td>0.213</td>
<td>0.01</td>
<td>0.012</td>
<td>0.028</td>
<td>0.456</td>
</tr>
<tr>
<td>T-PM2</td>
<td>0.182</td>
<td>0.107</td>
<td>0.011</td>
<td>0.012</td>
<td>0.022</td>
<td>0.322</td>
</tr>
<tr>
<td>IA-SELECT</td>
<td>0.258</td>
<td>0.161</td>
<td>0.008</td>
<td>0.009</td>
<td>0.02</td>
<td>0.376</td>
</tr>
<tr>
<td>PM2</td>
<td>0.295</td>
<td>0.192</td>
<td>0.011</td>
<td>0.011</td>
<td>0.025</td>
<td>0.444</td>
</tr>
<tr>
<td>MDIV</td>
<td>0.309</td>
<td>0.209</td>
<td>0.009</td>
<td>0.011</td>
<td>0.025</td>
<td>0.454</td>
</tr>
<tr>
<td>ONLYTIME</td>
<td>0.344</td>
<td>0.22</td>
<td>0.007</td>
<td>0.01</td>
<td>0.024</td>
<td>0.482</td>
</tr>
<tr>
<td>HISTDIV</td>
<td>0.351</td>
<td>0.275</td>
<td>0.01</td>
<td>0.012</td>
<td>0.031</td>
<td>0.519</td>
</tr>
</tbody>
</table>

Table 2: Retrieval Effectiveness ($k = 10$)