Constructing Test Collections using Multi-armed Bandits and Active Learning

Md Mustafizur Rahman¹, Mucahid Kutlu², Matthew Lease¹
¹School of Information, ²Computer Engineering Department, ¹The University of Texas at Austin, ²TOBB University of Economics and Technology

Research Goal
Can we develop a test collection without organizing a shared task?

Shortcomings of a shared task
- Organizing a shared task is difficult, slow, and expensive
- Sometimes, it is impossible to gather enough participants, such as for less studied languages (e.g., Turkish) or search tasks (e.g., historical search).

Challenges
- How to allocate budget across topics?
- How to select documents for annotations?

Background
A test collection consists of
I. A collection of documents
II. A set of topics
III. A set of relevance judgments

Test collections are typically constructed
- by organizing a shared task, where multiple teams participate and submit their document rankings for the given document collection and the set of topics
- by applying the pooling, where the top-ranked \(K\) documents from each submitted ranking system are selected for relevance judging

Methods

Topic Selection
- Different topics need different number of relevance judgments.
- Allocating a pre-defined budget across topics will incur more cost than actually needed.
- To find out as many as relevant documents, we frame the problem as an exploration-exploitation phenomenon where we
  - Either exploit an already selected topic
  - Or explore a new topic.
- We solve the exploration-exploitation phenomenon using Multi-armed Bandits (MAB) technique [1].

Challenges
- Different topics need different number of relevance judgments.
- Allocating a pre-defined budget across topics will incur more cost than actually needed.
- To find out as many as relevant documents, we frame the problem as an exploration-exploitation phenomenon where we
  - Either exploit an already selected topic
  - Or explore a new topic.
- We solve the exploration-exploitation phenomenon using Multi-armed Bandits (MAB) technique [1].

Datasets

Table 1: Test collection statistics

<table>
<thead>
<tr>
<th>Track</th>
<th>Collection</th>
<th>Topics</th>
<th>#Docs</th>
<th>#Judged</th>
<th>%Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC-8[2]</td>
<td>Didees5-CR²</td>
<td>401-450</td>
<td>528K</td>
<td>86,830</td>
<td>5.4</td>
</tr>
</tbody>
</table>

Baselines

Topic Selection
- Oracle
- Round-robin (RR)
- Move-to-front (MTF) [3]
- MaxMean Non-Stationary (MM-NS) [4]

Results

Table 2: Avg. number of relevant documents found under varying budget per topic on TREC-8 for MTF [3], MM-NS [4], & MAB+CAL.

<table>
<thead>
<tr>
<th>Method</th>
<th>100</th>
<th>300</th>
<th>500</th>
<th>700</th>
<th>900</th>
<th>1100</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTF</td>
<td>34.06</td>
<td>58.84</td>
<td>71.78</td>
<td>79.22</td>
<td>84.5</td>
<td>87.58</td>
<td>94.04</td>
</tr>
<tr>
<td>MM-NS</td>
<td>66.64</td>
<td>66.64</td>
<td>66.64</td>
<td>66.64</td>
<td>66.64</td>
<td>66.64</td>
<td>66.64</td>
</tr>
<tr>
<td>MAB-CAL</td>
<td>46.3</td>
<td>78.6</td>
<td>86.5</td>
<td>90.3</td>
<td>91.3</td>
<td>93.5</td>
<td>94.04</td>
</tr>
</tbody>
</table>

References