Probabilistic Multileave for Online Retrieval Evaluation

Anne Schuth, Robert-Jan Bruintjes, Fritjof Büttner, Joost van Doorn, Carla Groenland, Harrie Oosterhuis, Cong-Nguyen Tran, Bas Veeling, Jos van der Velde, Roger Wechsler, David Woudenberg, Maarten de Rijke

anne.schuth@uva.nl  firstname.lastname@student.uva.nl  derijke@uva.nl

University of Amsterdam

A/B Testing

- User population split into two groups
- Many metrics are compared
- + Any metric can be measured using A/B testing
- - Not very sensitive, between subject design. Noise coming from differences between users and their queries.

Interleaved Comparisons

- Track team assignments
- System with most clicks wins comparison
- + Sensitive, within subject design.
  - About 100 times less interactions needed compared to A/B testing.
- - Only pairwise. Given a set of systems, quadratic comparisons are required. Often prohibitive.

Multileaved Comparisons (TDM)

- There are teams for each system
- Limited number of systems can be represented
- + Highly sensitive, within subject design.
  - Even more sensitive than interleaving, depending on the number of systems and result list length.
- +/- Many rankings at a time. But not many more than can be represented in the result list.
- - No reuse of historical interaction data.
  - Comparisons always involve a user.

Probabilistic Multileaved Comparisons (PM)

- Consider sample of all possible team assignments
- Outcome weighted by probability of assignment
- + Highly sensitive, within subject design.
  - As sensitive as TDM Multileaved comparisons.
- + Unlimited number of systems at a time.
- + Reuse of historical interaction data is possible.
  - Sets of new systems can be compared using historical clicks.

<table>
<thead>
<tr>
<th></th>
<th>perfect</th>
<th>navigational</th>
<th>informational</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI</td>
<td>0.085 (0.08)</td>
<td>0.137 (0.11)</td>
<td>0.363 (0.15)</td>
</tr>
<tr>
<td>TDM</td>
<td>0.037 (0.06)</td>
<td>0.038 (0.05)</td>
<td>0.099 (0.09)</td>
</tr>
<tr>
<td>PM((n=10^3))</td>
<td>0.062 (0.07)</td>
<td>0.073 (0.07)</td>
<td>0.162 (0.10)</td>
</tr>
<tr>
<td>PM((n=10^4))</td>
<td>0.054 (0.05)</td>
<td>0.060 (0.06)</td>
<td>0.117 (0.09)</td>
</tr>
<tr>
<td>PM((n=10^5))</td>
<td>0.046 (0.05)</td>
<td>0.054 (0.05)</td>
<td>0.090 (0.08)</td>
</tr>
<tr>
<td>PM((n=10^6))</td>
<td>0.046 (0.05)</td>
<td>0.039 (0.05)</td>
<td>0.087 (0.08)</td>
</tr>
</tbody>
</table>