Ranking Interruptus: When Truncated Rankings Are Better and How to Measure That

Enrique Amigó, Stefano Mizzaro, Damiano Spina
Which Ranking is Better?

Relevant Document

Non-relevant Document

Ranking A

Ranking B

Same effectiveness according to:
AP
RR
RBP
NDCG
...

Cutoff decided by the system!

Truncated Ranking

Which metric tells us that Ranking B > Ranking A?
Why Truncated Rankings?

Conversational Search

Small Screens
How to Pick the Right Metric?

“My system is the best one!”

Axiomatic Analysis

System Effectiveness vs. User Satisfaction
Axiomatic Analysis of Effectiveness Metrics

A general evaluation measure for document organization tasks

Axiomatic analysis for ad-hoc retrieval

Extended to diversity metrics + Rank-Biased Utility (RBU)

On the nature of information access evaluation metrics:
a unifying framework

Theoretical validation of the framework

Truncated Rankings
Contributions

1. Formal properties for truncated rankings
2. Extension of traditional metrics (ERR, NDCG, and RBP) by adding a user effort factor
3. Theoretical analysis of effectiveness metrics:
   • De-facto standard metrics do not satisfy desirable properties to evaluate truncated rankings
   • Observational Information Effectiveness (OIE) satisfies them all
4. Empirical evidence using 9 TREC test collections
5. Guidelines for metric selection in different ranking scenarios
### Desirable Properties

<table>
<thead>
<tr>
<th>Metric</th>
<th>Non-Truncated</th>
<th>Truncated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Priority</td>
<td>Top-W.</td>
</tr>
<tr>
<td>FULL RANKING</td>
<td>Spearman</td>
<td>✓</td>
</tr>
<tr>
<td>METRICS</td>
<td>Kendall</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>AUC-ROC</td>
<td>✓</td>
</tr>
<tr>
<td>BINARY</td>
<td>P@N</td>
<td>✓</td>
</tr>
<tr>
<td>RELEVANCE</td>
<td>R@N</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>R-p</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>F-measure</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>AP</td>
<td>✓</td>
</tr>
<tr>
<td>GRADED</td>
<td>NDCG</td>
<td>✓</td>
</tr>
<tr>
<td>RELEVANCE</td>
<td>Q-measure</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>BPref</td>
<td>✓</td>
</tr>
<tr>
<td>PROBABILISTIC</td>
<td>ERR</td>
<td>✓</td>
</tr>
<tr>
<td>USER MODEL</td>
<td>RBP</td>
<td>✓</td>
</tr>
<tr>
<td>BASED</td>
<td>iRBU</td>
<td>✓</td>
</tr>
<tr>
<td>TERMINAL</td>
<td>NDCGT</td>
<td>✓</td>
</tr>
<tr>
<td>DOCUMENT</td>
<td>ERR-T</td>
<td>✓</td>
</tr>
<tr>
<td>BASED</td>
<td>RBPT</td>
<td>✓</td>
</tr>
<tr>
<td>UTILTY</td>
<td>Flat Utility</td>
<td>✓</td>
</tr>
<tr>
<td>BASED</td>
<td>RBU</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>DCGU</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ERRU</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>RBPU</td>
<td>✓</td>
</tr>
<tr>
<td>INFORMATION</td>
<td>OIE</td>
<td>✓</td>
</tr>
</tbody>
</table>

- **Designed for non-truncated rankings**
- **Add one ‘relevant’ terminal document at the bottom of the ranking**
- **Add a component to take “effort” into account**
- **Based on Information Theory: quantity of information obtained from ranking and ground truth**
Empirical Validation

• Nine TREC test collections
• Optimal Truncation Points
• Truncated vs. Full Ranking Effectiveness
• Validation of the aspects captured by truncated rankings
(Some) Results

• Confirmation of theoretical analysis
  • Observational Information Effectiveness (OIE)
  • RBPU: extension of RBP that incorporates effort penalty

• For most metrics, optimal truncation is rather substantial

Figure 2: The distributions of optimal ranking lengths according to different truncation metrics. Results are averaged across the nine datasets.
Guidelines for Metric selection

To help practitioners identifying which effectiveness metric should be used for different ranking tasks / conditions
Summary

Evaluation of Truncated Rankings

Axiomatic Analysis + Empirical Validation

Future work (long road ahead!)
- Experiments with systems which incorporate a stopping criterion
- User studies to calibrate penalty effort
- ...

Guidelines to Choose the Appropriate Metric