Desirable Properties for Diversity and Truncated Effectiveness Metrics

Ameer Albahem1 Damiano Spina1 Falk Scholer1 Alistair Moffat2 Lawrence Cavedon1

The 23rd Australasian Document Computing Symposium
11-12 December 2018, Dunedin, New Zealand

Problem

- Complex search scenarios such as dynamic search are evaluated based on several dimensions: relevance, novelty and effort
- As metrics model multiple dimensions, understanding their behavior and suitability to search tasks is critical
- Existing meta-analysis approaches either do not model the dimensions or do not quantify the behavior

Contribution

A meta-analysis framework that can:
- Model desirable evaluation dimensions separately or simultaneously
- Quantify metrics behavior
- Provide cases for counter-intuitive behavior

Framework

Properties:
- **Relevance Monotonicity (Rel Mon):** If a ranking is extended by a single relevant document, then the computed effectiveness score should not decrease, but might remain the same
- **Irrelevance Monotonicity (Irrel Mon):** If a ranking is extended by a single non-relevant document, then the computed effectiveness score should not increase
- **Redundancy (Red):** A diversification metric should not favor a ranking that adds a document relevant to an already-seen aspect over a ranking that adds a relevant document to an unseen aspect
- **Induction:** If there is a directed path from the ranking Y to the ranking Z, then Y is inferior or equal to ranking Z

Case-analysis steps:
- Generate all cases of a truncated and diversified ranking of a maximum length of m
- Build tree-based relationships of rankings using properties.
- Evaluate all generated rankings using the metrics
- For every metric, check and count cases where a metric violates any of the properties proposed

Analysis

Setup:
- Truncated and Diversified rankings of up to m=10 documents
- Two aspects (A and B)

Metrics:
- **Ad hoc:** RR, P@5, P@10, NDCG@5, NDCG@10 and AP
- **Diversity:** Subtopic Recall; Intent-Aware: Precision (P-IA), Average Precision (AP-IA), and Diversity-aware Expected Reciprocal Rank, α-NDCG, and Novelty and Rank Biased Precision (NRBP)
- **Diversity and Truncation:** Cube Test (CT), Normalized Cube Test (nCT), and the Average Cube Test (ACT)

<table>
<thead>
<tr>
<th>Metric</th>
<th>ACT</th>
<th>AP-IA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>29,496</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Properties</th>
<th>Rel Mon</th>
<th>Irrel Mon</th>
<th>Red</th>
<th>Induction</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>(29,523)</td>
<td>(59,046)</td>
<td>(2,026)</td>
<td>(90,595)</td>
</tr>
<tr>
<td>AP-IA</td>
<td>0</td>
<td>0</td>
<td>2,026</td>
<td>6,892</td>
</tr>
<tr>
<td>Total</td>
<td>110,836</td>
<td>8,918</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of cases (number of edges)