The Effect of Pooling and Evaluation Depth on Metric Stability

William Webber  Alistair Moffat  Justin Zobel

Department of Computer Science and Software Engineering
The University of Melbourne

The 3rd International Workshop on Evaluating Information Access
Overview

What effect do:

- Evaluation depth
- Assessment depth
- Normalization
- Choice of metric

have upon discriminative power in assessment?
Motivation

Moffat and Zobel designed the RBP metric.

RBP has nice mathematical properties.

RBP also has an intuitive, plausible user model.
RBP poor’s discrimination

But studies showed RBP had poorer discrimination that AP, nDCG.

RBP and nDCG are very similar, rank-weighted metrics.
How RBP and nDCG differ

Main differences are:

- RBP is not normalized
- RBP weights decline smoothly, nDCG is steep–flat
- RBP typically not very deep
Hypotheses

nDCG is more discriminative because:

- of normalization
- because it validly makes use of more (deeper) relevance information
- because it is misled by evaluation beyond pooling depth
Evaluation and pooling depth

Allow that every system is pooled.

Evaluation frequently performed beyond pooling depth.

TREC: pool depth 100, evaluate depth 1,000.

82% of DCG’s rank weight to depth 1,000 falls beyond depth 100.
Rank weights for nDCG, RBP

![Graph showing rank weights for nDCG and RBP](image)

- **DCG**
- **RBP, p=0.95**
- **RBP, p=0.977**
- **RBP, p=0.9977**

**Webber, Moffat, and Zobel (UniMelb)**

**Depth and Metric Stability**

**EVIA 2010**
Discriminative power

For a metric.

Measured on a particular set of runs.

Proportion of run pairs whose difference in effectiveness is statistically significant.

Popularized by Sakai, using bootstrap. We use $t$ test.
Discriminative power

<table>
<thead>
<tr>
<th>Metric</th>
<th>T5</th>
<th>T8</th>
<th>T01</th>
<th>T04</th>
<th>T05</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>P@10</td>
<td>0.628</td>
<td>0.645</td>
<td>0.594</td>
<td>0.516</td>
<td>0.555</td>
<td>0.588</td>
</tr>
<tr>
<td>RBP, p=0.8</td>
<td>0.638</td>
<td>0.657</td>
<td>0.602</td>
<td>0.517</td>
<td>0.562</td>
<td>0.595</td>
</tr>
<tr>
<td>RBP, p=0.95</td>
<td>0.661</td>
<td>0.691</td>
<td>0.627</td>
<td>0.598</td>
<td>0.658</td>
<td>0.647</td>
</tr>
<tr>
<td>AP@1000</td>
<td>0.638</td>
<td>0.725</td>
<td>0.627</td>
<td>0.680</td>
<td>0.748</td>
<td>0.683</td>
</tr>
<tr>
<td>nDCG@1000</td>
<td>0.693</td>
<td>0.718</td>
<td>0.673</td>
<td>0.673</td>
<td>0.762</td>
<td>0.704</td>
</tr>
<tr>
<td>mean</td>
<td>0.651</td>
<td>0.687</td>
<td>0.624</td>
<td>0.597</td>
<td>0.657</td>
<td>0.643</td>
</tr>
</tbody>
</table>

Table: Discriminative power of standard metrics on different TREC collections. The most discriminative metric for each collection is highlighted.
Metric similarity

<table>
<thead>
<tr>
<th>R@10</th>
<th>AP@10</th>
<th>nDCG</th>
<th>RBP</th>
<th>P@1k</th>
<th>R@1k</th>
<th>AP@1k</th>
<th>nDCG</th>
<th>RBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.88</td>
<td>0.90</td>
<td>0.94</td>
<td>0.93</td>
<td>0.74</td>
<td>0.69</td>
<td>0.83</td>
<td>0.83</td>
<td>0.80</td>
</tr>
<tr>
<td>0.90</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.71</td>
<td>0.68</td>
<td>0.83</td>
<td>0.82</td>
<td>0.77</td>
</tr>
<tr>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.73</td>
<td>0.70</td>
<td>0.86</td>
<td>0.85</td>
<td>0.79</td>
</tr>
<tr>
<td>0.98</td>
<td></td>
<td>0.98</td>
<td></td>
<td>0.71</td>
<td>0.66</td>
<td>0.80</td>
<td>0.81</td>
<td>0.78</td>
</tr>
<tr>
<td>RBP</td>
<td>0.72</td>
<td>0.67</td>
<td>0.81</td>
<td>0.81</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P@1k</td>
<td>0.88</td>
<td>0.81</td>
<td>0.85</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R@1k</td>
<td>0.79</td>
<td>0.84</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP@1k</td>
<td>0.91</td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nDCG</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: Kendall’s $\tau$ between system rankings on the TREC 8 AdHoc track participant systems, using different metrics.
**nDCG at depths**

![Graphs showing relationships between system mean nDCG scores at different pooling and evaluation depths, for the TREC 8 AdHoc runset.](image)

**Figure:** Relationship of system mean nDCG scores at different pooling and evaluation depths, for the TREC 8 AdHoc runset.
Cutoff depths for normalization

\[ AP \equiv eAP \]

**Figure:** AP normalized by \( R \) versus AP normalized by \( \max(k, R) \)
Cutoff, pooling, discrimination

Figure: Pooling and evaluation depth, and discriminative power

\[ \text{RBP, resid} = 0.1 \]

\[ \text{aAP} \]
## Correlation of significance

<table>
<thead>
<tr>
<th>Pool</th>
<th>Eval</th>
<th>Metric</th>
<th>Pool@10 Eval@10</th>
<th>Pool@10 Eval@10</th>
<th>Pool@100 Eval@100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>nDCG RBP</td>
<td>nDCG RBP</td>
<td>aAP nDCG RBP</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>aAP</td>
<td>0.89 0.88</td>
<td>0.73 0.72 0.67</td>
<td>0.74 0.74 0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nDCG</td>
<td>0.96</td>
<td>0.73 0.75 0.70</td>
<td>0.73 0.76 0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RBP</td>
<td></td>
<td>0.72 0.74 0.69</td>
<td>0.72 0.75 0.73</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>aAP</td>
<td>0.88 0.84</td>
<td>0.86 0.86 0.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>nDCG</td>
<td>0.88</td>
<td>0.79 0.88 0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RBP</td>
<td></td>
<td>0.75 0.81 0.85</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>aAP</td>
<td>0.87 0.82</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>nDCG</td>
<td></td>
<td>0.88</td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Kendall’s $\tau$ between $p$ values assigned to TREC 8 AdHoc system pairs by paired, two-tailed $t$ tests.
Recapitulating hypotheses

The original hypotheses:

- Normalization helps discriminative power
- Evaluating beyond pooling depth misleadingly helps discriminative power
- Greater evaluation depth helps discriminative power
Normalization doesn’t help discriminative power

**Figure:** Pooling and evaluation depth, and discriminative power
Evaluation beyond pooling depth is not misleading

Figure: (a) short and full pooling similar scores; (b) short and full pooling similar discrimination
Deep evaluation picks up useful information

Figure: Evaluation depth the most important, consistent determinant of discriminative power.
New hypothesis: DCG weights good

Figure: Effect of increasing evaluation depth on discriminative power
New hypothesis: DCG weights good

Deepening evaluation by raising RBP $p$ harms discriminative power with short pooling.

Has little effect on DCG.

Steep–flat weighting of DCG may actually be (by chance?) well suited.
Questions