CIR at the NTCIR-17
Unbiased Learning to Rank Evaluation Task 2
(ULTRE-2)

Lulu Yu, Keping Bi, Jiafeng Guo, Xueqi Cheng
Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS)
CAS Key Lab of Web Data Science and Technology
Background
Background

• Learning to Rank

<table>
<thead>
<tr>
<th>Annotation dataset</th>
<th>Relevance label</th>
<th>Ranking model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query – Docs</td>
<td>4, 1, 3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Search logs</th>
<th>Click label</th>
<th>Non-click label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query – Docs</td>
<td>1, 2, 3</td>
<td></td>
</tr>
</tbody>
</table>

Expensive to obtain

Easy to collect & implicit relevance
Background

• Unbiased Learning to Rank

<table>
<thead>
<tr>
<th>search logs</th>
</tr>
</thead>
<tbody>
<tr>
<td>query – docs</td>
</tr>
<tr>
<td>- click</td>
</tr>
<tr>
<td>- non-click</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>- non-click</td>
</tr>
</tbody>
</table>

• Biases
  - Position bias
  - Sample selection bias
  - ...
Background

- NTCIR-17 ULTRE-2 task

Unbiased Learning to Rank

- synthetic click data
- real-world click data from a commercial Web search engine, Baidu

ranking model
Motivation & Methods
Motivation

- non-clicks do not mean irrelevant ⇔ false negative issue

Non-clicks: false negative issue
Methods

- Label Correction
  
  Correct the labels for non-clicked items by a relevance judgment model trained from DLA.
  
  \[
  \text{DLA} \quad \text{Ranking} \model
  \]

  \[
  DLA_0 \quad \text{reasonableness} \quad l_i = \begin{cases} 
  \text{sig: sigmoid}(o_i) & \text{if } o_i \geq \min(o_j, \ldots) \quad c_j = 1 \\
  \min: 1 & \text{otherwise}
  \end{cases}
  \]

  \[
  o: \text{the output of } DLA_0
  \]
## Methods

- **Negative Sampling**
  - through negative sampling → **reconstruct the original result lists**

  - “click-only” scheme: preserve clicked results
  - “last-click” scheme: preserve all the results before the last clicked result

\[
\mathcal{D} = \{ \text{doc}_1, \ldots, \text{doc}_i \}\text{-click}=1
\]

\[
\{ \text{random negatives}_{1, \ldots, j} \}\text{-label}=0
\]

\[
\{ \text{hard negatives}_{1, \ldots, k} \}\text{-label}=0
\]

\[
\text{Naive Algorithm}
\]

\[
\text{Ranking model}
\]

\[
\text{loss} = -\sum_{x \in \mathcal{D'}} \frac{e^f(x)}{\sum_{x \in \mathcal{D'}} e^f(x)}
\]
Experiments & Results
Experiments

• Experimental implementation
  • Input features
    • for label correction method
      • traditional features of 13 dimensions
    • pretrained score
  • for negative sampling method
    • extracted traditional word matching features (e.g. LM-DIR, BM25) of 24 dimensions

• Model architecture
  • Feature projection: project features to a higher dimension
  • Ranking model: a deep neural network with three hidden layers
Results

- **Effect of Negative Sampling**
  - We investigate the use of negative sampling on the validation set.
  - The nDCG@10 on the validation set indicates that this approach is effective in improving performance.

Performance curves of two schemes ("click-only" and "last-click") w.r.t. the number of random and hard negatives.
(a) Performance curves of two schemes w.r.t. the number of random negatives.
(b) The Performance curve of the "click-only" scheme w.r.t. the number of hard negatives.
Results

• Effect of Label Correction
  ➢ DLA with Label Correction **outperforms** the basic DLA model, under various strategies.
  ➢ The underline denotes the performance of the baseline DLA.

<table>
<thead>
<tr>
<th>Model</th>
<th>nDCG@10</th>
<th>DCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>test</td>
<td>total valid</td>
</tr>
<tr>
<td>Scratch-DLA-LC (sig)</td>
<td>0.5355</td>
<td>0.5019</td>
</tr>
<tr>
<td>Aux-DLA-LC (sig)</td>
<td>0.5326</td>
<td>0.5015</td>
</tr>
<tr>
<td>Scratch-DLA-LC (min)</td>
<td>unk</td>
<td>0.4816</td>
</tr>
<tr>
<td>Aux-DLA-LC (min)</td>
<td>unk</td>
<td>0.4947</td>
</tr>
<tr>
<td>DLA</td>
<td>0.5247</td>
<td>0.4920</td>
</tr>
<tr>
<td>lgbBase</td>
<td>0.5350</td>
<td>/</td>
</tr>
<tr>
<td>lgbBaseAdd</td>
<td>0.5333</td>
<td>/</td>
</tr>
</tbody>
</table>
Conclusion
Conclusion

• We focus on the false negative issue and propose two approaches to tackle this issue: label correction and negative sampling.

• Both methods can enhance the model performance and our best method (label correction) has achieved nDCG@10 of 0.5355, which is 2.66% better than the best score from the organizer.
Thanks!

yululu23s@ict.ac.cn