Top-K Learning to Rank: Labeling, Ranking and Evaluation

Shuzi Niu, Jiafeng Guo, Yanyan Lan, Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences
Outlines

• Motivation

• Top-K Learning to Rank Framework
  – Top-K Labeling Strategy
  – FocusedRank
  – Top-K Evaluation

• Experimental Results

• Conclusions & Future Work
Motivation

One great challenge for learning to rank: it is difficult to obtain reliable training data from human assessors!

**Absolute Relevance Judgment**

**Drawbacks:**

1. Choice of the specific of the gradations.
2. Increasing assessing burdens.
3. High level of disagreement on judgments.
Motivation (cont’)

Pairwise Preference Judgment

Pros:
(1) No need to determine the gradation specifications.
(2) Easier for an assessor to express a preference.
(3) Noise may be reduced.

Cons:
Complexity of judgment increases! (From $O(n)$ to $O(n^2)$, $O(n \log n)$.)

How to reduce the complexity of pairwise preference judgment?
Motivation (cont’)

• Do we really need to get a total ordering for each query? **NO!**

• Users mainly care about the top results in real web search application!

Take more effort to figure out the top results and judge the preference orders among them.

Top-K Ground-truth

Total ordering of top K results

Preferences between top K Documents and the other N-K documents
Motivation (cont’)

• Three Tasks:
  – How to design an efficient pairwise preference labeling strategy to get top-k ground-truth?
  – How to develop more powerful ranking algorithms in the new scenario?
  – How to define new evaluation measures for the new scenario?
Top-k Learning to Rank: Labeling

- Top-k Labeling Strategy
  - Pairwise preference judgment
  - HeapSort

Example: k=3, n=5

Step 1: O(k)
Step 2: O((n-k)log k)
Step 3: O(k log k)

Top-3 Ground-truth
Top-K Learning to Rank: Ranking

- New characteristics of top-k ground-truth

\[ L(f; q_i) = \beta L_{list}(f; T_i, y_i) + (1 - \beta) L_{pair}(f; P_i, y_i) \]

\[ FocusedRank \]

Listwise ranking algorithms

Pairwise ranking algorithms

- Struct-SVM
- AdaRank
- ListNet

- RankSVM
- RankBoost
- RankNet

- FocusedSVM
- FocusedBoost
- FocusedNet
Top-K Learning to Rank: Evaluation

• Traditional evaluation measures, e.g. MAP, NDCG, ERR, are mainly defined on absolute relevance scores.

• In the scenario of top-k ground-truth, define a position-aware relevance score:

\[ y_j^{(i)} = k + 1 - \pi_i(x_j^{(i)}), \text{ if } x_j^{(i)} \in T_i, \ y_j^{(i)} = 0, \text{ otherwise.} \]

- \( \kappa \)-NDCG

\[ \kappa - \text{NDCG}@l = \frac{1}{N'_l} \sum_{j=1}^{l} \frac{2^{y_j^{(i)}} - 1}{\log_2(1 + j)}, \]

- \( \kappa \)-ERR

\[ \kappa - \text{ERR} = \sum_{s=1}^{n} \frac{1}{n_i} R(y_s^{(i)}) \prod_{t=1}^{s-1} (1 - R(y_t^{(i)})), \quad R(r) = \frac{2^r - 1}{2y_m^{(i)}}, \]
Experiments

• Effectiveness and efficiency of top-k labeling strategy
  – Data Sets: all the 50 queries from Topic Distillation task of TREC 2003, for each query, sample 50 documents.
  – Labeling Tools: top-10 labeling tool T1 and five-graded relevance judgment tool T2.
  – Assessors: Five graduate students who are familiar with web search.
  – Assignment: Divided into five folds Q1,…Q5, Ui judges Qi with T1 and Qi+1 with T2, for i=1,2,3,4, and U5 judges Q5 with T1 and Q1 with T2.
Experimental Results I

• **Time Efficiency**

<table>
<thead>
<tr>
<th>Method</th>
<th>Time per judgment(s)</th>
<th>Time per query(min)</th>
<th>Judgment complexity</th>
<th>#Judgments per query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-k labeling</td>
<td>5.51</td>
<td>13.13</td>
<td>$O(n \log k)$</td>
<td>142.76</td>
</tr>
<tr>
<td>Five-grade judgment</td>
<td>13.87</td>
<td>11.78</td>
<td>$O(n)$</td>
<td>50</td>
</tr>
</tbody>
</table>

• **Agreement**

<table>
<thead>
<tr>
<th></th>
<th>A&gt;B</th>
<th>A~B</th>
<th>A&lt;B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&gt;B</td>
<td>0.6749</td>
<td>0.2766</td>
<td>0.0485</td>
</tr>
<tr>
<td>A~B</td>
<td>0.1138</td>
<td>0.8198</td>
<td>0.0664</td>
</tr>
<tr>
<td>A&lt;B</td>
<td>0.1047</td>
<td>0.3779</td>
<td>0.5174</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A&gt;B</th>
<th>A~B</th>
<th>A&lt;B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&gt;B</td>
<td>0.6272</td>
<td>0.2913</td>
<td>0.0815</td>
</tr>
<tr>
<td>A~B</td>
<td>0.2825</td>
<td>0.5232</td>
<td>0.1944</td>
</tr>
<tr>
<td>A&lt;B</td>
<td>0.1534</td>
<td>0.3826</td>
<td>0.4640</td>
</tr>
</tbody>
</table>

*Top 10 Labeling*  
*5 Graded Labeling*
Experiments (cont’)

• Performance of FocusedRank
  – Baselines:
    (1) Pairwise: RankSVM, RankBoost, RankNet,
    (2) Listwise: SVMMAP, AdaRank, ListNet,
    (3) Top-k: Top-k ListMLE
  – Data Sets:
    (1) MQ2007 (From LETOR): Graded MQ2007 and Top-k MQ2007
    (2) TD2003 (Previous constructed data): Graded TD2003 and Top-k TD2003
Experimental Results II

Performance comparison among FocusedRank, pairwise and listwise algorithms on Top-k datasets.
Performance comparison among FocusedRank, pairwise and listwise algorithms on Graded datasets.
Experimental Results II (cont’)

Performance comparison between FocusedRank and Top-k ListMLE on Top-k datasets.
Conclusions

• Top-K Learning to Rank Framework
  – Top-k labeling strategy: obtain reliable relevance judgments via pairwise preference judgment. Complexity is reduced to $O(n \log k)$.
  – FocusedRank: capture the characteristics of the top-k ground-truth.
  – Top-k evaluation measures

• Empirical studies show the efficiency and reliability of top-k labeling strategy, and demonstrate the effectiveness of FocusedRank.
Future Work

• Further reduce the complexity of top-k labeling strategy.
• Design new ranking models for top-k ranking.
• Rank aggregations of top-k ground-truth.
• Active learning in top-k labeling strategy.
Thanks for your Attention!

Thank SIGIR 2012 for providing Shuzi Niu the student travel grants!
Thank the committee for granting us the best student paper award!

lanyanyan@ict.ac.cn