Learning For Search Result Diversification

Yadong Zhu
Institute of Computing Technology,
Chinese Academy of Sciences
edgewind11@gmail.com

Co-authors: Yanyan Lan, Jiafeng Guo, Shuzi Niu, Xueqi Cheng
Outline

• Motivation
• Our Approach
• Experiments
• Conclusion
Motivation

• Different user needs
  – Ambiguous queries
    • Apple, Jaguar, Band...
  – Multi-faceted needs
    • Britney spears (news, videos, photos...)

• Information redundancy
  – Many duplicate or similar results
Motivation

• Existing approaches

Non-learning

✓ MMR
✓ IA-Select
✓ xQuAD
✓ PM-2
✓ ...

Learning-based

✓ SVMDIV
Non-learning Methods

• Typical methods
  – MMR: Maximal Marginal Relevance
    • Predefined utility function

\[
MMR \overset{\text{def}}{=} \arg \max_{d_i \in R \setminus S} [\lambda \text{Sim}_1(d_i, q) - (1 - \lambda) \max_{d_j \in S} \text{Sim}_2(d_i, d_j)]
\]

- Query Relevance
- Similarity with selected documents
Non-learning Methods

• Typical methods
  – xQuAD: Explicit query aspect diversification
  • Predefined utility function

\[
(1 - \lambda)P(d|q) + \lambda \sum_{q_i \in Q} [P(q_i|q)P(d|q_i) \prod_{d_j \in S} (1 - P(d_j|q_i))] 
\]

1) Heuristic predefined utility function;
2) Limited features incorporated;
Learning-based Methods

• Typical method
  – SVMDIV
    • Only focuses on diversity, discard the relevance
    • Propose to optimize subtopic coverage based on maximizing word coverage

How to model both relevance and diversity?
Outline

• Motivation
• **Our Approach**
• Experiments
• Conclusion
Our Approach

• Relational Learning-to-rank approach (R-LTR)
  – Considering both *content* of individual documents and *relations* among documents.

• Formalization
  – Four key components: input space, out space, ranking function $f$, loss function $L$

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^{N} L(f(X^{(i)}, R^{(i)}), y^{(i)}).$$
Challenges for R-LTR

- How to define ranking function
- How to define loss function
Definition of Ranking Function

Sequential Ranking Process

1) Top-down user browsing behavior;
2) NP-hard, greedy sequential approximation
Definition of Ranking Function

• Definition

\[ f_S(x_i, R_i) = \omega_r^T x_i + \omega_d^T h_S(R_i), \quad \forall x_i \in X \setminus S \]

• Relational function \( h_S(R_i) \)
  
  – Minimal Distance

\[ h_S(R_i) = (\min_{x_j \in S} R_{ij1}, \ldots, \min_{x_j \in S} R_{ijl}). \]

  – Average Distance

\[ h_S(R_i) = \left( \frac{1}{|S|} \sum_{x_j \in S} R_{ij1}, \ldots, \frac{1}{|S|} \sum_{x_j \in S} R_{ijl} \right). \]

  – Maximal Distance

\[ h_S(R_i) = (\max_{x_j \in S} R_{ij1}, \ldots, \max_{x_j \in S} R_{ijl}). \]
Definition of Ranking Function

• Relevance features
  – Traditional LTR relevance features, such as: TFIDF, bm25, LM, Proximity......

• Diversity features
  – Subtopic diversity: semantic distance based on topic distribution.
  – Text, title, anchor diversity based on cosine similarity;
  – ODP-based: existing ODP taxonomy
  – Link-based
  – url-based
  – ...
Definition of Loss Function

Sequential Ranking Process

Model the generation of the result list in a Sequential way

Loss function: likelihood loss of the generation probability

\[ L(f(X, R), y) = - \log P(y | X) \]

\[ P(y | X) = P(x_{y(1)} | X)P(x_{y(2)} | X \setminus S_1) \cdots P(x_{y(n)} | X \setminus S_{n-1}) \]
Definition of Loss Function

- **Plackett-Luce Model**

\[
P(\pi | \nu) = \prod_{i=1}^{M} \frac{v_{\pi(i)}}{v_{\pi(i)} + v_{\pi(i+1)} + \cdots + v_{\pi(M)}}
\]

- **Detailed definition**

\[
P(x_{y(1)} | X) = \frac{\exp\{f_\phi(x_{y(1)})\}}{\sum_{k=1}^{n} \exp\{f_\phi(x_{y(k)})\}} , \quad P(x_{y(j)} | X \setminus S_{j-1}) = \frac{\exp\{f_{S_{j-1}}(x_{y(j)}, R_{y(j)})\}}{\sum_{k=j}^{n} \exp\{f_{S_{k-1}}(x_{y(k)}, R_{y(k)})\}}.
\]

- **maximize the sum of the likelihood function**

\[
- \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log \left\{ \frac{\exp\{\omega_r^T x_{y(j)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}} (R_{y(j)})\}}{\sum_{k=j}^{n_i} \exp\{\omega_r^T x_{y(k)}^{(i)} + \omega_d^T h_{S_{k-1}^{(i)}} (R_{y(k)})\}} \right\}
\]
Learning

- Unconstrained optimization problem
  - Stochastic Gradient Descent

---

**Algorithm 2 Optimization Algorithm**

**Input:** training data \(\{(X^{(i)}, R^{(i)}, y^{(i)})\}_{i=1}^{N}\),
  parameter: learning rate \(\eta\), tolerance rate \(\epsilon\)

**Output:** model vector: \(\omega_r, \omega_d\)

1: Initialize parameter value \(\omega_r, \omega_d\)

2: repeat

3: Shuffle the training data

4: for \(i = 1, \ldots, N\) do

5: Compute gradient \(\Delta \omega_r^{(i)}\) and \(\Delta \omega_d^{(i)}\)

6: Update model:
   \[\omega_r = \omega_r - \eta \times \Delta \omega_r^{(i)},\]
   \[\omega_d = \omega_d - \eta \times \Delta \omega_d^{(i)}\]

7: end for

8: Calculate likelihood loss on the training set

9: until the change of likelihood loss is below \(\epsilon\)
## Prediction

- **Sequential Prediction Process**

<table>
<thead>
<tr>
<th>Algorithm 3 Ranking Prediction via Sequential Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> $X^{(t)}, R^{(t)}, \omega_r, \omega_d$</td>
</tr>
<tr>
<td><strong>Output:</strong> $y^{(t)}$</td>
</tr>
<tr>
<td><strong>1:</strong> Initialize $S_0 \leftarrow \emptyset, y^{(t)} = (1, \ldots, n_t)$</td>
</tr>
<tr>
<td><strong>2:</strong> for $k = 1, \ldots, n_t$ do</td>
</tr>
<tr>
<td><strong>3:</strong> bestDoc $\leftarrow \text{argmax}<em>{x \in X_t} f</em>{S_{k-1}}(x, R)$</td>
</tr>
<tr>
<td><strong>4:</strong> $S_k \leftarrow S_{k-1} \cup \text{bestDoc}$</td>
</tr>
<tr>
<td><strong>5:</strong> $y^{(t)}(k) \leftarrow \text{the index of bestDoc}$</td>
</tr>
<tr>
<td><strong>6:</strong> end for</td>
</tr>
<tr>
<td><strong>7:</strong> return $y^{(t)} = (y^{(t)}(1), \ldots, y^{(t)}(n_t))$</td>
</tr>
</tbody>
</table>
Outline

• Motivation
• Our Approach
• Experiments
• Conclusion
Experiments

• Dataset:

• Evaluation measures (K=20):
  – TREC official evaluation measures: ERR-IA, a-NDCG, NRBP;
  – Traditional diversity measures: Precision-IA, Subtopic Recall;

• Baseline methods:
  – QL, MMR, xQuAD, PM-2, ListMLE, SVMDIV

• Platform:
  – Indri toolkit (version 5.2)
TREC Official Measures

**Table 2: Performance comparison of all methods in official TREC diversity measures for WT2009.**

<table>
<thead>
<tr>
<th>Method</th>
<th>ERR-IA</th>
<th>α-NDCG</th>
<th>NRBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>0.1637</td>
<td>0.2691</td>
<td>0.1382</td>
</tr>
<tr>
<td>ListMLE</td>
<td>0.1913 (+16.86%)</td>
<td>0.3074 (+14.23%)</td>
<td>0.1681 (+21.64%)</td>
</tr>
<tr>
<td>MMR(_{list})</td>
<td>0.2022 (+23.52%)</td>
<td>0.3083 (+14.57%)</td>
<td>0.1715 (+24.09%)</td>
</tr>
<tr>
<td>xQuAD(_{list})</td>
<td>0.2316 (+41.48%)</td>
<td>0.3437 (+27.72%)</td>
<td>0.1956 (+41.53%)</td>
</tr>
<tr>
<td>PM-2(_{list})</td>
<td>0.2294 (+40.13%)</td>
<td>0.3369 (+25.20%)</td>
<td>0.1788 (+29.38%)</td>
</tr>
<tr>
<td>SVMDIV</td>
<td>0.2408 (+47.10%)</td>
<td>0.3526 (+31.03%)</td>
<td>0.2073 (+50.00%)</td>
</tr>
<tr>
<td><strong>R-LTR(_{min})</strong></td>
<td><strong>0.2714 (+65.79%)</strong></td>
<td><strong>0.3915 (+45.48%)</strong></td>
<td><strong>0.2339 (+69.25%)</strong></td>
</tr>
<tr>
<td><strong>R-LTR(_{avg})</strong></td>
<td><strong>0.2671 (+63.16%)</strong></td>
<td><strong>0.3964 (+47.31%)</strong></td>
<td><strong>0.2268 (+64.11%)</strong></td>
</tr>
<tr>
<td><strong>R-LTR(_{max})</strong></td>
<td><strong>0.2683 (+63.90%)</strong></td>
<td><strong>0.3933 (+46.15%)</strong></td>
<td><strong>0.2281 (+65.05%)</strong></td>
</tr>
<tr>
<td>TREC-Best</td>
<td>0.1922</td>
<td>0.3081</td>
<td>0.1617</td>
</tr>
</tbody>
</table>

**R-LTR approaches show better performance!**
### Table 3: Performance comparison of all methods in official TREC diversity measures for WT2010.

<table>
<thead>
<tr>
<th>Method</th>
<th>ERR-IA</th>
<th>α-NDCG</th>
<th>NRBPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>0.1980</td>
<td>0.3024</td>
<td>0.1549</td>
</tr>
<tr>
<td>ListMLE</td>
<td>0.2436 (+23.03%)</td>
<td>0.3755 (+24.17%)</td>
<td>0.1949 (+25.82%)</td>
</tr>
<tr>
<td>MMR&lt;sub&gt;list&lt;/sub&gt;</td>
<td>0.2735 (+38.13%)</td>
<td>0.4036 (+33.47%)</td>
<td>0.2252 (+45.38%)</td>
</tr>
<tr>
<td>xQuAD&lt;sub&gt;list&lt;/sub&gt;</td>
<td>0.3278 (+65.56%)</td>
<td>0.4445 (+46.99%)</td>
<td>0.2872 (+85.41%)</td>
</tr>
<tr>
<td>PM-2&lt;sub&gt;list&lt;/sub&gt;</td>
<td>0.3296 (+66.46%)</td>
<td>0.4478 (+48.08%)</td>
<td>0.2901 (+87.28%)</td>
</tr>
<tr>
<td>SVMDIV</td>
<td>0.3331 (+68.23%)</td>
<td>0.4593 (+51.88%)</td>
<td>0.2934 (+89.41%)</td>
</tr>
<tr>
<td>R-LTR&lt;sub&gt;min&lt;/sub&gt;</td>
<td><strong>0.3647 (+84.19%)</strong></td>
<td><strong>0.4924 (+62.83%)</strong></td>
<td><strong>0.3293 (+112.59%)</strong></td>
</tr>
<tr>
<td>R-LTR&lt;sub&gt;avg&lt;/sub&gt;</td>
<td>0.3587 (+81.16%)</td>
<td>0.4781 (+58.10%)</td>
<td>0.3125 (+101.74%)</td>
</tr>
<tr>
<td>R-LTR&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.3639 (+83.79%)</td>
<td>0.4836 (+59.92%)</td>
<td>0.3218 (+107.74%)</td>
</tr>
<tr>
<td>TREC-Best</td>
<td>0.2981</td>
<td>0.4178</td>
<td>0.2616</td>
</tr>
</tbody>
</table>

### Table 4: Performance comparison of all methods in official TREC diversity measures for WT2011.

<table>
<thead>
<tr>
<th>Method</th>
<th>ERR-IA</th>
<th>α-NDCG</th>
<th>NRBPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>0.3520</td>
<td>0.4531</td>
<td>0.3123</td>
</tr>
<tr>
<td>ListMLE</td>
<td>0.4172 (+18.52%)</td>
<td>0.5169 (+14.08%)</td>
<td>0.3887 (+24.46%)</td>
</tr>
<tr>
<td>MMR&lt;sub&gt;list&lt;/sub&gt;</td>
<td>0.4284 (+21.70%)</td>
<td>0.5302 (+17.02%)</td>
<td>0.3913 (+25.30%)</td>
</tr>
<tr>
<td>xQuAD&lt;sub&gt;list&lt;/sub&gt;</td>
<td>0.4753 (+35.03%)</td>
<td>0.5645 (+24.59%)</td>
<td>0.4274 (+36.86%)</td>
</tr>
<tr>
<td>PM-2&lt;sub&gt;list&lt;/sub&gt;</td>
<td>0.4873 (+38.44%)</td>
<td>0.5786 (+27.70%)</td>
<td>0.4318 (+38.26%)</td>
</tr>
<tr>
<td>SVMDIV</td>
<td>0.4898 (+39.15%)</td>
<td>0.5910 (+30.43%)</td>
<td>0.4475 (+43.29%)</td>
</tr>
<tr>
<td>R-LTR&lt;sub&gt;min&lt;/sub&gt;</td>
<td><strong>0.5389 (+53.10%)</strong></td>
<td><strong>0.6297 (+38.98%)</strong></td>
<td><strong>0.4982 (+59.53%)</strong></td>
</tr>
<tr>
<td>R-LTR&lt;sub&gt;avg&lt;/sub&gt;</td>
<td>0.5276 (+49.89%)</td>
<td>0.6219 (+37.25%)</td>
<td>0.4724 (+51.26%)</td>
</tr>
<tr>
<td>R-LTR&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.5285 (+50.14%)</td>
<td>0.6223 (+37.34%)</td>
<td>0.4741 (+51.81%)</td>
</tr>
<tr>
<td>TREC-Best</td>
<td>0.4380</td>
<td>0.5220</td>
<td>0.4070</td>
</tr>
</tbody>
</table>
R-LTR approaches also shown better performance!
Experiments

• Robustness Analysis
  – Win/Loss Ratio (vs QL)

Table 5: The robustness of the performance of all diversity methods in Win/Loss ratio

<table>
<thead>
<tr>
<th>Method</th>
<th>WT2009</th>
<th>WT2010</th>
<th>WT2011</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ListMLE</td>
<td>20/18</td>
<td>27/16</td>
<td>26/11</td>
<td>73/45</td>
</tr>
<tr>
<td>MMR&lt;sub&gt;list&lt;/sub&gt;</td>
<td>22/15</td>
<td>29/13</td>
<td>29/10</td>
<td>80/38</td>
</tr>
<tr>
<td>xQuAD&lt;sub&gt;list&lt;/sub&gt;</td>
<td>28/11</td>
<td>31/12</td>
<td>31/12</td>
<td>90/35</td>
</tr>
<tr>
<td>PM-2&lt;sub&gt;list&lt;/sub&gt;</td>
<td>26/15</td>
<td>32/12</td>
<td>32/11</td>
<td>90/38</td>
</tr>
<tr>
<td>SVM DIV</td>
<td>30/12</td>
<td>32/11</td>
<td>32/11</td>
<td>94/34</td>
</tr>
<tr>
<td>R-LTR&lt;sub&gt;min&lt;/sub&gt;</td>
<td>34/9</td>
<td>35/10</td>
<td>35/9</td>
<td>104/28</td>
</tr>
<tr>
<td>R-LTR&lt;sub&gt;avg&lt;/sub&gt;</td>
<td>33/9</td>
<td>34/11</td>
<td>34/10</td>
<td>101/30</td>
</tr>
<tr>
<td>R-LTR&lt;sub&gt;max&lt;/sub&gt;</td>
<td>33/10</td>
<td>35/10</td>
<td>34/10</td>
<td>102/30</td>
</tr>
</tbody>
</table>
Experiments

• Offline Training Time

ListMLE (~1.5h) < SVMDIV (~2h) < R-LTR (~3h)
Outline

• Motivation
• Our Approach
• Experiments
• Conclusion
Conclusion

• Contributions
  – Propose a novel relational learning-to-rank framework for search results diversification
  – The R-LTR is very general and can be easily extended to other fields such as summarization or recommendation.
  – Extensive experimental evaluation
http://www.yadongzhu.com

SIGIR Student Travel Grants 😊

Q&A