A Regularized Competition Model for Question Difficulty Estimation in Community Question Answering Services

Quan Wang†  Jing Liu‡  Bin Wang†  Li Guo†
† Institute of Information Engineering, CAS  ‡ Harbin Institute of Technology

Research Problem
- Question difficulty estimation in community question answering
- Applications
  - Question routing, incentive mechanism design, linguistic analysis

Previous Solutions
- Competition-based methods
  - Extract pairwise competitions from question answering threads
  - Estimate question difficulty based on extracted competitions
    - TrueSkill (Liu et al., 2013)
    - PageRank (Yang et al., 2008)
- Drawbacks
  - Data sparsity issue: each question gets only two competitions
  - Cold-start issue: cannot handle questions with no answers received

Our Solution
- Competitions + textual descriptions
  - For data sparsity issue: textual descriptions provide additional information
  - For cold-start issue: textual descriptions link cold-start questions to well-resolved ones

Regularized Competition Model

Assumption I: pairwise comparison assumption
- Question’s difficulty > asker’s skill
- Question’s difficulty < best answerer’s skill
- Best answerer’s skill > all other answerers’ skill

For assumption I: a margin-based loss
\[ t(\hat{\phi}, \phi) = \max(0, \phi - (\hat{\phi} - \theta))^p, \quad p = 1 \text{ or } 2 \]
- Express question difficulty and user skill on the same scale
- If estimation is consistent with assumption, the loss is zero
- Otherwise, the loss is proportional to the violation

Assumption II: smoothness assumption
- Questions close to each other in textual descriptions have similar difficulty

For assumption II: manifold regularization
- If textual descriptions are similar, difficulty gap will be small
- Can choose a variety of term weighting schemas
- Can choose a variety of similarity measures

Introduction

Evaluation for Resolved Questions

Results
- RCM preforms significant better on both datasets
- Improvements can be achieved by a variety of term weighting schemas and similarity measures
- Improvements on SO/Math are greater than those on SO/CPP

Evaluation for Cold-Start Questions

Procedures
- Select k well-resolved questions closest in textual descriptions as nearest neighbors
- Calculate average difficulty of nearest neighbors

Results
- RCM performs consistently better on both datasets with different k values

Difficulty Levels of Words

Procedures
- Split questions into buckets according to their difficulty
- Calculate the frequency of a word in each bucket

Results
- RCM might provide an automatic way to measure difficulty levels of words

Experimental Settings
-Datasets
  - SO/Math: 10528 questions and 6564 users
  - SO/CPP: 10164 questions and 14884 users
-For evaluation
  - 539 annotated SO/Math question pairs
  - 521 annotated SO/CPP question pairs
-Development/test/cold-start split

Baselines
- TrueSkill (TS), PageRank (PR), Competition Model (CM)

Evaluation metric
- Accuracy: proportion of question pairs that are correctly judged

Experiments

Evaluation for Resolved Questions

<table>
<thead>
<tr>
<th>Datasets</th>
<th>PR</th>
<th>TS</th>
<th>CM</th>
<th>RCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO/CPP</td>
<td>H</td>
<td>Q</td>
<td>H</td>
<td>Q</td>
</tr>
<tr>
<td>SO/Math</td>
<td>0.5876</td>
<td>0.6134</td>
<td>0.6480</td>
<td>0.6753</td>
</tr>
<tr>
<td></td>
<td>0.6067</td>
<td>0.6109</td>
<td>0.6227</td>
<td>0.6820</td>
</tr>
</tbody>
</table>

Evaluation for Cold-Start Questions

<table>
<thead>
<tr>
<th>Datasets</th>
<th>PR</th>
<th>TS</th>
<th>CM</th>
<th>RCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO/CPP</td>
<td>H</td>
<td>Q</td>
<td>H</td>
<td>Q</td>
</tr>
<tr>
<td>SO/Math</td>
<td>0.5870</td>
<td>0.5413</td>
<td>0.6120</td>
<td>0.6304</td>
</tr>
<tr>
<td></td>
<td>0.6411</td>
<td>0.6305</td>
<td>0.6653</td>
<td>0.7265</td>
</tr>
</tbody>
</table>