4-27-2018

Quantifying Error in Recommender System Evaluations

Mucun Tian
Boise State University
Quantifying Error in Recommender System Evaluations

**GOAL**
- Quantify and mitigate offline evaluation error in recommender systems.

**RESEARCH QUESTIONS**
- How often does this happen?
- What is the impact of this error case on our evaluation results?
- Simulations allow us quantify the evaluation error in a controlled environment.

**SIMULATION ARCHITECTURE**
- User-Item Simulator: Generating the complete ground-truth data about user preference (Uniform generator and Indian Buffet Process).
- Observable Data Sampler: Sampling the ground-truth data to produce a simulated user consumption data (Uniform and Popular Sampler).
- Simulated Recommenders: Random, Most-Popular and Oracle recommenders
- Evaluator: Computing evaluation metrics using both the ground-truth data and the observed data.

<table>
<thead>
<tr>
<th>Test Data</th>
<th>Movie Recommender</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>???</td>
<td>1. Zootopia</td>
<td>X</td>
</tr>
<tr>
<td>Thumbs Up</td>
<td>2. The Iron Giant</td>
<td>✓</td>
</tr>
<tr>
<td>Thumbs Up</td>
<td>3. Frozen</td>
<td>✓</td>
</tr>
<tr>
<td>???</td>
<td>4. Seven</td>
<td>X</td>
</tr>
<tr>
<td>???</td>
<td>5. Tangled</td>
<td>X</td>
</tr>
</tbody>
</table>

**PROBLEMS**
- If the user would like Zootopia but has not yet seen it, this would be a very good recommender. But the evaluation penalizes it.
- The recommender’s job is to find this kind of items, and the evaluation should account for this.

**PRELIMINARY RESULTS**
- Early Results show a strong skew of the error distribution.

---

Mucun Tian & Michael Ekstrand
People and Information Research Team
Department of Computer Science