April 2018

Quantifying Error in Recommender System Evaluations

Mucun Tian
Boise State University

Quantifying Error in Recommender System Evaluations

Abstract

Researchers and practitioners often use information retrieval and machine learning metrics (the better the recommendation algorithm can predict users' historical consumption data, the higher the metric score is) to evaluate candidate recommendation algorithms in offline experiments. However, these metrics are biased towards recommenders that favor popular items and provide safe, known recommendations, and they incorrectly assume that items users have not consumed are not relevant. To measure the extent of these problems, we are conducting simulation experiments to quantify the metric error of offline evaluation and investigate the impact of data set parameters on the recommender metric error.

Quantifying Error in Recommender System Evaluations

GOAL

- Quantify and mitigate offline evaluation error in recommender systems.

Test Data	Movie Recommender	Evaluation
???	1. Zootopia	✗
👍	2. The Iron Giant	✓
👍	3. Frozen	✓
👎	4. Seven	✗
???	5. Tangled	✗

PROBLEMS

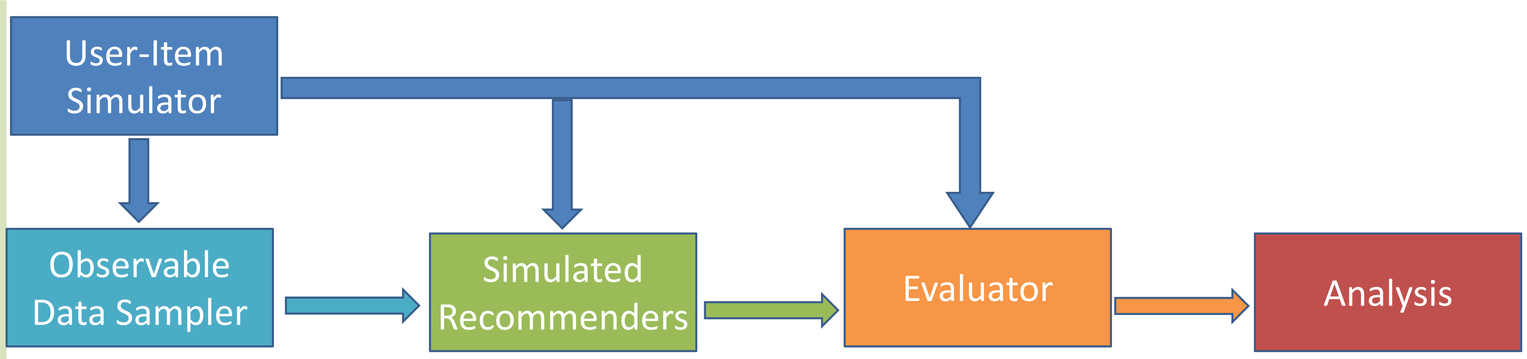
- If the user would like Zootopia but has not yet seen it, this would be a very good recommender. But the evaluation penalizes it.
- The recommender's job is to find this kind of items, and the evaluation should account for this.

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 ARCHITETURE

- 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.



PRELIMINARY RESULTS

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

BOISE STATE UNIVERSITY
COLLEGE OF ENGINEERING
Department of Computer Science

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

People and Information Research Team