Experiment Design for Data Science

Group 49

Kartik Arya, Abbas Manzar, Nurul Alam, Raja Wajahatali, Diana Safarpour

Paper 10: Measuring Item Global Residual Value for Fair Recommendation

Need of GRV & TaFR

Why Do We Need GRV?

- Recommendation systems often neglect item-side fairness, focusing only on user preferences.
- Older items dominate due to the "Snowball Effect," creating unfair exposure, while newer items face visibility challenges (cold-start problem).
- **GRV (Global Residual Value):** A timeliness score that ensures fairer exposure for newer content while balancing accuracy.

Plan to Implement GRV

Enhance backbone models (e.g., NeuMF, GRU4Rec, TiSASRec) by:

- Calculating GRV scores from interaction data.
- Re-ranking items using GRV.
- Validating on MIND and Kuai datasets

Role of TaFR

- TaFR integrates GRV with recommendation models to balance fairness and accuracy.
- It improves visibility for newer items while keeping recommendations relevant and equitable.

Experiment workflow

1. **Understanding the Paper**
   → Analyze methodology, datasets, and experimental setup.
   → Identify key components: GRV module, TaFR framework, evaluation metrics.

2. **Reproducing Experimental Setup**
   → Recreate preprocessing pipelines for MIND and KuaiRec datasets.
   → Implement GRV module and integrate it with backbone models (NeuMF, GRU4Rec, TiSASRec).

3. **Analyzing and Validate Results**
   → Compare reproduced metrics (HR@k, NDCG@k, N_Cov@k, Cov@k) with original findings.
   → Conduct statistical tests to confirm significance.

4. **Documentation**
   → Summarize findings, challenges, and next steps.

What’s done so far?

Milestone 1: Establish the claim that pre-existing recommendation systems are unfair.

Progress

1. Dataset Preparation:
   a. Processed the MIND dataset (behaviors.tsv and news.tsv) to extract user interactions and item information.
   b. Mapped news items to their upload times for timeliness analysis.

2. Initial Analysis:
   a. Began exploring the exposure distribution of news items uploaded at different times.
   b. Grouped items based on upload time to assess exposure trends and started measuring performance metrics.

3. Challenges:
   a. GRV score integration with NeuMF is in progress.
   b. Timeliness analysis is partially complete, with further refinement needed to compare results with snowball effect insights.

Next Steps

- Finalize GRV calculations and validate initial results to identify unfairness trends.
- Expand experiments to cover diverse timeliness characteristics of items.

Encountered difficulties and key findings

We implemented NeuMF (Neural Collaborative Filtering), one of the framework backbone algorithms suggested in the Experiment Settings.

- While we obtained metrics like MAP, NDCG, Precision, and Recall, we are still working on integrating GRV score calculations.
- The GRV score will help analyze how quickly news items become irrelevant based on their posting time, confirming the snowball effect.
- This comparison will also clarify the need to integrate the TaFR model into existing recommender systems.
- Complete results are pending due to limited expertise in recommendation systems. Once GRV scores are derived for one algorithm, applying TaFR and testing others will be faster.

Comments on the datasets:

1. **MIND**: Straightforward. ‘behaviors.tsv’ and ‘news.tsv’ provide user interaction and news details, while .vec files contain entity and relation embeddings (from WikiData KG). We are using MIND_small (6 days of logs) due to high processing time, compared to the full dataset (~7 days in the paper).
2. **kuaiRec**: Diverse but unclear which parts to use. It may be excluded unless TaFR needs additional experimentation.

Note: Reproducing results with MIND should suffice to test GRV relevance. KuaiRec will only be considered if further validation is needed.