UNIVERSITÄT KLAGENFURT

A Worrying Reproducibility Study of Intent-Aware Recommendation Models

Faisal Shehzad, Maurizio Ferrari Dacrema, Dietmar Jannach
Emails: Faisal.Shehzad@aau.at, Maurizio.Ferrari@polimi.it, Dietmar.Jannach@aau.at

Introduction and Motivation

Recommender systems

Traditional recommender systems

Intent aware recommender systems (IARS)

- **Intent aware recommender systems**
  1. Analyze the user's current needs or situational context to provide more relevant recommendations
  2. Developed for both sequential and top-N recommender systems [1, 2, 3]
  3. Examples: browsing vs. ready to buy on e-commerce site, exploring destinations vs. ready to book, etc., [4]
- **From a technical point of view**
  1. Most IARS are based on neural models [1]
  2. Key neural architectures: Graph neural networks (GNN), Graph contrastive learning, Attention mechanism, etc.

Introduction and Motivation

- Due to substantial computational complexity and carbon footprint, one can imagine the superiority of such complex models over simpler models
  a. However, the literature work presents a different picture
  b. Studies [5, 6] demonstrate the superiority of simpler models over state-of-the-art top-N recommenders like NeuMF and ConvMF
  c. Another recent study [7] shows similar results for GNN-based session-based recommender systems
  d. Another study [8] shows shared artifacts are insufficient to reproduce reported results in a paper
- Various factors [5, 7] contribute to this "virtual progress", such as comparing only one family of algorithms (neural models), poorly tuned baseline models, etc.

Research Questions

- Although we do not expect this pattern to reoccur in the emerging and promising area of IARS, we are still proposing two questions to thoroughly evaluate it

  a. Is it possible to reproduce the reported results of published papers using the provided artifacts, such as the source code, datasets, shared hyperparameter values, and so on?

  b. Could simple models, such as kNN, outperform IARS under the same configuration used in a published paper?

Research Methodology

- **Identification of papers**
  1. Queried Google Scholar and IEEE Xplore for papers containing the terms ‘intent’ or ‘intent awareness’ along with the keywords ‘recommend’
  2. Identified 88 papers and retained those proposing a top-N model, published in the last four years in A* venues or top journals
  3. Thirteen papers met this criteria
- **Baselines:** Chose six baseline models, which showed good performance in the previous reproducibility studies [6, 7, 8]
- **Evaluation methodology:** Simpler models are compared with IARS under the same configuration used in the published papers [6]

Results (1/3)

- RQ1: Is it possible to reproduce the reported results of published papers using the provided artifacts, such as the source code, datasets, shared hyperparameter values, and so on?

- Here are the findings our study
  a. Practice of sharing reproducibility packages: 7 out of 13 (~53%)
  b. Reproducibility packages in working condition: 5 out of 13 (~38%)
  c. Reproducibility packages with reproducible results: 2 out of 13 (~15%)

Results (2/3)

- RQ2: Could simple models, such as kNN, outperform IARS under the same configuration used in a published paper?

- Here are the key findings of the experiments
  a. On all accuracy measures, the simpler models outperform all IARS
  b. In one case, the KGIN model outperforms the simpler models on Recall@20 for the Last.fm dataset. However, in this case, we observe a severe data leakage issue in the shared train-test splits
  c. For the DCCF and BIGCF models, the authors used an unusual recommendation list length, such as 40, without providing a valid reason
  d. We found no single winner among the simpler models. However, ItemKNN and $RP^3\beta$ demonstrate competitive performance

Results (3/3)

- Additional observations regarding the training time (T-time) and prediction time (P-time) of the IARS and simpler models?

- The T-time and P-time of KGIN and the simpler models are measured using the largest Alibaba iFashion dataset
- KGIN takes 1 day and 8 hours per iteration, with several iterations required for hyperparameter tuning
- The Alibaba iFashion dataset is significantly smaller than the Netflix Prize dataset (100M), which was released 15 years ago
- ItemKNN, which performs best on this dataset, requires only 2 minutes to build the lookup tables
- In terms of P-time, KGIN model is also 50% slower than ItemkNN

Alibaba iFashion dataset	
Users	114, 000
Items	30,000
Interactions	1.7M

Conclusions

- Reproducibility challenges in recommender systems keep returning, even after fifteen years
  1. Many existing studies show that general level of reproducibility is low
  2. None of examined papers share the code of baselines, data preprocessing, etc., to ensure full reproducibility
  3. Sometimes, authors are unresponsive when approached for guidance
  4. Worryingly, the state of reproducibility is alarming in top-ranked venues
  5. Lastly, we are not against research on complex algorithms. Many studies have demonstrated their effectiveness in recommender systems, computer vision, and natural language processing

UNIVERSITÄT KLAGENFURT

Thank you for your attention

SCAN ME to access the source code

UNIVERSITÄT KLAGENFURT

References

1. Jannach, Dietmar, and Markus Zanker. “A Survey on Intent-aware Recommender Systems.” ACM Transactions on Recommender Systems 3.2 (2024): 1-32.
2. Wang, Xiang, et al. “Disentangled Graph Collaborative Filtering.” Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.
3. Jin, Di, et al. “Dual Intent Enhanced Graph Neural Network for Session-based New Item Recommendation.” Proceedings of the ACM Web Conference 2023. 2023.
4. Loepp, Benedikt. “Multi-list Interfaces for Recommender Systems: Survey and Future Directions.” Frontiers in Big Data 6 (2023): 1239705.
5. Anelli, Vito Walter, et al. “Top-N Recommendation Algorithms: A Quest for the State-of-the-art.” Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization. 2022.
6. Dacrema, Maurizio Ferrari, Michael Benigni, and Nicola Ferro. “Reproducibility and Artifact Consistency of the SIGIR 2022 Recommender Systems Papers based on Message Passing.” arXiv preprint arXiv:2503.07823 (2025).
7. Shehzad, Faisal, and Dietmar Jannach. “Performance Comparison of Session-based Recommendation Algorithms based on GNNs.” European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2024.
8. Shehzad, Faisal, and Dietmar Jannach. “Everyone’s a Winner! on Hyperparameter Tuning of Recommendation Models.” Proceedings of the 17th ACM Conference on Recommender Systems. 2023.
9. Wang, Xiang, et al. "Disentangled Graph Collaborative Filtering." Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.
10. Wang, Xiang, et al. “Learning Intents behind Interactions with Knowledge Graph for Recommendation.” Proceedings of the Web Conference 2021.
11. Qian, Tieyun, et al. “Intent Disentanglement and Feature Self-supervision for Novel Recommendation.” IEEE Transactions on Knowledge and Data Engineering 35.10 (2022): 9864-9877.
12. Ren, Xubin, et al. “Disentangled Contrastive Collaborative Filtering.” Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.
13. Zhang, Yi, Lei Sang, and Yiwen Zhang. “Exploring the Individuality and Collectivity of Intents behind Interactions for Graph Collaborative Filtering.” Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.