Finite Rank-Biased Overlap (FRBO): A New Measure for Stability in Sequential Recommender Systems (Extended Abstract)*

Filippo Betello[1], Federico Siciliano[1], Pushkar Mishra[2] and Fabrizio Silvestri[1]

[1]Sapienza University of Rome, Rome, Italy
[2]AI at Meta, London, UK

Abstract

Sequential Recommender Systems (SRSs) are widely used to model user behavior over time but they often face a critical challenge: they can fail when faced with perturbations in their training data. While the conventional Rank-Biased Overlap (RBO) measure is widely used, it does not properly address this issue, especially when dealing with finite rankings. To fill this gap, we introduce the Finite Rank-Biased Overlap (FRBO) measure. We study the impact of removing elements at the beginning, in the middle, and at the end of the sequence: the latter removal has a negative impact on performance of up to 60% in NDCG. Surprisingly, removing elements from the beginning or middle of sequences has minimal impact on performance. These results shed light on the crucial role of element positioning within the training data and highlight the urgent need for improved robustness in SRSs. We make available our code implementation[1] for FRBO and invite further exploration and adoption by the research community.

Keywords

Recommender Systems, Evaluation of Recommender Systems, Model Stability, Input Data Perturbation

1. Introduction

Recommender systems are now ubiquitous, crucial for helping users navigate the vast online information [2, 3, 4]. Despite their success, the robustness of SRSs against training data perturbations remains an open research question [5, 6]. In real-world scenarios, users may employ different services for similar purposes, leading to fragmented data between competitors. Providers must train robust recommender systems with this incomplete data. Previous assessments [7] use Rank-Biased Overlap (RBO) [8], designed for infinite lists and fail to converge to 1 when applied to finite-length lists. Therefore, we propose the Finite Rank-Biased Overlap (FRBO)[1] measure to address this limitation. We empirically analyze the effects of removing items from user interaction sequences on SRS performance. Our results indicate that removing the most recent items in user interaction sequences leads to a significant decrease in recommendation accuracy.

[1]https://github.com/siciliano-diag/finite_rank_biased_rbo.git

IIR2024: The 14th Italian Information Retrieval Workshop, 5th - 6th September 2024, Udine, Italy

*This work is an extended abstract based on the publication "Investigating the Robustness of Sequential Recommender Systems Against Training Data Perturbations" at the 46th European Conference on Information Retrieval [1].

betello@diag.uniroma1.it (F. Betello); siciliano@diag.uniroma1.it (F. Siciliano); pushkarmishra@meta.co (P. Mishra); fsilvestri@diag.uniroma1.it (F. Silvestri)
0009-0006-0945-9688 (F. Betello); 0000-0003-1339-6983 (F. Siciliano); 0000-0002-1653-6198 (P. Mishra); 0000-0001-7669-9055 (F. Silvestri)

2. Methodology

2.1. Setting

In Sequential Recommendation, each user u is represented by a sequence of items $S_u = (I_1, I_2, ..., I_j, ..., I_{L_u-1}, I_{L_u})$ with which they have interacted, where L_u is the sequence length. We investigate this type of removal: **Beginning**: $S_u = (I_{n+1}, \ldots, I_{L_u-1})$, **Middle**: $S_u = (I_1, \ldots, I_{\lfloor \frac{L_u-1-n}{2} \rfloor}, I_{\lfloor \frac{L_u-1+n}{2} \rfloor} ..., I_{L_u-1})$, **End**: $S_u = (I_1, \ldots, I_{L_u-1-N})$, with $L_u \leq 10$.

2.2. Metrics

To evaluate model performance, we use traditional metrics: Precision, Recall, MRR and NDCG. For stability assessment, we employ the Rank List Sensitivity (RLS) [7], which compares two ranking lists $\mathcal{X}$ and $\mathcal{Y}$, derived from the model trained under standard and perturbed conditions, respectively. The RLS measure is defined as: $\mathbf{RLS} = \frac{1}{|\mathcal{X}|} \sum_{k=1}^{|\mathcal{X}|} \text{sim}(R^{X_k}, R^{Y_k})$, where X_k and Y_k represent the k-th ranking inside $\mathcal{X}$ and $\mathcal{Y}$ respectively. The similarity measure sim can be either Jaccard Similarity (JAC) [9] or Rank-Biased Overlap (RBO) [8].

$$\mathbf{JAC(X, Y)} = \frac{|X \cap Y|}{|X \cup Y|} \quad \mathbf{RBO(X, Y)} = (1-p) \sum_{d=1}^{+\infty} p^{d-1} \frac{|X[1:d] \cap Y[1:d]|}{d} \tag{1}$$

In recommendation systems, metrics are often computed using finite-length rankings, indicated by appending "@k" to the metric name, like NDCG@k. Traditional metrics (e.g. NDCG, MRR) adapt well to this format, but RLS does not when using RBO due to a key limitation: it does not converge to one for identical finite-length lists. To address this, we introduce Finite Rank-Biased Overlap FRBO@k, designed to ensure a convergence value of 1 for identical lists and 0 for completely dissimilar lists.

Theorem 1. *Given a set of items $I = \{I_1, ..., I_{N_I}\}$, two rankings $X = (x_1, ..., x_k)$ and $Y = (y_1, ..., y_k)$, such that $x_i, y_i \in I$, and $k \in \mathbb{N}^+$*

$$\mathbf{FRBO(X, Y)@k} = \frac{\text{RBO(X, Y)@k} - min_{X,Y}\text{RBO@k}}{max_{X,Y}\text{RBO@k} - min_{X,Y}\text{RBO@k}} \tag{2}$$

$$\min_{X,Y} \text{FRBO(X, Y)@k} = 0, \ \max_{X,Y} \text{FRBO(X, Y)@k} = 1$$

Finding the minimum and maximum values of RBO is crucial for normalizing it when summing up to the top k items of the ranking. We also need to show that these values are not necessarily limited to 0 and 1.

Lemma 1. *Given a set of items $I = \{I_1, ..., I_{N_I}\}$, two rankings $X = (x_1, ..., x_k)$ and $Y = (y_1, ..., y_k)$, such that $x_i, y_i \in I$, and $k \in \mathbb{N}^+$, the following holds:*

$$\min_{X,Y} \text{RBO@k} = \begin{cases} 0, & \text{if } k \leq \lfloor \frac{N_I}{2} \rfloor \\ (1-p)\left(2\frac{p^{\lfloor \frac{N_I}{2} \rfloor} - p^{N_I}}{1-p} - N_I \ell\right) & \text{otherwise} \end{cases} \tag{3}$$

$$\textit{where } \mathrm{RBO(X,Y)@k} = (1-p)\sum_{d=1}^{k} p^{d-1}\frac{|X[1:d]\cap Y[1:d]|}{d}$$

$$\textit{and } \ell = p^{\left\lfloor \frac{N_I}{2} \right\rfloor}\Phi(p,1,\lfloor\frac{N_I}{2}\rfloor+1) - p^{N_I}\Phi(p,1,N_I+1)$$

Lemma 2. *Given a set of items $I = \{I_1, ..., I_{N_I}\}$, two rankings $X = (x_1, ..., x_k)$ and $Y = (y_1, ..., y_k)$, such that $x_i, y_i \in I$ and $k \in \mathbb{N}^+$, the following holds:* $\max_{X,Y} \mathrm{RBO@k} = 1 - p^k$.

2.3. Experimental setup

We use four different datasets: MovieLens (1M and 100K versions) [10] and Foursquare (New York City and Tokyo) [11]; we use two different architectures to validate the results: SASRec [12] and GRU4Rec [13]. The RecBole library [14] was utilized for conducting all the experiments, encompassing data preprocessing, model configuration, training, and testing. The code required to replicate the experiments is accessible in our GitHub repository[1].

3. Results

3.1. Intrinsic Models Instability (RQ1)

The inherent resilience of the models when using alternative starting seeds is shown in the Baseline row in Tab. 1. In general, the deviation is almost always less than 1%. However, the RLS shows a large deviation from the optimal value of 1, indicating quite different ranks. The aggregated results suggest that the models achieve a sufficient level of performance regardless of the initialization seed, but the generated rankings are significantly affected by it.

3.2. Comparison of the position of removal (RQ2)

The performance and stability of keeping all elements in the training set with a constant initialization seed versus removing 10 elements are compared in Tab. 1. It can be seen that the performance of the model is not significantly affected by removing items from the beginning or the middle of the sequence. Rather, we can see how drastically the metrics are reduced by removing items from the end of the sequence: for example, when SASRec is applied to the MovieLens 1M dataset, the NDCG decreases by more than 50%. In addition, the Jaccard similarity and FRBO values approach 0, indicating that very few items are shared by the generated ranks.

3.3. Effect of the number of elements removed (RQ3)

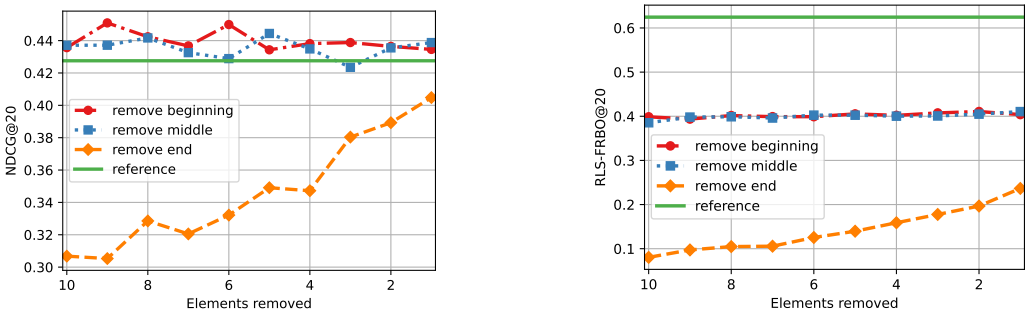
As seen before, removing items at the beginning or in the middle of the sequence has a negligible impact on performance, as confirmed by Fig. 1. However, removing elements from the end of the sequence leads to a noticeable decrease in the metrics, proportional to the number of elements removed. This effect is consistent across both models tested and across all datasets evaluated. In particular, Fig. 1 illustrate how the SASRec model performs on the MovieLens datasets under

[1] https://github.com/siciliano-diag/finite_rank_biased_rbo.git

Table 1
Variations in metrics for ten-item removal in SASRec on ML-1M and GRU4Rec on FS-NYC, with baseline percentage variations from two initialization seeds, highlighting less robust models in bold and statistically significant results indicated by †.

Model	Removal	Prec.	Recall	MRR	NDCG	FRBO	JAC
SASRec ML-1M	Baseline	0.3%	0.4%	0.5%	0.5%	.549	.569
	Beginning	-0.23%	-0.23%	-0.15%	-0.07%	.399†	.368†
	Middle	-0.35%	-0.29%	-1.09%	-0.73%	.385†	.356†
	End	**-15.5%**†	**-15.3%**†	**-45.9%**†	**-56.0%**†	**.080**†	**.106**†
GRU4Rec FS-NYC	Baseline	0.1%	0.1%	0.6%	0.1%	.110	.083
	Beginning	-0.23%	-0.23%	-1.47%	-0.94%	.105†	.075†
	Middle	-0.93%	-0.93%	-0.18%	-0.44%	.110†	.074†
	End	**-4.92%**†	**-4.86%**†	**-8.42%**†	**-7.39%**†	**.089**†	**.062**†

these conditions, highlighting in particular a significant decrease in metrics. Interestingly, even removals from the beginning and middle of sequences in ML-1M show a significant decrease in RLS-FRBO, suggesting considerable variation in the rankings generated despite overall stable performance levels, likely influenced by the large user base and interaction volume of the dataset.



(a) NDCG@20 SASRec ML-100k

(b) FRBO@20 SASRec ML-1M

Figure 1: Plots of NDCG and FRBO for SASRec on the ML datasets illustrate the baseline as a solid line and show the variations of the metrics with changing item removal across three scenarios as dashed lines.

4. Conclusion

This study investigates the effect of item position within a temporally ordered sequence in SRSs. First, it introduces Finite RBO, a variant of RBO tailored for finite lists, which has been shown to normalize within the [0,1] range. Second, it shows that removing items at the end of the sequence significantly affects all performances, while removing items at the beginning or in the middle of the sequence has a less pronounced effect. Future research aims to extend these results to more models and datasets, and to explore strategies to improve model robustness to missing training data possibly through different training approaches, robust loss functions [15], or different optimization goals [6].

Acknowledgments

This work was partially supported by projects FAIR (PE0000013) and SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. This work has also been supported by the NEREO (Neural Reasoning over Open Data) project funded by the Italian Ministry of Education and Research (PRIN) Grant no. 2022AEFHA.

References

[1] F. Betello, F. Siciliano, P. Mishra, F. Silvestri, Investigating the robustness of sequential recommender systems against training data perturbations, in: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2024, p. 205–220. URL: https://doi.org/10.1007/978-3-031-56060-6_14. doi:10.1007/978-3-031-56060-6_14.

[2] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering 17 (2005) 734–749. doi:10.1109/TKDE.2005.99.

[3] A. Purificato, G. Cassarà, F. Siciliano, P. Liò, F. Silvestri, Sheaf4rec: Sheaf neural networks for graph-based recommender systems, 2023.

[4] F. Betello, A. Purificato, F. Siciliano, G. Trappolini, A. Bacciu, N. Tonellotto, F. Silvestri, A reproducible analysis of sequential recommender systems, arXiv preprint arXiv:2408.03873 (2024).

[5] Y. Li, H. Chen, Z. Fu, Y. Ge, Y. Zhang, User-oriented fairness in recommendation, in: Proceedings of the Web Conference 2021, WWW '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 624–632. URL: https://doi.org/10.1145/3442381.3449866. doi:10.1145/3442381.3449866.

[6] A. Bacciu, F. Siciliano, N. Tonellotto, F. Silvestri, Integrating item relevance in training loss for sequential recommender systems, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1114–1119. URL: https://doi.org/10.1145/3604915.3610643. doi:10.1145/3604915.3610643.

[7] S. Oh, B. Ustun, J. McAuley, S. Kumar, Rank list sensitivity of recommender systems to interaction perturbations, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1584–1594. URL: https://doi.org/10.1145/3511808.3557425. doi:10.1145/3511808.3557425.

[8] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Trans. Inf. Syst. 28 (2010). URL: https://doi.org/10.1145/1852102.1852106. doi:10.1145/1852102.1852106.

[9] P. Jaccard, The distribution of the flora in the alpine zone. 1, New Phytologist 11 (1912) 37–50.

[10] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2015). URL: https://doi.org/10.1145/2827872. doi:10.1145/2827872.
[11] D. Yang, D. Zhang, V. W. Zheng, Z. Yu, Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns, IEEE Transactions on Systems, Man, and Cybernetics: Systems 45 (2014) 129–142.
[12] W.-C. Kang, J. McAuley, Self-attentive sequential recommendation, in: 2018 IEEE International Conference on Data Mining (ICDM), IEEE, 2018, pp. 197–206.
[13] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, 2016. arXiv:1511.06939.
[14] W. X. Zhao, S. Mu, Y. Hou, Z. Lin, Y. Chen, X. Pan, K. Li, Y. Lu, H. Wang, C. Tian, Y. Min, Z. Feng, X. Fan, X. Chen, P. Wang, W. Ji, Y. Li, X. Wang, J.-R. Wen, Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms, 2021. arXiv:2011.01731.
[15] M. S. Bucarelli, L. Cassano, F. Siciliano, A. Mantrach, F. Silvestri, Leveraging inter-rater agreement for classification in the presence of noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3439–3448.