

Replication of Recommender Systems with Impressions

Discussion Paper

Fernando B. Pérez Maurera^{1,2,*}, Maurizio Ferrari Dacrema¹ and Paolo Cremonesi¹

¹Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

²ContentWise, Via Simone Schiaffino 11, Milano, 20158, Milano, Italy

Abstract

Impressions are a novel data type in Recommender Systems containing the previously-exposed items, i.e., what was shown on-screen. Due to their novelty, the current literature lacks a characterization of impressions, and replications of previous experiments. Also, previous research works have mainly used impressions in industrial contexts or recommender systems competitions, such as the ACM RecSys Challenges. This work is part of an ongoing study about impressions in recommender systems. It presents an evaluation of impressions recommenders on current open datasets, comparing not only the recommendation quality of impressions recommenders against strong baselines, but also determining if previous progress claims can be replicated.

Keywords

Recommender Systems, Impressions, Exposure, Replication, Collaborative Filtering

1. Introduction


A recurrent and fundamental task in Recommender System (RS) is the empirical evaluation of recommendation models with varied data sources. One particular novel and modestly explored data source in RS research are *impressions*. These contains not only the previous interactions (e.g., purchases and clicks) of users but also the items they were presented with (e.g., recommendations and search results). Previous research works [1, 2, 3, 4] have proposed recommendations models that leverage impressions data, called *impressions recommenders*. To current date, no previous work has tried to replicate these models on open datasets.¹


The replication of previous works is fundamental to measure the current status of recommendation models across different domains and data sources. Previous research works have highlighted the importance of replication works for the RS community [5, 6, 7, 8, 9]. To address this existing gap in the literature, this work presents a replication study of four impressions

IIR2022: 12th Italian Information Retrieval Workshop, June 29 - June 30th, 2022, Milan, Italy

* Corresponding author.

✉ fernandobenjamin.perez@polimi.it (Fernando B. Pérez Maurera); maurizio.ferrari@polimi.it (M. Ferrari Dacrema); paolo.cremonesi@polimi.it (P. Cremonesi)

ORCID  0000-0001-6578-7404 (Fernando B. Pérez Maurera); 0000-0001-7103-2788 (M. Ferrari Dacrema); 0000-0002-1253-8081 (P. Cremonesi)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹The concept of “replicability” is the same as in the ACM Artifact Review and Badging, version 1.1, available online at <https://www.acm.org/publications/policies/artifact-review-and-badging-current>.

recommenders.² First, this work presents a brief categorization of impressions data, as the current literature does not have one. Second, this work empirically evaluates the recommendation quality of several baseline and impressions recommenders on current open-source impressions datasets and compares the obtained results with the claims given in the original works.

2. Impressions in Recommender Systems

Impressions are a novel and modestly used data source that contains the items shown on-screen to users, e.g., the items that users were presented when browsing an e-commerce service. Similar to interactions data in RS, an impression is characterized as an user-item pair (u, i) , indicating that user u has been impressed with item i . Importantly, previous research works with impressions have been in the context of industrial settings or RS competitions. Hence, progress in impressions research has been mostly slow. The following presents a brief categorization of impressions:

Signals: The signals within impressions are mixed, i.e., impressions may reflect both positive and negative users preferences toward items, mostly depending on the provenance of the impressions, e.g., a recommender system or business rules. There is no consensus in the current literature regarding the meaning of impressions. For instance, in the same context, previous research works have used impressions as positive [10] or negative [11] signals.

Challenges: Three main considerations should be taken into account when working with impressions data. First, the heterogeneous signals within impressions. Second, scalability as the number of impressions records might be orders of magnitude greater than interactions. Third, the effects of feedback loops between users and recommendation systems.

Impressions Recommenders: Two types of impressions recommenders have been proposed in previous research works: re-ranking and impressions as user profiles recommenders. The first group re-scores the preference scores of an existing recommendation model based on impressions data and features extracted from impressions [3, 1, 12, 13]. The second group expands the user profiles (interactions) with impressions data [14].

Impressions Datasets: Three datasets from different recommendation domains are open-source and can be used in research activities: CONTENTWISE IMPRESSIONS (TV and movies), MIND (news), and FINN.NO SLATES (e-commerce). Private and non-distributable datasets also exist and have been used in previous works [1, 12, 13, 15, 16]. However, due to their nature or license agreements, it is not possible to use them in newer research works.

Evaluation of Impressions: No evaluation and comparison of impressions recommenders on open datasets exists in the current literature. Currently, research works with impressions have worked on two contexts: recommendation challenges [14, 17, 18, 11] or industrial scenarios [13, 1, 19, 12]. In the former, complex recommendation models are built and tested against a specific dataset without assessing the generalization aspects of impressions on other areas or domains.

²This work is part of an ongoing study about impressions in recommender systems

In the latter, impressions are studied on private data and recommendation systems. No previous work have performed ablation studies to assess the impact of impressions.

3. Experimental Methodology

This work presents several experiments on impressions recommenders, particularly, when used as a *plug-in* to existing recommendation models, i.e., impressions recommenders alter the preference scores of recommendation models. The goal of these experiments is two-fold. First, to determine the recommendation quality of impressions recommenders on open-source impressions datasets. Second, to replicate, if possible, the progress achieved by impressions recommenders in their original works. The experiments followed the following experimental methodology:

Datasets, Processing, and Splits: The three available open-source datasets with impressions were used in the experiments: CONTENTWISE IMPRESSIONS, MIND, and FINN.NO SLATES. The following processing was applied to all datasets: (i) data records were sorted in ascending order by their time attribute; (ii) duplicated user-item interactions were aggregated into a single one, keeping the data of the first interaction; (iii) interactions and impressions of users without a minimum of three interactions were removed; (iv) the training, validation, and testing splits were created following a traditional leave-last-interaction out.

Evaluation: All recommenders were evaluated on traditional accuracy and beyond-accuracy metrics [5] in the standard top-N recommendation scenario. Hyper-parameters were searched using bayesian search with 16 random cases, 50 total cases, and optimizing NDCG [5] on the validation set.

Baseline Recommenders: Neighborhood-based (ITEM KNN and USER KNN) [5], graph-based (RP_{β}^3) [20], auto-encoders (SLIM ELASTICNET [21] and EASE R [22]), machine learning (PURESVD [23] and MF BPR [24]), and factorization machines recommenders (LIGHT FM) [25]. The description of these recommenders, their hyper-parameters, and their ranges is found in [5].

Impressions Recommenders: Re-ranking (CYCLING [3] and IMPRESSIONS DISCOUNTING [1]), and impressions as user profiles recommenders (ITEM WEIGHTED PROFILES and USER WEIGHTED PROFILES) [14].³

4. Results and Discussion

The accuracy and beyond accuracy of impressions recommenders varied by dataset, baseline, and impressions recommender. All impressions recommenders achieved higher NDCG than baselines on the FINN.NO SLATES dataset. On other datasets, impressions recommenders achieved slightly higher NDCG than baseline recommenders in some cases. Such cases are shown in Table 1. This shows the NDCG of the base and impressions recommenders on the MIND dataset.⁴

³Due to space limitations, this work omits the list of hyper-parameters of impressions recommenders.

⁴Recommenders were evaluated on more metrics. Due to space limitations Table 1 only contains the results on NDCG.

Table 1

Top-20 ranking accuracy measured with NDCG of base and impressions recommenders on the MIND dataset. MF BPR, NMF, and PURESVD are folded recommenders [23]. Values in **bold** mean higher accuracy than **Baseline**. **ID** refers to impressions discounting using the frequency of impressions. **IUP** refers to impressions as user profiles. **x** means the case was not explored due to incompatibility of the base recommender and the impressions recommender. - means explored cases yielded the same results.

	Baseline	Cycling	ID	IUP
ITEM KNN	0.00868	0.00693	0.00028	0.00012
USER KNN	0.00766	0.01797	0.01118	0.06681
MF BPR	0.00002	0.00680	0.00424	-
NMF	0.00116	0.00797	-	0.00098
PURESVD	0.00010	0.00728	-	0.00015
RP_{β}^3	0.01643	0.00720	0.00015	0.00009
SLIM ELASTICNET	0.01493	0.00699	0.00060	0.00010
LIGHT FM	0.00160	0.00705	0.00101	x

From the table, a notable case is the use of impressions as user profiles (**IUP**) with **USER KNN** on the MIND dataset. Particularly, this case obtained *eight* and *four* times higher NDCG than the base (**USER KNN**) and best (RP_{β}^3) baseline recommender, respectively.

When looking at each impressions recommender, the **CYCLING** recommender achieved higher NDCG on the **FINN.NO SLATES** and **MIND** datasets. Although, on the latter, this only occurred on matrix factorization and factorization machines recommenders. The **IMPRESSIONS DISCOUNTING**, **ITEM WEIGHTED PROFILES**, and **USER WEIGHTED PROFILES** recommenders did not have such consistent results. For instance, the former achieved higher NDCG than **USER KNN** but obtained lower NDCG than **ITEM KNN** on the **MIND**.

Regarding the replicability of impressions recommenders, **CYCLING** recommended less accurate but more diverse items on the **CONTENTWISE IMPRESSIONS** dataset. This result is aligned with the conclusions of [3], which performed experiments on a different dataset of the same domain. For **IMPRESSIONS DISCOUNTING**, only the results on the **FINN.NO SLATES** dataset are aligned with the conclusions of [1]. However, in the reference article, the experimental methodology was on error prediction (RMSE) instead of top-N recommendations. The remaining impressions recommenders could not be replicated due to lack of replicability information.

Regarding to the signals within impressions, the results varied mostly by dataset while the recommenders did not play a major role. For the **CONTENTWISE IMPRESSIONS** dataset, impressions cannot be considered as positive or negative, as substantially higher NDCG was not achieved by any recommender treating impressions as positive or negative signals. For the **MIND** and **FINN.NO SLATES** datasets, impressions were considered as positive signals in most recommenders while at the same time achieving higher NDCG than the base recommender.

References

- [1] P. Lee, L. V. S. Lakshmanan, M. Tiwari, S. Shah, Modeling impression discounting in large-scale recommender systems, in: S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, R. Ghani (Eds.), The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014, ACM, 2014, pp. 1837–1846. URL: <https://doi.org/10.1145/2623330.2623356>. doi:10.1145/2623330.2623356.
- [2] D. Zibriczky, A combination of simple models by forward predictor selection for job recommendation, in: Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016, ACM, 2016, pp. 9:1–9:4. URL: <https://doi.org/10.1145/2987538.2987548>. doi:10.1145/2987538.2987548.
- [3] Q. Zhao, G. Adomavicius, F. M. Harper, M. C. Willemsen, J. A. Konstan, Toward better interactions in recommender systems: Cycling and serpentine approaches for top-n item lists, in: C. P. Lee, S. E. Poltrock, L. Barkhuus, M. Borges, W. A. Kellogg (Eds.), Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017, ACM, 2017, pp. 1444–1453. URL: <https://doi.org/10.1145/2998181.2998211>. doi:10.1145/2998181.2998211.
- [4] M. Aharon, Y. Kaplan, R. Levy, O. Somekh, A. Blanc, N. Eshel, A. Shahar, A. Singer, A. Zlotnik, Soft frequency capping for improved ad click prediction in yahoo gemini native, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019, ACM, 2019, pp. 2793–2801. URL: <https://doi.org/10.1145/3357384.3357801>. doi:10.1145/3357384.3357801.
- [5] M. F. Dacrema, S. Boglio, P. Cremonesi, D. Jannach, A troubling analysis of reproducibility and progress in recommender systems research, *ACM Trans. Inf. Syst.* 39 (2021) 20:1–20:49. URL: <https://doi.org/10.1145/3434185>. doi:10.1145/3434185.
- [6] M. F. Dacrema, P. Cremonesi, D. Jannach, Are we really making much progress? A worrying analysis of recent neural recommendation approaches, in: T. Bogers, A. Said, P. Brusilovsky, D. Tikk (Eds.), Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019, ACM, 2019, pp. 101–109. URL: <https://doi.org/10.1145/3298689.3347058>. doi:10.1145/3298689.3347058.
- [7] J. Lin, The neural hype and comparisons against weak baselines, *SIGIR Forum* 52 (2019) 40–51. doi:10.1145/3308774.3308781.
- [8] J. Lin, The neural hype, justified! a recantation, *SIGIR Forum* 53 (2021) 88–93. doi:10.1145/3458553.3458563.
- [9] W. Yang, K. Lu, P. Yang, J. Lin, Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ACM, 2019, pp. 1129–1132. doi:10.1145/3331184.3331340.
- [10] C. Zhang, X. Cheng, An ensemble method for job recommender systems, in: F. Abel, A. A. Benczúr, D. Kohlsdorf, M. A. Larson, R. Pálovics (Eds.), Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016, ACM, 2016, pp. 2:1–2:4. URL: <https://doi.org/10.1145/2987538.2987545>.

doi:10.1145/2987538.2987545.

- [11] T. D. Pessemier, K. Vanhecke, L. Martens, A scalable, high-performance algorithm for hybrid job recommendations, in: Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016, ACM, 2016, pp. 5:1–5:4. URL: <https://doi.org/10.1145/2987538.2987539>. doi:10.1145/2987538.2987539.
- [12] M. Hristakeva, D. Kershaw, M. Rossetti, P. Knoth, B. Pettit, S. Vargas, K. Jack, Building recommender systems for scholarly information, in: Proceedings of the 1st Workshop on Scholarly Web Mining, SWM@WSDM 2017, Cambridge, United Kingdom, February 10, 2017, ACM, 2017, pp. 25–32. URL: <https://doi.org/10.1145/3057148.3057152>. doi:10.1145/3057148.3057152.
- [13] D. Agarwal, B. Chen, R. Gupta, J. Hartman, Q. He, A. Iyer, S. Kolar, Y. Ma, P. Shivaswamy, A. Singh, L. Zhang, Activity ranking in linkedin feed, in: S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, R. Ghani (Eds.), The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014, ACM, 2014, pp. 1603–1612. URL: <https://doi.org/10.1145/2623330.2623362>. doi:10.1145/2623330.2623362.
- [14] M. Polato, F. Aiolli, A preliminary study on a recommender system for the job recommendation challenge, in: Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016, ACM, 2016, pp. 1:1–1:4. URL: <https://doi.org/10.1145/2987538.2987549>. doi:10.1145/2987538.2987549.
- [15] F. Abel, A. A. Benczúr, D. Kohlsdorf, M. A. Larson, R. Pálovics, Recsys challenge 2016: Job recommendations, in: S. Sen, W. Geyer, J. Freyne, P. Castells (Eds.), Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016, ACM, 2016, pp. 425–426. URL: <https://doi.org/10.1145/2959100.2959207>. doi:10.1145/2959100.2959207.
- [16] P. Knees, Y. Deldjoo, F. B. Moghaddam, J. Adamczak, G.-P. Leyson, P. Monreal, Recsys challenge 2019: Session-based hotel recommendations, in: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 570–571. URL: <https://doi.org/10.1145/3298689.3346974>. doi:10.1145/3298689.3346974.
- [17] E. D'Amico, G. Gabbolini, D. Montesi, M. Moreschini, F. Parroni, F. Piccinini, A. Rossettini, A. R. Introito, C. Bernardis, M. F. Dacrema, Leveraging laziness, browsing-pattern aware stacked models for sequential accommodation learning to rank, in: P. Knees, Y. Deldjoo, F. B. Moghaddam, J. Adamczak, G. P. Leyson, P. Monreal (Eds.), Proceedings of the Workshop on ACM Recommender Systems Challenge, Copenhagen, Denmark, September 2019, ACM, 2019, pp. 7:1–7:5. URL: <https://doi.org/10.1145/3359555.3359563>. doi:10.1145/3359555.3359563.
- [18] J. I. Honrado, O. Huarte, C. Jimenez, S. Ortega, J. R. Pérez-Agüera, J. Pérez-Iglesias, Á. Polo, G. Rodríguez, Jobandtalent at recsys challenge 2016, in: F. Abel, A. A. Benczúr, D. Kohlsdorf, M. A. Larson, R. Pálovics (Eds.), Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016, ACM, 2016, pp. 3:1–3:5. URL: <https://doi.org/10.1145/2987538.2987547>. doi:10.1145/2987538.2987547.
- [19] Q. Zhao, M. C. Willemsen, G. Adomavicius, F. M. Harper, J. A. Konstan, Interpreting user inaction in recommender systems, in: Proceedings of the 12th ACM Conference on

- Recommender Systems, ACM, Vancouver British Columbia Canada, 2018, pp. 40–48. URL: <https://dl.acm.org/doi/10.1145/3240323.3240366>. doi:10.1145/3240323.3240366.
- [20] F. Christoffel, B. Paudel, C. Newell, A. Bernstein, Blockbusters and wallflowers: Accurate, diverse, and scalable recommendations with random walks, in: H. Werthner, M. Zanker, J. Golbeck, G. Semeraro (Eds.), Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16-20, 2015, ACM, 2015, pp. 163–170. URL: <https://doi.org/10.1145/2792838.2800180>. doi:10.1145/2792838.2800180.
- [21] X. Ning, G. Karypis, SLIM: sparse linear methods for top-n recommender systems, in: D. J. Cook, J. Pei, W. Wang, O. R. Zaïane, X. Wu (Eds.), 11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011, IEEE Computer Society, 2011, pp. 497–506. URL: <https://doi.org/10.1109/ICDM.2011.134>. doi:10.1109/ICDM.2011.134.
- [22] H. Steck, Embarrassingly shallow autoencoders for sparse data, in: L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, L. Zia (Eds.), The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, 2019, pp. 3251–3257. URL: <https://doi.org/10.1145/3308558.3313710>. doi:10.1145/3308558.3313710.
- [23] P. Cremonesi, Y. Koren, R. Turrin, Performance of recommender algorithms on top-n recommendation tasks, in: X. Amatriain, M. Torrens, P. Resnick, M. Zanker (Eds.), Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010, ACM, 2010, pp. 39–46. URL: <https://doi.org/10.1145/1864708.1864721>. doi:10.1145/1864708.1864721.
- [24] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: bayesian personalized ranking from implicit feedback, in: J. A. Bilmes, A. Y. Ng (Eds.), UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009, AUAI Press, 2009, pp. 452–461. URL: https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1630&proceeding_id=25.
- [25] M. Kula, Metadata embeddings for user and item cold-start recommendations, in: T. Bogers, M. Koolen (Eds.), Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 16-20, 2015, volume 1448 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015, pp. 14–21. URL: <http://ceur-ws.org/Vol-1448/paper4.pdf>.