Towards the Evaluation of Recommender Systems with Impressions

Fernando B. Pérez Maurera
Politecnico di Milano & ContentWise
Milan, Italy
fernandobenjamin.perez@polimi.it

Maurizio Ferrari Dacrema
Politecnico di Milano
Milan, Italy
maurizio.ferrari@polimi.it

Paolo Cremonesi
Politecnico di Milano
Milan, Italy
paolo.cremonesi@polimi.it

ABSTRACT

In Recommender Systems, impressions are a relatively new type of information that records all products previously shown to the users. They are also a complex source of information, combining the effects of the recommender system that generated them, search results, or business rules that may select specific products for recommendations. The fact that the user interacted with a specific item given a list of recommended ones may benefit from a richer interaction signal, in which some items the user did not interact with may be considered negative interactions. This work presents a preliminary evaluation of recommendation models with impressions. First, impressions are characterized by describing their assumptions, signals, and challenges. Then, an evaluation study with impressions is described. The study's goal is two-fold: to measure the effects of impressions data on properly-tuned recommendation models using current open-source datasets and disentangle the signals within impressions data. Preliminary results suggest that impressions data and signals are nuanced, complex, and effective at improving the recommendation quality of recommenders. This work publishes the source code, datasets, and scripts used in the evaluation to promote reproducibility in the domain.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Collaborative filtering*; • **General and reference** → *Evaluation.*

KEYWORDS

recommender systems, impressions, exposure, evaluation, real-time recommendations

ACM Reference Format:
Fernando B. Pérez Maurera, Maurizio Ferrari Dacrema, and Paolo Cremonesi. 2022. Towards the Evaluation of Recommender Systems with Impressions. In *Sixteenth ACM Conference on Recommender Systems (RecSys '22), September 18–23, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3523227.3551483

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
RecSys '22, September 18–23, 2022, Seattle, WA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9278-5/22/09…$15.00
https://doi.org/10.1145/3523227.3551483

1 INTRODUCTION

Research in Recommender Systems (RS) mainly builds recommendation models using historical feedback (e.g., clicks, purchases, watching actions) of products collected from users, but the community is always looking to improve the recommendation quality by leveraging other data sources such as content, social, and contextual information.

Impressions are an interesting, novel, and modestly explored source of information that is available for researchers and practitioners. Impressions, also called exposure data or past exposures, contain the previous recommendations to users, meaning the items or products displayed on users' screens. These items usually come from the existing recommendation system, a search function, or business rules.

Impressions have only been used to a limited extent in RS research, mainly due to the absence of publicly available datasets. However, this is rapidly changing as more impression datasets have been published and competitions encouraged their use. For instance: (i) interest in impressions research and use has risen [6, 22], (ii) several RS challenges included impressions in their datasets, (iii) industries have presented case studies highlighting the effects of impressions data in their recommendations [4, 19], and (iv) three open datasets were recently-published ContentWise Impressions [26], MIND [32], and FINN.no Slates [10, 11]. Due to the limited use of impressions in RS, there exist several open research questions, e.g., whether to evaluate recommendation models with impressions data requires to develop a specific methodology as is the case in other scenarios [12, 14], characterization of signals and biases in impressions, and challenges regarding the use of impressions. Previous studies have addressed a few of these questions [34]. In contrast, others have presented case studies in private settings or are evaluated in the setting of a challenge [1, 16, 18, 19], in which the analysis has been narrow and specific to the context of the study, also without a discussion of how presented results can be applied to other domains or evaluation scenarios. Consequently, many research questions remain unexplored.

This work aims to raise awareness of the existence and use of impressions in recommender systems. Towards this goal, this work contextualizes impressions and discusses their signals, bias, and challenges. Consequently, this work presents the first evaluation study of impressions recommender under the same evaluation methodology on open-source datasets. The study focuses on real-time recommendations with impressions, specifically, on plug-in recommenders, i.e., those that receive the recommendations generated by another model and use them to build the final recommendation list. The main advantage of plug-in recommenders is that the underlying recommendation model does not need to be

re-trained. The study aims to measure the recommendation quality of plug-in recommenders using impressions data and disentangle the signals within impressions. Under real-time recommendations, the simplicity of impressions models is key to meeting serving time constraints.

2 OVERVIEW OF IMPRESSIONS IN RECOMMENDER SYSTEMS

Open Source Datasets. In recent years, several research works have encouraged the research of impressions recommenders and impressions data by open-sourcing sourced the following datasets with impressions: ContentWise Impressions, MIND, and FINN.no Slates. The **ContentWise Impressions [26]** dataset contains interactions and impressions of users with a media provider of movies and TV series over the internet. Users are the registered accounts with the over-the-top media service while items are media content: movies and clips in series, TV movies and shows, movies, and episodes of TV series. This dataset was constructed by collecting the interactions and impressions of a subset of users during a period of four months. The **MIND [32]** dataset contains interactions and impressions of users with the Microsoft News service. Users are registered accounts of the service, while items are news articles, where impressions are personalized recommendations of items. Although its collection period spans six weeks. Impressions data is only available for the last two weeks, i.e., weeks five and six. Week six only does not contains interaction, only impressions data. Lastly, the **FINN.no Slates [10, 11]** dataset contains interactions and impressions of users with the finn.no site; an online service that allows users to sell goods and services. Users are registered accounts on the platform, while items are products listed in the second-hand marketplace. Differently from the others, this dataset does not contain the date and time of interactions and impressions, however, the attribute *time step* establishes an order relation between data points. Only this dataset has impressions from recommendations and search results.

Impressions Recommenders. Originally, impressions recommenders were first described in industrial studies or competitions, such as the ACM RecSys Challenge. From previous works, these recommenders can be placed into two categories: re-ranking and impressions as user profiles. The **re-ranking** recommenders reorder a recommendation list created by a base recommender. In this category, two impressions recommender exist: the *impressions discounting framework* and *cycling*. The first re-orders the recommendation list using features extracted from impressions data, e.g., frequency and position of impressions [19]. The second re-orders the recommendation list by sorting items by their impressions frequency, and in case of ties, by the score given by the base recommender [33]. On the other hand, the **impressions as user profiles** recommenders may treat impressions as user interactions are traditionally used. For instance, in the ACM RecSys 2016, impressions were used instead of interactions as user profiles, and in some solutions, both data sources were used, i.e., impressions to compute similarities while using interactions as user profiles or vice versa [5, 20, 21, 24, 27, 28, 36].

Signals. Impressions mostly contain mixed signals, e.g., both positive and negative traits of user preferences. Determining if such signals can be used to model the preferences of users is still an open research question in the RS domain [34]. Several factors affect the type of signal that exists within impressions. The main aspect to consider is that impressions, as interactions data, are strongly related to the system that generated the impression (e.g., a recommender system or a search function), to the characteristics of the system (e.g., position of items in single list, carousel, or grid layout), and to the context of the impressions (e.g., enforcement of business rules or editorial selections). Currently, no previous research work has studied the signals within current open impressions datasets and this constitutes an open research question of significant importance.

Challenges. Different properties of impressions datasets become challenges when evaluating recommendation models with impressions data, specifically, challenges regarding scalability and information leak are of utmost importance. Regarding **scalability** challenges, recommendation pipelines must account for scalability when working with impressions data as their size might be several orders of magnitude higher than interactions data. Hence, solutions that are both efficient in training time and space requirements that can work with this vast amount of data points must be more positively favored when compared to their complex counterparts. Regarding **information leak** challenges, they particularly emerge when using impressions as part of the recommendation process. The goal of the recommender that is being developed is to generate a recommendation list that contains the interactions in the user's ground truth. Hence, this very recommendation list must not be used in the form of impressions during the training process. In many systems, users only interact with impressed items. Therefore if the recommendation list shown to the user when the test interaction occurred is available during training, the recommendation model will know that the correct recommendation is necessarily within that list. Consequently, overestimating its recommendation quality.

This challenge implies, for instance, discarding methodologies such as traditional random-holdouts, as they introduce time-travel effects [17] and may leak impressions and interactions from future actions into training sets.

Evaluation with Impressions. Two different categories of previous impressions research exist: (i) solutions to recommendation challenges, such as ACM RecSys Challenges or Kaggle competitions, and (ii) case-studies in private organizations on private datasets. For the first category, research works have the main goal of using the provided data to develop algorithms with the best ranking on the challenge leaderboard [9, 15, 27, 28]. A limitation of these works is that they represent an analysis of a specific domain with the characteristics present in the available dataset, with no analysis on how the propose method could generalize to other data sources and scenarios. The second category of works instead analyzes the impact of impression models within a specific recommender system [3, 16, 19, 34]. To the best of our knowledge, only Lee et al. [19] presents a comparison of an impression model on several datasets using both an offline and online evaluation setup. However, there are two main drawbacks in the work's evaluation setting. First,

the evaluation was performed on two private datasets and in one dataset with a non-redistribute clause. Second, only one impression work was tested, instead of several approaches.

3 EXPERIMENTAL METHODOLOGY

This section describes the evaluation study proposed in this work. The goal of the study is two-fold. First, to benchmark the recommendation quality of base and impressions recommenders on a single evaluation methodology using open-source datasets. Second, to disentangle the signals within impressions. To achieve these goals, impressions data are used to enhance the recommendation quality of already available and properly-tuned recommendation models. To the best of our knowledge, this is the first comprehensive evaluation of impressions data in the domain of RS. Note that it is beyond the scope of this work to develop new recommendation models that use impressions. [1]

Recommendation Task. The evaluation study consist on experiments on recommender systems, in which the recommendation task is the traditional top-N recommendations. All experiments follow a traditional *leave last interaction out* approach, where recommendation models are trained on the training split and their hyper-parameters are optimized on the validation split. This strategy was selected due to its compliance with the arguments presented in Section 2, i.e., evaluations must account for time-travel effects.[2] Final recommendation quality is obtained by training recommendation models on the union of the training and validation splits, using their best hyper-parameters, and evaluating against the test split. The evaluation reports traditional accuracy and beyond-accuracy metrics for all recommenders.

Datasets. The evaluation evaluates recommenders on the three existing open-source datasets with impressions, i.e., the ContentWise Impressions [26], MIND [32], and FINN.no Slates [10, 11] datasets. For the ContentWise Impressions dataset all interactions and impressions that produced an interaction are used. For the MIND dataset, the *MIND-SMALL* version is used. For the FINN.no Slates dataset, due to its very large size a reduced version is created by following the sampling strategy used to build MIND-SMALL [32], i.e., selecting at random all interactions and impressions of 5 % of users. The datasets were processed as follows: (i) data points were sorted in ascending order by their time-related attribute, (ii) duplicated user-item interactions are aggregated into a single record, keeping the attributes of the first interaction, (iii) users without a minimum of 3 records were removed, and (iv) splits were created.

Base Recommenders. The baselines used in the evaluation of this work are standard collaborative filtering recommendation models trained on historical interactions data: neighborhood based heuristic models **ItemKNN CF**, **UserKNN CF** with asymmetric cosine similarity and shrinkage [30]; graph-based models **RP3beta** [25]; machine learning models **SLIM ElasticNet** [23], **MF BPR** [29], **PureSVD** [8], **NMF** [7]; and the auto-encoder **EASE R** [31]. A description of these recommendation models, their hyper-parameters and ranges, can be found in [13].[3]

Impressions Recommenders. This work evaluates the following impressions recommenders on top of the previously-mentioned base recommenders: re-ranking **Cycling** [33] and **ID** [19]; and the impressions as user profiles **IUP** [28]. The evaluation overrides the original signals assumptions held about impressions data of each recommender, i.e., in this work the hyper-parameter optimizer decides whether impressions should be considered positive or negative signals. The supplemental material provides a description of the hyper-parameters of impressions recommenders.

Hyper-Parameter Tuning. This work uses a traditional Bayesian Search [13] with 16 initial random cases and 50 maximum total cases. Each recommender is allotted a maximum of 14 days to complete its hyper-parameter search.[4] If the search is not completed after this time frame, then the search halts and the recommender is evaluated on the best hyper-parameters found thus far. If no sufficient number of cases are explored by the 14 days mark, then the recommender is not included in the results. [5] Hyper-parameters range and distributions of base recommendation models are the same reported in [13] whereas for impression models are summarized in the supplemental material. For the analysis of signals within impressions, impressions recommenders were optimized three times: (i) the optimizer decides the signal of impressions, (ii) the signal is manually set as positive, and (iii) the signal is manually set as negative.

4 RESULTS AND DISCUSSIONS

Overall, as impressions recommenders (3) are tested in combination with base recommenders (10), which themselves are optimized, and the evaluation was done on three datasets, this analysis required to fit and evaluate 4500 models. Table 1 and Table 2 present the top-20 ranking accuracy and beyond-accuracy of all recommenders on the ContentWise Impressions, MIND, and FINN.no Slates datasets. Due to space limitations, the tables only presents the recommendation quality measured by NDCG and Item Coverage. Results on other metrics, e.g., precision, novelty [35], and diversity gini [2] are provided in the supplemental material of this work. The results are preliminary as some recommenders could not be evaluated due to memory or time constraints. For instance, the recommenders **SLIM BPR** and **SLIM ElasticNet** could not be trained on the FINN.no Slates dataset due to out of memory errors.

4.1 Recommendation Quality

From a recommender-agnostic perspective, the results suggest that impressions positively impact both ranking accuracy and item coverage on the FINN.no Slates dataset by increasing both metrics on all recommendation models with impressions. On the other side, on

[1] Supplemental material, including source code, scripts, datasets, and results are permanently placed in: https://github.com/recsyspolimi/recsys-2022-evaluation-of-recsys-with-impressions

[2] Further studies may select different evaluation methodologies depending on their context, see Castells and Moffat [6].

[3] Matrix Factorization recommenders were folded-in [8] for compatibility with some impressions recommenders.

[4] The experiments were executed on a single `m6i.4xlarge` Amazon Web Services virtual machine with 16 vCPUs, 64 GB of RAM, and using Ubuntu 20.04 as operating system.

[5] Due to time limitations, this resulted in the exclusion of 20 out of 150 base-impressions recommender pairs.

Table 1: Top-20 ranking accuracy of base and impressions recommenders on the ContentWise Impressions dataset. MF BPR, NMF, and PureSVD are folded recommenders. Values in bold mean higher accuracy or item coverage than Baseline. ID refers to impressions discounting. IUP refers to impressions as user profiles. Suffix + indicates the impressions were selected to be positive signals, while - to be negative signals. Results on different metrics are available in the supplemental material.

	NDCG					
Recommender	**Baseline**	**Cycling**	**ID**	**IUP**	**IUP+**	**IUP-**
ItemKNN CF	0.09569	0.08883	**0.09607**	**0.09584**	0.09568	**0.09583**
UserKNN CF	0.09327	0.08690	**0.09344**	**0.09437**	0.09315	**0.09446**
MF BPR	0.06931	0.06305	**0.06940**	**0.06952**	**0.06953**	0.06928
NMF	0.03847	0.03567	**0.03917**	0.03841	0.03793	**0.03847**
PureSVD	0.06924	0.06386	0.06920	**0.07162**	**0.07191**	0.06924
RP3beta	0.09756	0.09040	0.09753	**0.09759**	**0.09757**	**0.09759**
SLIM BPR	0.08626	0.07888	0.08601	**0.08635**	**0.08627**	**0.08642**
SLIM ElasticNet	0.11194	0.10427	**0.11203**	0.11185	**0.11198**	0.11187

the ContentWise Impressions dataset beyond-accuracy metrics are slightly negatively affected when achieving higher accuracy and vice-versa. Lastly, on the MIND dataset, recommendation models that reached higher accuracy with impressions, achieved it by drastically decreasing their coverage. The diverse results on each dataset suggest that using impressions may be beneficial in some domains and data, while in others more sophisticated strategies must be developed.

The **Cycling** recommender is expected to provide more diverse recommendations by its definition [33], as it penalizes frequently impressed items. Hence, the results on the ContentWise Impressions and MIND datasets are expected. It is not the case for the FINN.no Slates dataset, which achieved higher ranking accuracy and item coverage on all base recommenders. For instance, the **UserKNN CF** and **NMF** recommenders achieved relative improvements of 31.53 % and 572.98 %, respectively. Furthermore, this impressions recommender made **UserKNN CF** the most accurate recommender; without it, both **ItemKNN CF** and **RP3beta** are more accurate than **UserKNN CF**.

For the **ID** recommender, the results on the ContentWise Impressions dataset show comparable accuracy metrics on some base recommenders. When looking at the hyper-parameters chosen by the optimizer, the importance of impressions sit between 3.6^{-2} and 2.5^{-1} on those impressions recommenders with higher accuracy than the base recommender. Hence, there is slight importance of the impressions in the improvement of recommendation quality. On the MIND dataset, the recommender achieves higher accuracy on only two base recommenders (**UserKNN CF** and **PureSVD**), however, at the cost of much lower item coverage. On the FINN.no Slates dataset, instead, the **ID** recommender achieved higher accuracy on all tested base recommenders. Its item coverage also increases on all base recommenders except for **ItemKNN CF**. Furthermore, on all recommenders, the importance of impressions was set to the maximum value, i.e., 1.0.

For the **IUP** recommender, on the ContentWise Impressions dataset, only some base recommenders achieved higher accuracy while also increasing its item coverage. For the latter, only two base recommenders (**MF BPR** and **PureSVD**) translated into lower item coverage and higher accuracy. When inspecting the hyper-parameters chosen by the optimizer, the importance of impressions sit between 4.4^{-2} and 1.9^{-1}, indicating slight importance of the impressions in the improvement of recommendation quality. On the MIND dataset, higher accuracy was achieved only on two recommenders (**UserKNN CF** and **PureSVD**) using impressions. Similar to the **ID** recommender, higher accuracy on this dataset translated to lower item coverage. On the FINN.no Slates dataset, instead, all tested recommenders achieved higher accuracy—however, at the cost of item coverage, which decreased. Impressions have slight importance in this dataset, which ranges between 4.7^{-2} and 1.8^{-1}. The exception is **NMF**, as the importance of impressions is close to the lower end of the range, i.e., 1.02^{-5}.

4.2 Impressions Signals

Now the discussion shifts to categorizing the signals within the impressions data for the general case. This discussion uses the following categories to classify impressions signals in terms of what can be inferred as the user preference for the item: *positive* and *negative*. Overall, the evaluation's preliminary results indicate that the information contained in the impressions is nuanced and varies according to the dataset and what type of recommendation model is used. The evaluation of impressions in different scenarios or methodologies impacts how these may model users' preferences.

The evaluation of signals within impressions was only carried using the ContentWise Impressions with the **IUP** recommender due to time constraints (**IUP+** and **IUP-** on Table 1). The results suggest that this dataset's impressions signal is *negative*. The optimizer sets the signal as *positive* for the **MF BPR**, **PureSVD**, and **NMF** recommenders and as *negative* for the rest of the recommenders. When manually choosing the signal, two cases arise. First, when the sign matches that selected by the optimizer, then the importance of the impressions feature has similar value, and the changes in accuracy or beyond-accuracy are similar. Second, when the sign does not match what the optimizer selected, the importance of impressions is set to a low value, and the recommendation quality is lower or similar to the base recommender.

Interestingly, the evaluation results on the FINN.no Slates dataset suggest that impressions may be used as a positive signal of

Table 2: Top-20 ranking accuracy (NDCG) and beyond-accuracy (Item Coverage) metrics of base and impressions recommenders. NMF, and PureSVD are folded recommenders. Values in bold mean higher accuracy or item coverage than Baseline. ID refers to impressions discounting. IUP refers to impressions as user profiles. Due to memory and time limitations, some recommenders are not included. Results on different metrics are available in the supplemental material.

		NDCG				Item Coverage			
Dataset	**Recommender**	**Baseline**	**Cycling**	**ID**	**IUP**	**Baseline**	**Cycling**	**ID**	**IUP**
MIND	**ItemKNN CF**	0.00868	0.00693	0.00028	0.00012	0.54686	**0.62265**	**0.66200**	**0.66069**
	UserKNN CF	0.00766	**0.01797**	**0.01118**	**0.06681**	0.22501	0.12687	0.06214	0.02298
	NMF	0.00116	**0.00797**	0.00116	0.00098	0.11202	**0.25060**	0.02259	**0.57219**
	PureSVD	0.00010	**0.00728**	**0.00015**	**0.00015**	0.14163	**0.26647**	0.00580	0.12583
	RP3beta	0.01643	0.00720	0.00013	0.00009	0.60989	0.53792	0.52827	0.53557
	SLIM ElasticNet	0.01493	0.00699	0.00060	0.00010	0.40652	**0.48159**	**0.44872**	0.38962
FINN.no Slates	**ItemKNN CF**	0,03933	**0,04538**	**0,04568**	**0,04018**	0,35260	**0,52645**	0,34601	0,33779
	UserKNN CF	0,03841	**0,05052**	**0,04440**	**0,04055**	0,26051	**0,47073**	**0,27781**	0,25993
	NMF	0,00681	**0,03902**	**0,00881**	**0,00718**	0,02563	**0,52985**	**0,03791**	0,02542
	PureSVD	0,01502	**0,04024**	**0,02044**	**0,01682**	0,04639	**0,51562**	**0,06872**	0,03958
	RP3beta	0,03883	**0,04614**	**0,04538**	**0,03929**	0,30450	**0,51916**	**0,30493**	0,30194

user preference towards items, as in all cases, the optimizer selected impressions as positive signals, and all cases achieved higher accuracy. On the other hand, on the MIND dataset, impressions were mostly considered chosen as positive by the optimizer, however, this does not translated to better accuracy in the majority of cases.

5 CONCLUSIONS

This work aims to raise awareness of impressions and encourage their use in further RS research. The work presents an overview of impressions in RS and the different signals and challenges that must be considered when working with impressions data. Also, this work presents the first evaluation study of recommender systems with impressions. An evaluation like the one presented here is lacking in the current literature. Previous research works with impressions were published using private datasets, recommenders, or in the context of recommendation challenges, such as the ACM RecSys Challenge. The preliminary results of this evaluation study are promising. They suggest that impressions data is complex, and their signal is nuanced. Nonetheless, including impressions may be beneficial in terms of accuracy and beyond-accuracy metrics. Hence, translating into higher-quality recommendations to users.

REFERENCES

[1] Fabian Abel, András A. Benczúr, Daniel Kohlsdorf, Martha A. Larson, and Róbert Pálovics. 2016. RecSys Challenge 2016: Job Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016*, Shilad Sen, Werner Geyer, Jill Freyne, and Pablo Castells (Eds.). ACM, 425–426. https://doi.org/10.1145/2959100.2959207

[2] Gediminas Adomavicius and YoungOk Kwon. 2012. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Trans. Knowl. Data Eng.* 24, 5 (2012), 896–911. https://doi.org/10.1109/TKDE.2011.15

[3] Deepak Agarwal, Bee-Chung Chen, Rupesh Gupta, Joshua Hartman, Qi He, Anand Iyer, Sumanth Kolar, Yiming Ma, Pannagadatta Shivaswamy, Ajit Singh, and Liang Zhang. 2014. Activity ranking in LinkedIn feed. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani (Eds.). ACM, 1603–1612. https://doi.org/10.1145/2623330.2623362

[4] Michal Aharon, Yohay Kaplan, Rina Levy, Oren Somekh, Ayelet Blanc, Neetai Eshel, Avi Shahar, Assaf Singer, and Alex Zlotnik. 2019. 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*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 2793–2801. https://doi.org/10.1145/3357384.3357801

[5] Tommaso Carpi, Marco Edemanti, Ervin Kamberoski, Elena Sacchi, Paolo Cremonesi, Roberto Pagano, and Massimo Quadrana. 2016. Multi-stack ensemble for job recommendation. In *Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016*, Fabian Abel, András A. Benczúr, Daniel Kohlsdorf, Martha A. Larson, and Róbert Pálovics (Eds.). ACM, 8:1–8:4. https://doi.org/10.1145/2987538.2987541

[6] Pablo Castells and Alistair Moffat. 2022. Offline recommender system evaluation: Challenges and new directions. *AI Magazine* 43, 2 (2022), 225–238. https://doi.org/10.1002/aaai.12051 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12051

[7] Andrzej Cichocki and Anh Huy Phan. 2009. Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* 92-A, 3 (2009), 708–721. https://doi.org/10.1587/transfun.E92.A.708

[8] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010*, Xavier Amatriain, Marc Torrens, Paul Resnick, and Markus Zanker (Eds.). ACM, 39–46. https://doi.org/10.1145/1864708.1864721

[9] Edoardo D'Amico, Giovanni Gabbolini, Daniele Montesi, Matteo Moreschini, Federico Parroni, Federico Piccinini, Alberto Rossettini, Alessio Russo Introito, Cesare Bernardis, and Maurizio Ferrari Dacrema. 2019. Leveraging laziness, browsing-pattern aware stacked models for sequential accommodation learning to rank. In *Proceedings of the Workshop on ACM Recommender Systems Challenge, Copenhagen, Denmark, September 2019*, Peter Knees, Yashar Deldjoo, Farshad Bakhshandegan Moghaddam, Jens Adamczak, Gerard Paul Leyson, and Philipp Monreal (Eds.). ACM, 7:1–7:5. https://doi.org/10.1145/3359555.3359563

[10] Simen Eide, David S. Leslie, and Arnoldo Frigessi. 2021. Dynamic Slate Recommendation with Gated Recurrent Units and Thompson Sampling. *CoRR* abs/2104.15046 (2021), 30 pages. arXiv:2104.15046 https://arxiv.org/abs/2104.15046

[11] Simen Eide, David S. Leslie, Arnoldo Frigessi, Joakim Rishaug, Helge Jenssen, and Sofie Verrewaere. 2021. FINN.no Slates Dataset: A new Sequential Dataset Logging Interactions, all Viewed Items and Click Responses/No-Click for Recommender Systems Research. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Hurnink, and Even Oldridge (Eds.). ACM, 556–558. https://doi.org/10.1145/3460231.3474607

[12] Nicolò Felicioni, Maurizio Ferrari Dacrema, and Paolo Cremonesi. 2021. A Methodology for the Offline Evaluation of Recommender Systems in a User Interface with Multiple Carousels. In *Adjunct Publication of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June 21-25, 2021*, Judith Masthoff, Eelco Herder, Nava Tintarev, and Marko Tkalcic (Eds.). ACM, 10–15. https://doi.org/10.1145/3450614.3461680

[13] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *ACM Trans. Inf. Syst.* 39, 2 (2021), 20:1–20:49. https://doi.org/10.1145/3434185

[14] Maurizio Ferrari Dacrema, Nicolò Felicioni, and Paolo Cremonesi. 2022. Offline Evaluation of Recommender Systems in a User Interface With Multiple Carousels. *Frontiers Big Data* 5 (2022), 910030. https://doi.org/10.3389/fdata.2022.910030

[15] Jose Ignacio Honrado, Oscar Huarte, Cesar Jimenez, Sebastian Ortega, José R. Pérez-Agüera, Joaquín Pérez-Iglesias, Álvaro Polo, and Gabriel Rodríguez. 2016. Jobandtalent at RecSys Challenge 2016. In *Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016*, Fabian Abel, András A. Benczúr, Daniel Kohlsdorf, Martha A. Larson, and Róbert Pálovics (Eds.). ACM, 3:1–3:5. https://doi.org/10.1145/2987538.2987547

[16] Maya Hristakeva, Daniel Kershaw, Marco Rossetti, Petr Knoth, Benjamin Pettit, Saúl Vargas, and Kris Jack. 2017. 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, 25–32. https://doi.org/10.1145/3057148.3057152

[17] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2022. A Critical Study on Data Leakage in Recommender System Offline Evaluation. arXiv:2010.11060 [cs.IR]

[18] Peter Knees, Yashar Deldjoo, Farshad Bakhshandegan Moghaddam, Jens Adamczak, Gerard Paul Leyson, and Philipp Monreal. 2019. RecSys challenge 2019: session-based hotel recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 570–571. https://doi.org/10.1145/3298689.3346974

[19] Pei Lee, Laks V. S. Lakshmanan, Mitul Tiwari, and Sam Shah. 2014. Modeling impression discounting in large-scale recommender systems. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani (Eds.). ACM, 1837–1846. https://doi.org/10.1145/2623330.2623356

[20] Vasily A. Leksin and Andrey Ostapets. 2016. Job recommendation based on factorization machine and topic modelling. In *Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016*, Fabian Abel, András A. Benczúr, Daniel Kohlsdorf, Martha A. Larson, and Róbert Pálovics (Eds.). ACM, 6:1–6:4. https://doi.org/10.1145/2987538.2987542

[21] Kuan Liu, Xing Shi, Anoop Kumar, Linhong Zhu, and Prem Natarajan. 2016. Temporal learning and sequence modeling for a job recommender system. In *Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016*, Fabian Abel, András A. Benczúr, Daniel Kohlsdorf, Martha A. Larson, and Róbert Pálovics (Eds.). ACM, 7:1–7:4. https://doi.org/10.1145/2987538.2987540

[22] James McInerney, Ehtsham Elahi, Justin Basilico, Yves Raimond, and Tony Jebara. 2021. Accordion: A Trainable Simulator forLong-Term Interactive Systems. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge (Eds.). ACM, 102–113. https://doi.org/10.1145/3460231.3474259

[23] Xia Ning and George Karypis. 2011. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu (Eds.). IEEE Computer Society, 497–506. https://doi.org/10.1109/ICDM.2011.134

[24] Andrzej Pacuk, Piotr Sankowski, Karol Wegrzycki, Adam Witkowski, and Piotr Wygocki. 2016. RecSys Challenge 2016: job recommendations based on preselection of offers and gradient boosting. In *Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, Boston, Massachusetts, USA, September 15, 2016*, Fabian Abel, András A. Benczúr, Daniel Kohlsdorf, Martha A. Larson, and Róbert Pálovics (Eds.). ACM, 10:1–10:4. https://doi.org/10.1145/2987538.2987544

[25] Bibek Paudel, Fabian Christoffel, Chris Newell, and Abraham Bernstein. 2017. Updatable, Accurate, Diverse, and Scalable Recommendations for Interactive Applications. *ACM Trans. Interact. Intell. Syst.* 7, 1 (2017), 1:1–1:34. https://doi.org/10.1145/2955101

[26] Fernando B. Pérez Maurera, Maurizio Ferrari Dacrema, Lorenzo Saule, Mario Scriminaci, and Paolo Cremonesi. 2020. ContentWise Impressions: An Industrial Dataset with Impressions Included. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 3093–3100. https://doi.org/10.1145/3340531.3412774

[27] Toon De Pessemier, Kris Vanhecke, and Luc Martens. 2016. 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*, Fabian Abel, András A. Benczúr, Daniel Kohlsdorf, Martha A. Larson, and Róbert Pálovics (Eds.). ACM, 5:1–5:4. https://doi.org/10.1145/2987538.2987539

[28] Mirko Polato and Fabio Aiolli. 2016. 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*, Fabian Abel, András A. Benczúr, Daniel Kohlsdorf, Martha A. Larson, and Róbert Pálovics (Eds.). ACM, 1:1–1:4. https://doi.org/10.1145/2987538.2987549

[29] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, Jeff A. Bilmes and Andrew Y. Ng (Eds.). AUAI Press, 452–461. https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1630&proceeding_id=25

[30] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5, 2001*, Vincent Y. Shen, Nobuo Saito, Michael R. Lyu, and Mary Ellen Zurko (Eds.). ACM, 285–295. https://doi.org/10.1145/371920.372071

[31] Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 3251–3257. https://doi.org/10.1145/3308558.3313710

[32] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 3597–3606. https://doi.org/10.18653/v1/2020.acl-main.331

[33] Qian Zhao, Gediminas Adomavicius, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2017. Toward Better Interactions in Recommender Systems: Cycling and Serpentining Approaches for Top-N Item Lists. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, Charlotte P. Lee, Steven E. Poltrock, Louise Barkhuus, Marcos Borges, and Wendy A. Kellogg (Eds.). ACM, 1444–1453. https://doi.org/10.1145/2998181.2998211

[34] Qian Zhao, Martijn C. Willemsen, Gediminas Adomavicius, F. Maxwell Harper, and Joseph A. Konstan. 2018. Interpreting user inaction in recommender systems. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 40–48. https://doi.org/10.1145/3240323.3240366

[35] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515. https://doi.org/10.1073/pnas.1000488107

[36] Dávid Zibriczky. 2016. 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*, Fabian Abel, András A. Benczúr, Daniel Kohlsdorf, Martha A. Larson, and Róbert Pálovics (Eds.). ACM, 9:1–9:4. https://doi.org/10.1145/2987538.2987548