Warming up From Extreme Cold start Using Stereotypes with Dynamic User and Item Features

Nourah A. AlRossais[1,†]

[1]King Saud University, College of Computer and Information Sciences, Information Technology Department

Abstract

A demanding operation regime of Reccomender Systems (RS) is that of extreme cold start followed by a 'warming up' phase, in which a new user begins to interact with the items in the catalogue, or a new item receives the first interactions by existing users. In the RS literature the majority of approaches and techniques have been proposed for a fully 'warm' RS; only a small subset of the research addresses the cold start and within that subset, the 'warming up' phase is an area that has received little attention. During warm up new user (new item) interactions begin to appear but they are too few for a collaborative filtering technique to work, while a smart content based filtering approach (those of extreme cold start) may not be able to capture the emerging personalization traits. In this paper, starting from a stereotype driven approach developed for pure cold starts, we formulate and discuss a dynamic model that uses the few arising user/item interactions to adjust simple personalization features that are embedded in the proposed model. We demonstrate how the little personalization introduced by the dynamic model improves substantially a range of performance metrics during warm up.

Keywords

Recommender system, cold start, warming up, new item problem, new user problem, stereotypes, dynamic bias

1. Introduction

One of the most challenging areas of operation for a Recommender System (RS) is that of cold start, namely when the RS is required to produce content recommendations to new unknown users, or when new content is added to the catalogue and the task is to recommend the new unknown content to its existing users. Different solutions have been proposed to address cold start, some data driven and some technique driven, as reviewed in [1]. The author in [2] demonstrated how creating rating agnostic stereotypes for both users and items lead to better recommendation metrics during extreme cold start (i.e. zero interactions available from the new user or concerning the new item); only deep learning architectures can achieve comparable levels of accuracy, serendipity and fairness during extreme cold start to that of stereotyped metadata features [3].

When a new user begins interacting with the items catalogue, or when new content begins to be rated by some users, the RS can make use of such information to update its extreme cold start algorithm in light of the new evidence. During such "warm up" phase, too few interactions are available to switch to a collaborative filtering approach, but such little interactions should not be omitted as they may give important personalization traits. Different lines of research have been dedicated to the warming up phase, some using a Bayesian construct (Markov chains) to incorporate incoming information [4, 5], or modeling sequential sessions via recurrent neural networks [6]. Meta learning and reinforcement learning have been applied to cold start problems, [7, 8]. More recently deep learning recommendation problems have been solved adding an embedding layer before the deep learning layers, the embedding is also subject to cold start and there are multiple proposed approaches on how initialization and updating of the embedding could be handled at cold start and during the warming phase, [9, 10].

In the recent work [11] the authors introduce temporal effects into a latent factor model, without the aim to address cold start, but rather to be able to adapt the proposed factorization to time changing preferences of users.

In this work we elaborate the idea of [11] and apply it away from their intended factorization approach. We introduce dynamic time varying user and item factor biases in the extreme cold start framework of stereotypes explained and validated in details in [2]. We asses the performance of the models for both the new user and new item experiments and discuss possible future improvements.

2. Approach

In a generic RS framework, user u consumes content/item i and attributes an explicit or implicit liking via a rating metric r_{iu}. A RS can be represented by a function (po-

Fifth Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2023, September 18–22 2023, Singapore.

*Corresponding author.

nalrossais@ksu.edu.sa (N. A. AlRossais)

https://ksu.edu.sa/ (N. A. AlRossais)

0009-0008-1062-9001 (N. A. AlRossais)

tentially non linear or stochastic, e.g. a deep learning architecture) that models r_{iu} via the user's describing feature vector U_u, the item's describing feature vector I_i, via all previous ratings $r_{i\omega}$ provided to item i by all other users ($\omega \neq u$) as well as such other users' features, and finally all previous ratings $r_{\eta u}$ given by the user u to other items η and all such other item's features. In our previous works [2, 3] we modeled further the functional form of our RS as in equation (1):

$$r_{iu} = F(I_i, \lambda_i, U_u, \mu_u) \tag{1}$$

Equation (1) simplifies further the problem by condensing the explanatory effects provided by all other users having consumed item i prior to user u in a simplified term, λ_i, that we call characteristic item bias. In a similar fashion, the explanatory effects provided by all other items rated by user u prior to encountering item i, are incorporated in a simplified term, μ_u, the characteristic user bias. All previous user to item interactions, not directly involving user u and item i, together with their metadata features, are used in the generation of the functional form F, i.e. training the model (1).

Sterotypization of the features was introduced via standard algorithms that can take any metadata feature type and transform it into a stereotyped version of the original metadata with substantially smaller dimensions. Doing so for both user and item features aids the sparsity of the problem, and simultaneously creates a basis, $\tilde{I}$ and $\tilde{U}$, that improves the training process and predictive power of model (1), as illustrated in [2].

During the new item and new user extreme cold starts, using stereotyped features, the RS can be rewritten for the new user extreme cold start problem as:

$$r_{iu} = \phi(\tilde{I}_i, \lambda_i, \tilde{U}_u, \bar{\mu}(\tilde{U}_u)) \tag{2}$$

And for the new item extreme cold start problem as:

$$r_{iu} = \psi(\tilde{I}_i, \bar{\lambda}(\tilde{I}_i), \tilde{U}_u, \mu_u) \tag{3}$$

In the new user cold start, given that no prior rating information is available about the new user, $\bar{\mu}(\tilde{U}_u)$ represents the typical user bias of all previously observed users belonging to the same stereotypes. Likewise in the new item problem, $\bar{\lambda}(\tilde{I}_i)$ represents the typical item bias across all items belonging to the same stereotype.

In the present research we extend the models (2,3) to adapt dynamically during the warming up phase using the concept of dynamic adaptive features introduced by [11]. In particular, we can write for the new user warming up phase, when there are exactly k ratings available for user u and we are modeling the $k+1$ consumption of item i:

$$\begin{aligned} r_{iu}^{(k+1)} &= \phi(\tilde{I}_i, \lambda_i, \tilde{U}_u, \mu_u^{(k+1)}) \\ \mu_u^{(k+1)} &= \alpha \cdot \bar{\mu}(\tilde{U}_u) + (1-\alpha) \cdot < \mu_u >_{1,..,k} \\ \alpha &= (k/N_u)^{\gamma} \end{aligned} \tag{4}$$

In (4) $\mu_u^{(k+1)}$ represents the dynamic user bias. $< \mu_u >_{1,..,k}$ is the average observed bias of the particular user u, over its first k reviews. When k is small the observed bias is not trustworthy, and it receives a low weight compared to $\mu_u^{(k+1)}$. As the user advances in his interactions with the items a stronger personalization arises, and when k grows to the value of N_u the user bias is fully personalized. In this model N_u and γ are also model's parameters optimized during training.

The new item warming up phase can be modeled similarly:

$$\begin{aligned} r_{iu}^{(k+1)} &= \psi(\tilde{I}_i, \lambda_i^{(k+1)}, \tilde{U}_u, \mu_u) \\ \lambda_i^{(k+1)} &= \alpha \cdot \bar{\lambda}(\tilde{I}_i) + (1-\alpha) \cdot < \lambda_i >_{1,..,k} \\ \alpha &= (k/N_i)^{\gamma} \end{aligned} \tag{5}$$

In (5) $\lambda_i^{(k+1)}$ is the dynamic item bias, expressed as a dynamic weighted sum of $\bar{\lambda}(\tilde{I}_i)$ and the actual item i average bias for its first k reviews, $< \lambda_i >_{1,..,k}$. The model improves the item characterisation as its number of reviews k increases, obtaining full characterization after N_i interactions.

3. Experimental settings and results

The ideal dataset to demonstrate the application of models (4, 5), must exhibit several characteristics; firstly the ratings (interactions) should be timestamped, so that the full history of how new users, new items and user/item interactions occurred is available. Secondly it should be rich in metadata features for both users and items, and finally it should be publicly available to make findings reproducible. Based on these requirements the dataset used is the Movielens/IMDb integrated dataset described in [12].

3.1. Cold start warming experiments

Beginning from pure cold start experiments, as described in [2, 3], when a user (item) is left out to represent a new user (item), all its ratings and interactions are blanked out and temporal consistency is enforced during the experiments. Only the users/items/ratings that had been expressed at a time prior to the new user (item) joining

the online platform are retained to train the RS functional shapes, i.e. the ϕ and the ψ of (2,3,4,5). In addition, the temporal order of the interactions provided by the new user (or to the new item) must be maintained when training and evaluating the dynamic biases of (4), (5).

Following the works and findings of [2, 3] we train two algorithms for both ϕ and ψ of problems (2,3,4,5), the first one is Extreme Gradient Boosting, [13] (XGB). This is chosen as our benchmark, because in the previous work [2] it represented the best extreme cold start model among the various machine learning / statistical driven models discussed in [2]. The second is a deep neural network algorithm, (DNN) whose deep layers, coupled by dropout layers, and performance during extreme cold starts have been the subject of the research [3]. The main focus of the paper is not to detail which of the functional algorithms exhibits better performance metrics during cold start and during the warm up transition, instead the objective is to investigate the effect of the dynamic bias model, how that improves the warming up phase and whether the findings can be ascribed to any particular RS functional form, or if they are independent of the RS chosen.

In the results that follow we discuss the behaviour of the extreme cold start model and of the dynamic bias warm up model as the new user begins rating existing items (or existing users begin rating the new item). In particular we examine the performance metrics at several increasing review counts, W, i.e. how the metrics change when we take away the first W reviews of the user for the new user case (or reviews to the item for the new item) and use these to train the dynamic bias models (4), (5). When assessing performance only the interactions that follow, from $W + 1$, are utilized to compute the metric for both warming model as well as for the extreme cold start model.

In the experiments presented in this paper we restrict our attention to users that have more than 100 interactions with the catalogue (approximately 4,000 users), and items that have been interacted with at least 100 times (approximately 2,500 items).

In addition to well known accuracy metrics of single predictions and ranked lists we will also discuss how the warming up phase affects serendipity (SER) and fairness (FRN) of ranked lists. SER is the property of generating ranked lists that contain useful and unexpected items, therefore exposing users to larger parts of the catalogue. Fairness (FRN), which is related to SER, focuses on the tail of the distribution of the items that are recommended the fewest, it measures how often the least recommended items are injected in some lists. The latter metric is a very actual research issue in RS, after ethical concerns raised by some industry RS [14].

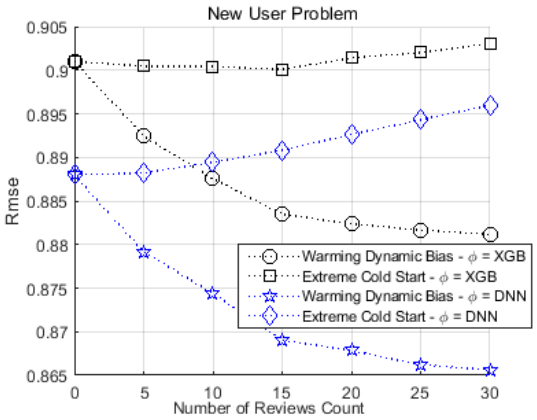

Figure 1: New user cold start model (2) and new user dynamic bias model (4) prediction root mean square error. Functional form XGB (square and circle markers), DNN (diamond and star markers). The figure demonstrates that independently of the functional form adopted for the RS, warming up the RS with a dynamic bias quickly improves the prediction quality as measured by Rmse, with already few interactions provided by the new user.

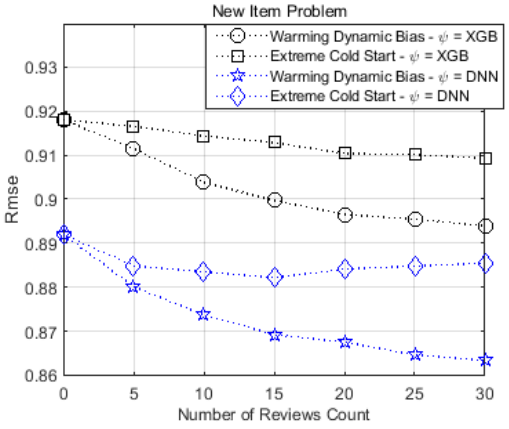

Figure 2: New item cold start model (3) and new item dynamic bias model (5) prediction root mean square error. Functional form XGB (square and circle markers), DNN (diamond and star markers). The figure demonstrates that independently of the functional form adopted for the RS, warming up the RS with a dynamic bias quickly improves the prediction quality as measured by Rmse, with already few interactions provided to the new item.

3.2. Results

In figures (1) and (2) we show the effect on the Rmse of the predicted ratings when warming the RS's personalization via the dynamic bias models based on stereotypes, versus continuing the recommendations using the stereotyped

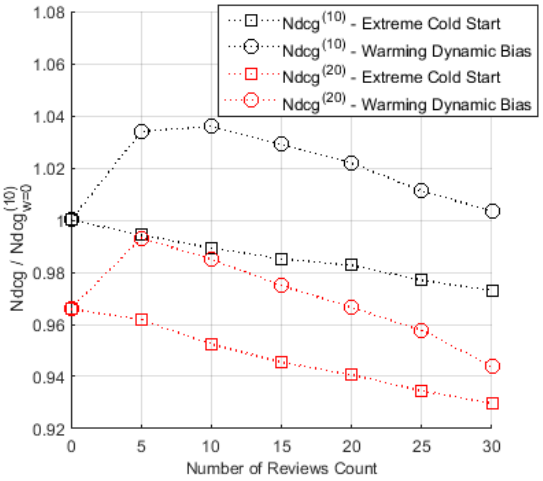

Figure 3: New User Problem. Functional form ϕ = XGB (DNN results are similar and not shown for brevity). Ndcg standardized by the value of the Ndcg of the top 10 list evaluated when the user has zero expressed ratings ($Ndcg^{(10)}_{w=0}$). The figure shows how the dynamic bias provides a boost in the ranked list quality by about 5% materializing during the first 5 to 10 user to items interactions.

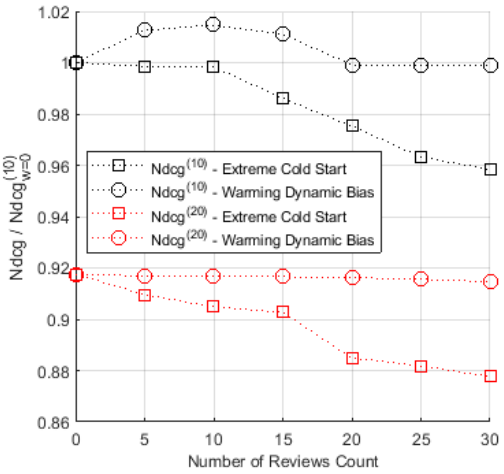

Figure 4: New Item Problem. Functional form ψ = DNN (XGB results are similar and not shown for brevity). Ndcg standardized by the value of the Ndcg of the top 10 list evaluated when there are no expressed ratings to items ($Ndcg^{(10)}_{w=0}$). The figure shows how the dynamic bias provides a boost in the ranked list quality that increases steadily as the item are reviewd.

extreme cold start model. Three distinct interesting facts can be inferred from these results; firstly the dynamic personalization improves the Rmse recommendation performance by 3 to 4% for both the new user and new item experiments. Secondly, the improvement obtained is independent of the functional form of the RS chosen in our experiments XGB vs DNN. Each RS has its own cold start base prediction ability and the warming up of the model bias is beneficial in a similar manner to both functional forms and across experiments. Thirdly, there is a different behaviour in the new user vs the new item experiments. As the user begins interacting with the catalogue it becomes more difficult for the pure stereotype based cold start model to predict the ratings of the items encountered further down the catalogue interactions path. This is potentially explained by the fact that items that are well known and widely interacted with are easier to predict and usually come earlier in the review histories. With new items the situation is slightly different, the personalization effect of a dynamic bias improves performance versus the cold start base model, but also the base algorithm improves (at a slower rate) as the reviews increase, hence indicating that the first temporal reviews to a new item are the most difficult to predict.

When moving from accuracy of single rating predictions to a description of how well a RS proposes ranked lists of items to users, there are several metrics that can be adopted as discussed in [2]. One of the most relevant and popular measure is the Normalized Discounted Cumulative Gain (Ndcg) whose definition can be found in [2] and references within. In this context it is sufficient to remind that the higher the Ndcg, the more valuable the content of the ranked list to the user (both in terms of items present and how they are ranked).

Figures (3) and (4) show the standardized Ndcg of both the top-10 and top-20 ranked lists for the new user and new item experiments. Given that the Ndcg tends to decay as the top-N list grows, we standardize all the data by the $Ndcg^{(10)}_{w=0}$ (the Ndcg of the top 10 list when there are no user to item interactions, i.e. the extreme cold start case). In all cases when there are W reviews available the assessment of the $Ndcg$ takes the items reviewed into account, excluding them from the lists, therefore also the ranked lists of the extreme cold start base model changes with W. In both experiments we notice the tendency of the $Ndcg$ to decay as the number of reviews W increases. In both experiments and for both functional forms tested for our warming up RS, the dynamic bias substantially improves the ranked list quality as measured by the $Ndcg$, albeit in a slightly different manner. In the case of the new user experiment the improvement due to personalization effects to the dynamic bias happens relatively quickly in this dataset, where an approximately 5% improvement over cold start base model is obtained within the first 5, 10 interactions lifting the Ndcg curve. In the new

Table 1
New user warm up experiment, metrics for serendipity (SER) and fairness (FRN) of the top N list as a function of the warm up reviews W, functional form of the RS DNN (XGB results are similar and not shown for brevity). The table shows how, compared to the pure cold start lists, the dynamic bias model offers a marked improvement in SER with a small reduction in FRN at small W which is quickly recovered at larger W.

SER Top 10, New user (ϕ = DNN)					
	W=0	W=5	W=10	W=15	W=20
Cold Base	1.000	0.992	1.012	1.020	1.036
Warm Bias	1.000	1.044	1.057	1.116	1.124
FRN Top 10, New user (ϕ = DNN)					
Cold Base	1.000	0.226	0.067	0.034	0.012
Warm Bias	1.000	0.147	0.041	0.042	0.077

item case the improvement over pure cold start is more distributed over the entire number of interactions. These two distinct behaviours may be related to the different statistical nature of users and items.

Our previous works [2, 3] introduced operative metrics for serendipity (SER) and fairness (FRN) of ranked lists, in this context we refer to such definitions. Table (1) shows the effect of warming up the dynamic bias on SER and FRN, for brevity we only report a single top-N for the new user case with the DNN functional shape, as a function of the new user growing interactions W (for brevity we do not show XGB results which are very similar). We can see that dynamic bias approach improves substantially the SER of the top N list as the personalization increases the lists cover more and more catalogue. However, the specialization also reduces the FRN for the new user case. As specialization occurs the probability of being recommended for the least recommended items decreases initially and only at larger W the FRN of the top-10 ranked list surpasses that of the base cold start model.

4. Conclusions and future work

The model proposed in this paper addresses the warming up of RS from cold start, it presents a unified construct for both the new user and new item problems in a stereotype driven framework. The model allows for a degree of personalization via dynamic user/item bias terms. The findings discussed in the paper demonstrate how for the new user problem the adoption of the bias allows for a rapid improvement in accuracy metrics, while for the new item problem the improvement is more gently spread over the initial incoming interactions. The only metric that using dynamic biases is observed to under perform the pure content based stereotype model is the fairness (FRN) metric for the new user problem. Such small under-performance in fairness may be the price to pay for a rapid user specialization. The new user and new item dynamic bias are effective with two alternative formulation of the recommendation algorithms, and that is indicating that they can be a source of improved accuracy independently of the RS functional shape chosen.

As a future work the author plans to enrich the models presented with a collaborative filtering element, to be modeled via a Markov chain. The dynamic bias presented in this work, coupled with a Bayesian effect (to be formulated) should provide a construct capable of smoothly transitioning from a content based cold start RS, based on stereotyped features, to a collaborative filtering model when there are enough user to items interactions.

Finally, this work presents results on an integrated dataset rich in user and item metadata, which also fulfils the necessary requirement of having interactions timestamped. Part of our future research will be focused on finding data sets that have the required characteristics and extending our findings to different domains.

References

[1] D. K. Panda, S. Ray, Approaches and algorithms to mitigate cold start problems in recommender systems: a systematic literature review, Journal of Intelligent Information Systems 59 (2022) 341–366.

[2] N. AlRossais, D. Kudenko, T. Yuan, Improving cold-start recommendations using item-based stereotypes, User Model User-Adap Inter 31 (2021) 867–905.

[3] N. AlRossais, Improving cold start stereotype-based recommendation using deep learning, 2023.

[4] S. Rendle, C. Freudenthaler, L. Schmidt-Thieme, Factorizing personalized markov chains for next-basket recommendation, in: Proceedings of the 19th international conference on World wide web, 2010, pp. 811–820.

[5] C. Cai, R. He, J. McAuley, Spmc: Socially-aware personalized markov chains for sparse sequential recommendation, arXiv preprint arXiv:1708.04497 (2017).

[6] J. Li, Y. Wang, J. McAuley, Time interval aware self-attention for sequential recommendation, in: Proceedings of the 13th international conference on web search and data mining, 2020, pp. 322–330.

[7] M. Vartak, A. Thiagarajan, C. Miranda, J. Bratman, H. Larochelle, A meta-learning perspective on cold-start recommendations for items, Advances in neural information processing systems 30 (2017).

[8] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International conference on machine learning, PMLR, 2017, pp. 1126–1135.

[9] Y. Zhu, R. Xie, F. Zhuang, K. Ge, Y. Sun, X. Zhang, L. Lin, J. Cao, Learning to warm up cold item embeddings for cold-start recommendation with meta scaling and shifting networks, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1167–1176.

[10] F. Pan, S. Li, X. Ao, P. Tang, Q. He, Warm up cold-start advertisements: Improving ctr predictions via learning to learn id embeddings, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 695–704.

[11] G. Behera, N. Nain, Collaborative filtering with temporal features for movie recommendation system, Procedia Computer Science 218 (2023) 1366–1373.

[12] N. A. ALRossais, D. Kudenko, isynchronizer: A tool for extracting, integration and analysis of movielens and imdb datasets, in: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, 2018, pp. 103–107.

[13] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, et al., Xgboost: extreme gradient boosting, R package version 0.4-2 1 (2015) 1–4.

[14] A. A. Kodiyan, An overview of ethical issues in using ai systems in hiring with a case study of amazon's ai based hiring tool (2019).