Unifying Explicit and Implicit Feedback for Rating Prediction and Ranking Recommendation Tasks

Amir H. Jadidinejad, Craig Macdonald, Iadh Ounis
University of Glasgow
{Amir.Jadidinejad,Craig.Macdonald,Iadh.Ounis}@glasgow.ac.uk

ABSTRACT

The two main tasks addressed by collaborative filtering approaches are rating prediction and ranking. Rating prediction models leverage explicit feedback (e.g. ratings), and aim to estimate the rating a user would assign to an unseen item. In contrast, ranking models leverage implicit feedback (e.g. clicks) in order to provide the user with a personalized ranked list of recommended items. Several previous approaches have been proposed that learn from both explicit and implicit feedback to optimize the task of ranking or rating prediction at the level of recommendation algorithm. Yet we argue that these two tasks are not completely separate, but are part of a unified process: a user first interacts with a set of items and then might decide to provide explicit feedback on a subset of items. We propose to bridge the gap between the tasks of rating prediction and ranking through the use of a novel weak supervision approach that unifies both explicit and implicit feedback datasets. The key aspects of the proposed model is that (1) it is applied at the level of data pre-processing and (2) it increases the representation of less popular items in recommendations while maintaining reasonable recommendation performance. Our experimental results – on six datasets covering different types of heterogeneous user's interactions and using a wide range of evaluation metrics – show that, our proposed approach can effectively combine explicit and implicit feedback and improve the effectiveness of the baseline explicit model on the ranking task by covering a broader range of long-tail items.

ACM Reference format:
Amir H. Jadidinejad, Craig Macdonald, Iadh Ounis. 2019. Unifying Explicit and Implicit Feedback for Rating Prediction and Ranking Recommendation Tasks. In *Proceedings of The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval, Santa Clara, CA, USA, October 2–5, 2019 (ICTIR '19),* 8 pages.
https://doi.org/10.1145/3341981.3344225

1 INTRODUCTION

Many recommendation algorithms are based on the notion of collaborative filtering, in that they are based on the interactions of users and items, obtained from either *explicit* feedback – where users directly express the preferences (e.g. star ratings) or *implicit* feedback – where users indirectly reveal their interests through interactions (e.g. clicks). Explicit feedback is more difficult to collect from users, hence more scarce, but, on the other hand it is more precise in reflecting the users' preferences [22]. In contrast, implicit feedback is much easier to collect although it is less accurate in reflecting the user preferences, as there is no explicit judgement by the users as to their liking of the item [17]. The difference between explicit and implicit feedback has led researchers to develop different models and techniques to address each of their distinct properties [23]. Multiple types of explicit and implicit feedback may be available in real-world recommendation systems and could potentially complement each other. Indeed, we argue that it is desirable to unify these two heterogeneous forms of user's interactions in order to generate more accurate recommendations.

On the other hand, recommendation systems are typically concerned with two tasks: *rating prediction* and *ranking* [3, 13, 22]. The goal of a *rating prediction task* is to predict the possible rating a user would assign to a particular item. In contrast, the goal of a *ranking task* is to provide the user with a personalized rank list of items (also known as top-k recommendation). Despite the logical relation between these two tasks, different models and techniques have been developed in the literature and in fact researchers often distinguish between the two broad categories when measuring recommendation effectiveness [22]: the rating prediction task is usually quantified in terms of the Root Mean Square Error [20]; while the ranking task effectiveness is measured using Information Retrieval (IR) metrics [1]. Despite the popularity of the ranking task in the field of recommendation systems, ranking can be considered to be a sub-problem of the rating prediction task [22], since accurate rating prediction models can enable the ranking of the items with the highest predicted ratings.

Heterogeneous explicit and implicit feedback provide key indicators for different versions of latent factor models [6, 17], e.g. explicit and implicit Matrix Factorization (MF), which are well-known approaches in collaborative filtering-based recommendation systems. Explicit MF models leverage the explicit feedback in the form of a user-item rating matrix in order to map both users and items into a latent space, while Implicit MF models use implicit feedback in the form of user-item interactions. It is well-known in both academia [13, 20] and industry [22] that explicit models are more effective for the rating prediction task while implicit models are well

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

suited to the ranking task. Therefore, when evaluating a system based on the construction of a ranking over all items, modeling implicit feedback is crucial [1].

Since providing explicit feedback in the form of ratings usually requires additional cognitive effort by users, the resulting explicit rating matrix is often extremely sparse. Sparsity refers to the problem that users typically rate only a small fraction of all available items, hence the observed ratings are very few. As a consequence, it is challenging to build a predictive model solely based on the observed ratings to approximate the user's preferences [6, 20]. We argue that the lower performance of explicit models in the task of ranking is due to the severe sparsity feature of explicit feedback.

The main difference between the rating prediction and ranking tasks is due to the user-item interactions that are taken into account [3, 22]. Models of rating prediction leverage only the subset of user-item interactions where a rating is observed while the ranking models leverage all user-item interactions whether the user deliberately chose to assign a rating value or not. The users of a real-world recommendation system follow 'monotonic behaviour chains' [23], i.e. the user behaviour is represented as a chain of implicit feedback (e.g. clicks), which *sometimes* leads to an explicit feedback (e.g. a rating) [3]. Once a user decides to 'stop' at a stage for a given item, the subsequent interactions will not be observed. If we represent these stages of interactions as a sequence, most of the missing values would have occurred at the *tail* of the sequence. Therefore, the dense *head* of the sequence of interactions (implicit feedback) provides valuable information in predicting the sparse tail of the sequence (explicit feedback). Yet, because of the popular nature of the *head* of this implicit feedback, recommendations learned from this data can be biased towards popularity [2], which may lack surprise or serendipity for users [19, 24].

In this paper, we propose to leverage *weak supervision at the level of data pre-processing* with the aim of unifying the tasks of ranking and rating prediction. With the aid of weak supervision, we propose to annotate some of the *missing values* in the explicit dataset by augmenting the implicit interactions. The proposed model uses the underlying explicit matrix factorization model trained on explicit feedback in order to provide a weak supervision signal to annotate the missing rating values from the head of the interactions' sequence. Therefore, the contributions of the paper are as follows:

- We bridge the gap between rating prediction and ranking tasks by providing a novel weak supervision approach for using *implicit* feedback in the *explicit* matrix factorization model. The key aspect of the proposed model is that the performance of the underlying explicit model significantly increases in the task of *ranking*. Our experiments on six datasets, covering different range of both explicit and implicit feedback and a wide range of evaluation metrics, show that the recommendation performance of the baseline explicit matrix factorization model can significantly be improved by using the less sparse weakly annotated dataset instead of the original explicit dataset.
- We show that explicit models are less vulnerable to the popularity bias than implicit models. By augmenting the explicit feedback dataset with the proposed weak supervision approach, we can increase the representation of less popular items in recommendations.
- The additional benefit of our proposed approach is that with the aid of weak supervision we can *decouple* the model of user's preferences from the recommendation model. Compared to previous studies [8, 9, 15] that proposed new *recommendation models/algorithms* to represent both explicit and implicit feedback simultaneously, our approach is unique by working at the *lower level of data pre-processing*. This is important from a practical perspective as existing deployments can retain their underlying recommendation model and leverage the proposed approach in order to improve the quality of recommendations.

The remainder of the paper is organized as follows: In Section 2, we present related work, and position our contributions in comparison to the existing literature. Section 3 presents the details of the proposed model. Our experimental setup and results are presented in Sections 4 and 5 respectively. Finally, Section 6 summarizes our findings.

2 RELATED WORK

The fundamental difference between explicit and implicit feedback has led researchers to develop different models and techniques that address each of their distinct properties. In this section, previous key related works are summarized. In Section 2.1 and Section 2.2, we review previous models, which take into account explicit and implicit feedback for the tasks of rating prediction and ranking, respectively. Finally, in Section 2.3 we describe the previous works that aimed to unify both the explicit and implicit feedback at the level of recommendation algorithm itself.

2.1 Explicit Feedback in Rating Prediction Tasks

Traditional recommendation systems often rely on collaborative filtering techniques to learn from explicit feedback (e.g. ratings). Latent factor models including the popular Matrix Factorization model are the dominant model-based approaches in this field of study, well-known from the Netflix Prize competition [6]. These models seek to accurately estimate unseen ratings by mapping both users and items into a lower-dimensional embedding space inferred from the ratings patterns. In doing so, the main goal is to generalize those previous rating patterns in a way that predicts unseen ratings. Therefore, these models are well suited to the task of rating prediction [22].

Nevertheless, users of real-world recommendation systems are reluctant to provide explicit feedback. As a result, the performance of explicit models is affected by the large proportion of missing values in the rating matrix [20]. Previous research [13, 20] revealed that unobserved ratings are Missing Not At Random (MNAR) because users of recommendation systems are more likely to provide ratings for items that they have interacted or observed before. Therefore, we investigate whether methods that augment this (non-random) missing data can enhance both the prediction and ranking tasks. Indeed, selection bias in the observed data [2] combined with data sparsity is the main barrier for adapting explicit models into the real-world rating prediction or ranking tasks [22].

2.2 Implicit Feedback for Ranking Tasks

The importance of modeling implicit feedback in ranking-based recommendation systems has been emphasized in both academia [13] and industry [22]. One of the main characteristics of latent models is that they can adapt to implicit feedback [6]. The classical matrix factorization model has been extended to incorporate the abundance of implicit feedback [4, 17]. Pairwise algorithms are popular in this context. For each user, they aim to discriminate between a relatively small set of interacted items (i.e. implicit feedback) and the large set of remaining items (i.e. the assumed negative items). The main challenge is that the interacted items can only, at best, represent positive observations. A common approach in this respect is to leverage random items as negative instances for training, an approach denoted as negative sampling [17].

Bayesian Personalized Ranking [17] is a well-known optimization framework that uses such negative sampling. Indeed, BPR is a pairwise ranking function trained based on an underlying pointwise predictive function (usually matrix factorization, denoted as $\mathrm{MF}_{\mathrm{bpr}}$), which randomly samples unclicked items as negative instances. Since BPR was introduced, many enhancements have been proposed [10, 12]. Indeed, while BPR is limited to only one type of implicit feedback, Loni et al. [10] extended the original model to incorporate multiple types of implicit feedback. The main limitation of the aforementioned approaches is that they can only model implicit feedback, while multiple types of explicit and implicit feedback are common in real-world recommendation systems and could complement each other. In the next section, we review the previous works that aimed to unify both explicit and implicit models at the level of recommendation.

2.3 Unifying Approaches

For recommendation scenarios with both implicit and explicit interaction data, it is desirable to unify both forms of users' interactions in order to generate more accurate recommendations. A line of work has emerged that incorporates both implicit feedback and explicit ratings for either ranking or rating prediction tasks. For example, ChainRec [23] – a recent approach that we use as a baseline – represents the sequence of implicit and explicit feedback as a *monotonic behaviour chain*; i.e. it is not possible to observe an explicit feedback without observing a chain of implicit feedback beforehand. Liu et al. [9] proposed a collaborative filtering model that can be simultaneously learned from both explicit and implicit feedback. Zhang et al. [25] proposed a model that learns the corresponding user and item embeddings individually for each type of feedback and integrates them to obtain a joint representation of users and items. SVD++ [5] is another work proposed for the task of rating prediction that considers all items for which a user made implicit feedback, in order to learn a representation for the user. Also, previous research [3, 7, 18] proposed training simultaneously a ranking and a rating prediction algorithm with a shared representation for users and items in a multi-task learning framework. Although the connection between explicit and implicit interactions has been well-studied in previous research [8, 9, 14, 15, 23], most approaches have focused on modifying the current recommendation models or have proposed a new model that considers other feedback as auxiliary information.

The main barrier in unifying explicit and implicit feedback is that they are heterogeneous in terms of both representations and distributions. Therefore, the identification of a single model to represent both of them simultaneously at the level of recommendation is a challenging task. Instead, in this paper, we propose to tackle the problem of unifying explicit and implicit feedback from a completely different perspective. In particular, the most important aspect of our proposed approach compared to the most related work [3, 8, 9, 23, 25] is that we tackle the problem at the level of data pre-processing: instead of suggesting a new recommendation model/algorithm that learns from both explicit and implicit feedback, we propose a weak supervision approach to augment the implicit feedback into the underlying model at the lower level of data pre-processing.

3 THE WEAK SUPERVISION MODEL

In this section, we present our proposed weak supervision approach. Section 3.1 reviews the notation used in this paper. Section 3.2 presents the baseline explicit matrix factorization model and Section 3.3 presents the proposed weak supervision approach for unifying explicit and implicit feedback.

3.1 Notation

The notations used in this paper are defined as follows: we use $U = \{u_1, u_2, ..., u_m\}$ to denote the set of m users and $I = \{i_1, i_2, ..., i_n\}$ to denote the set of n items. Both explicit and implicit feedback are observable from the interactions, i.e. from $U \times I$. The explicit feedback dataset is defined as $D_e = \langle U, I, R\rangle$ where $r \in R$ is usually in the form of a numeric rating assigned to item $i \in I$ by user $u \in U$[1]. On the other hand, the implicit feedback dataset is defined as $D_i = \langle U, I\rangle$, which indicates that user $u \in U$ interacted with item $i \in I$. The weakly annotated dataset is defined as $D_i^* = \langle U, I, \hat{R}\rangle$ where each $r \in \hat{R}$ is a predicted rating value provided by a weakly supervised signal. We use the notation Φ_{D_e} and $\Phi_{D_i^*}$ to denote the same model (Φ) – e.g. matrix factorization – that is trained on the explicit feedback dataset (D_e) and the weakly annotated dataset (D_i^*), respectively.

3.2 Explicit Recommendation Model

Most explicit recommendation algorithms are based on designing either a parametric or a nonparametric form of a scoring function $\hat{r} = f(u, i)$ that estimates the rating of item i by user u. Usually it takes the form of a regression model that aims to fit the parameters of the function $f(u, i)$ with

[1] Typically $R = \{\langle u_i, i_k, r\rangle ...\}$ is a set of all provided rating values, where r is the rating value.

the explicit rating values $r \in R$. We leverage the matrix factorization model [6] that represents both users and items as latent vectors denoted by $\vec{u}$ and $\vec{i}$; the predicted score for a specific user-item pair ($\hat{r}$) is given by the dot product of the user and item latent vectors:

$$\hat{r} = f(u,i) = \vec{u} \cdot \vec{i} = \sum_{k=1}^{d} u_k \times i_k \tag{1}$$

where d is the size of the latent space (also known as the embedding size). Typically, the whole model is trained with Stochastic Gradient Descent and the factor matrices U and I are optimized by solving the following equation:

$$\Phi = \operatorname*{arg\,min}_{U,I} \mathcal{L}(U,I) + \alpha \cdot \Omega(U,I) \tag{2}$$

where the loss function $\mathcal{L}(\cdot,\cdot)$ quantifies how good the approximation is; Ω is the regularization term, usually parameterized using least square errors (L_2 norm); and $\alpha \in [0,\infty]$ is a hyperparameter that weights the relative contribution of the norm regularisation term.

In particular, for explicit feedback, we want the factorization model to approximate well the observed ratings in D_e. It is common to leverage the widely used Mean Square Error as the loss function:

$$\mathcal{L}(U,I) = \frac{1}{n} \sum (r - \hat{r})^2 \tag{3}$$

where n is the number of data points, r is the actual rating value and $\hat{r}$ is the predicted rating value for user-item interaction $\langle u,i \rangle \in \langle U,I \rangle$[2]. In particular, we can learn a model based on explicit feedback D_e using the general model described in Equation (2):

$$\Phi_{D_e} = \operatorname*{arg\,min}_{U,I,R \in D_e} \mathcal{L}(U,I) + \alpha \cdot \Omega(U,I) \tag{4}$$

where $\mathcal{L}(\cdot,\cdot)$ is the loss function described in Equation (3) and $\Omega(\cdot,\cdot)$ is the L_2 regularization term.

As mentioned before, the Explicit Matrix Factorization model is used for the task of rating prediction [6, 22], using explicit feedback interaction, which we denote as Φ_{D_e}. Indeed, because there are no rating values in D_i, it is impossible to train explicit matrix factorization on implicit data – i.e. Φ_{D_i} cannot exist. On the other hand, previous studies [1, 2, 20] have emphasized the importance of modeling implicit feedback as part of the recommendation model. In the following section, we leverage the explicit model described in Equation (4) as a weakly supervised signal and augment it with the implicit feedback at the level of data pre-processing.

3.3 Unifying Explicit and Implicit Feedback Datasets

The explicit feedback dataset is represented as $D_e = \langle U,I,R \rangle$, which contains the explicit users' preferences, while the implicit feedback dataset is represented as $D_i = \langle U,I \rangle$, where the rating values are missing [14, 20]. We do not assume any further constraints regarding the explicit and implicit datasets. Therefore, as we show in our experiments, our approach can be applied to a broad range of real-world datasets.

```
Input: $D_e = \langle U,I,R \rangle$; $D_i = \langle U,I \rangle$
Output: $D_i^* = \langle U,I,\hat{R} \rangle$
initialize $D_i^* \leftarrow \emptyset$ ;
//train the baseline model $\Phi_{D_e}$;
do
    foreach $\langle u,i,r \rangle \in D_e^{train}$ do
        optimize $\Phi_{D_e}$ based on Equation (4);
    end
    evaluate $\Phi_{D_e}$ on the validation set $D_e^{\text{valid}}$;
while not converge;
//annotate the implicit dataset based on $\Phi_{D_e}$;
foreach $\langle u,i \rangle \in D_i$ do
    $\hat{r} \leftarrow \Phi_{D_e}(u,i)$;
    $D_i^* \leftarrow D_i^* \cup \{\langle u,i,\hat{r} \rangle\}$;
end
```

Algorithm 1: Generates the weakly annotated dataset D_i^* based on the trained model on the explicit dataset Φ_{D_e} as the weak supervision signal.

Our aim is to leverage weak supervision in order to transfer the knowledge of the explicit feedback dataset and adapt it to the abundant implicit feedback. Therefore, after training the base model Φ_{D_e} (described in Equation (4)), we can leverage the predicted ratings by this model in order to weakly annotate all interactions in the implicit feedback dataset and build a new weakly annotated dataset:

$$D_i^* = \langle U,I,\hat{R} \rangle \tag{5}$$

where $\hat{r} \in \hat{R}$ is the predicted rating value for a specific user's interaction $\langle u,i \rangle \in \langle U,I \rangle$ based on the explicit model Φ_{D_e}:

$$\hat{r} = \Phi_{D_e}(u,i) \tag{6}$$

The full process of generating the weakly annotated dataset D_i^* is presented in Algorithm 1. The algorithm takes as input the explicit and implicit datasets and generates the new weakly annotated dataset D_i^* unifying both explicit and implicit feedback. The effectiveness of weak supervision as a special case of transfer learning approaches depends on the similarity between the user interactions within the labeled and unlabeled datasets - i.e. D_e and D_i.

After generating the new weakly annotated dataset D_i^* as described in Algorithm 1, we leverage it to train exactly the *same model* as used in Equation (4) but with the new training dataset:

$$\Phi_{D_i^*} = \operatorname*{arg\,min}_{U,I,\hat{R} \in D_i^*} \mathcal{L}(U,I) + \alpha \cdot \Omega(U,I) \tag{7}$$

It is important that both models Φ_{D_e} and $\Phi_{D_i^*}$ in Equation (4) and Equation (7) use the *same* architecture and hyperparameters but *different* training sets. To demonstrate the benefit of this proposed approach for both the rating prediction and ranking tasks, we aim to answer the following two research questions:

[2] Both user and item bias vectors are a part of loss function as an additional variable.

Research Question 1. *Does the proposed weak supervision approach enhance recommendation accuracy for (a) rating prediction and (b) ranking tasks?*

It is known in the literature [13, 20] that explicit models do not perform well for ranking tasks, while the implicit models (e.g. BPR [17]) are effective. However, our explicit matrix factorization ($\Phi_{D_i^*}$) is instantiated using both the users' explicit rating values and the estimated missing values predicted by Φ_{D_e}. Therefore, the proposed model trained on the weakly annotated dataset ($\Phi_{D_i^*}$) is hypothesised to perform better than the original model trained on the explicit dataset (Φ_{D_e}). Consequently, we investigate the viability of weak supervision for both rating prediction and ranking tasks.

Research Question 2. *How much popularity bias is exhibited by our proposed weak supervision approach compared with the other baseline recommendation approaches?*

Indeed, recommending serendipitous items from the long tail is generally considered to be a key function of a recommendation system. Previous research [2, 21, 24] has revealed that collaborative filtering algorithms typically emphasize popular items over long-tail items. In this research question, we address the capability of the proposed model in suggesting unpopular items compared to the baseline explicit and implicit models. In the following, we investigate the above research questions based on common benchmark datasets. In particular, the next section describes our experimental setup. The observed experimental results and a corresponding discussion follow in Section 5.

4 EXPERIMENTAL SETUP

We aim to determine whether the quality of recommendations can be improved by unifying the implicit and explicit feedback into a weakly annotated dataset. In particular, we aim to answer the two research questions mentioned in Section 3. Therefore, we leverage the matrix factorization [6] as a widely used, robust recommendation algorithm in our experiments and evaluate the proposed weak supervision approach on six real-world datasets, where both explicit and implicit interactions are available.

4.1 Datasets and Evaluation Metrics

We consider six public datasets[3], which cover different types of user's behaviour in the form of both explicit and implicit feedback and vary markedly in data sparsity:

GoodReads. [23] is a large dataset from a popular book review website. The original collection contains 229,154,523 feedback from 876,145 users and 2,360,655 items. Different categorical subsets have been provided for research experiments[4]. In our experiments, we leverage the first three subsets provided by the original authors ('Children', 'Comics & Graphic' and 'Fantasy & Paranormal'). In addition, we leverage 'recommend/rate' as the explicit feedback and 'read' as the implicit feedback in our experiments.

Steam. [16] dataset contains the interactions of 24,110 Australian users on the Steam video game distribution network (8,696 video games). This dataset includes purchase information, play time, reviews, and recommends [23]. We leverage 'recommend' as the explicit feedback and 'play' as the implicit feedback in our experiments.

Douban. [8] dataset contains the interactions of 12,770 users and 22,002 items from another popular book review website. In our experiments, we leverage 'recommend' as the explicit feedback and 'reading', 'read', 'tag' and 'comment' as the implicit feedback.

Dianping. [8] dataset contains 10,549 users and 17,707 items from a restaurant review website. In our experiments, we use the overall ranking as the explicit feedback and 'taste', 'environment' and 'service' as the implicit feedback.

We follow a uniform preprocessing step for all datasets. For the Steam dataset, explicit feedback (i.e. recommend) is a binary variable, while the corresponding value for the GoodReads, Douban and Dianping datasets is an integer $r \in [0, 5]$. Our aim is to predict the most relevant items for each user, therefore, we binarize all explicit rating values by keeping the highly recommended items ($r \geq 4$). We also filter all users with less than 5 explicit interactions. Finally, for each user, we randomly split the explicit feedback, D_e, into training, validating and testing subsets, keeping 80% for training (denoted D_e^{train}), 10% for validation (denoted D_e^{valid}) and 10% for testing (D_e^{test}).

The task of rating prediction is measured by comparing the predicted ratings with the ground truth; metrics such as Root Mean Square Error (RMSE) are useful in this context. On the other hand, a ranking task is usually measured based on Information Retrieval (IR) metrics including Mean Reciprocal Rank (MRR) and normalized Discounted Cumulative Gain (nDCG) [1, 2]. We evaluate the performance of the proposed model on both the rating prediction and ranking tasks. Therefore, we rank items based on the *actual* rating score r from the explicit test subset (D_e^{test}), and thereby calculate both RMSE and rank-based measures, namely MRR, nDCG and MAP, as the evaluation measures [1, 2]. Significance testing of differences between ranking measure performances are measured using the paired t-test ($p < 0.01$)[5]. Finally, in order to evaluate the effectiveness of models in mitigating popularity bias and covering long tail items we use Average Recommendation Popularity, namely ARP [24]:

$$ARP = \frac{1}{|U|} \sum_{u \in U} \frac{\sum_{i \in L_u} \phi(i)}{|L_u|} \tag{8}$$

where $\phi(i)$ is the number of times item i has been rated in the explicit training set (D_e^{train}). L_u is the recommended list of items for user u and $|U|$ is the total number of users in the explicit test set (D_e^{test}). Lower values of ARP indicate the recommendation of less popular items.

[3] As explained below, three of these datasets represent subsets of the larger GoodReads dataset.

[4] https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home

[5] For the rating prediction task, as RMSE is a non-linear aggregation of squared absolute error, a significance test cannot be conducted.

4.2 Recommendation Models

We consider four different groups of recommendation models for comparisons:

Popularity model. we count the observed interactions in the training set D_e^{train} and rank items based on their popularity, denoted as Item_{pop}.

Explicit model. our baseline model is the explicit model trained on the explicit feedback dataset D_e^{train}, denoted as Φ_{D_e}. This model is trained based on Equation (4).

Implicit model. In order to compare the effectiveness of the proposed approach with the models that leverage only the implicit feedback dataset, we compare the performance of the proposed model with Bayesian Personalized Ranking [17] (denoted MF_{bpr}), which is a robust pairwise ranking model for implicit feedback. The BPR model is trained based on a uniform negative sampling, i.e. we randomly sample items not interacted with as negative instances for each user.

Unifying model. In order to compare the effectiveness of the proposed approach with the models that leverage both the explicit and implicit feedback, we compare the performance of the proposed model with [23] (denoted ChainRec). We leverage the provided source code[6] to reproduce the results on our datasets.

The baseline explicit matrix factorization model (described in Equation (4) and denoted as Φ_{D_e}) is used to estimate the missing rating values in the implicit feedback dataset D_i (described in Equation (7) and denoted as $\Phi_{D_i^*}$). Note that both models leverage the *same architecture and hyperparameters* but different input data. Φ_{D_e} is trained with the explicit dataset D_e while $\Phi_{D_i^*}$ is trained with the weakly annotated dataset D_i^*. For reproducibility, our code and the used datasets are available from https://github.com/amirj/unifying_explicit_implicit.

5 RESULTS AND DISCUSSION

Table 1 shows the obtained performances of the models on our used datasets. For each dataset, we compare the performance of the proposed model trained on the weakly annotated dataset (denoted as $\Phi_{D_i^*}$) with the popularity baseline (denoted as Item_{pop}), the baseline explicit and implicit models (denoted as Φ_{D_e} and MF_{bpr} respectively) and a recently proposed hybrid model [23] that unifies both explicit and implicit feedback (denoted as ChainRec). All models are compared on both the rating prediction and ranking tasks, using RMSE and a set of rank-based measures, respectively. In the following, we analyse Table 1 with respect to the two research questions stated in Section 3.3, concerning recommendation effectiveness (Section 5.1) and popularity bias (Section 5.2).

5.1 RQ 1. Recommendation Effectiveness

On analysing Table 1, we firstly note the low effectiveness of the baseline explicit matrix factorization Φ_{D_e} model trained upon the explicit rating matrix D_e. Indeed, Φ_{D_e} is trained to approximate the observed rating values, hence this model performs best in the rating prediction task – in terms of the RMSE – while its performance on the ranking task is very low. On the other hand, given that the implicit model MF_{bpr} is optimized for pairwise ranking, it can be expected that such a model performs better than the explicit baseline model (Φ_{D_e}) on the ranking task. In addition, ChainRec exhibits performance close to MF_{bpr}, as might be expected from the original paper [23] – ChainRec is inherently optimized based on pairwise preferences while leveraging both explicit and implicit feedback to select more informative pairs.

Comparing the performances of $\Phi_{D_i^*}$ and Φ_{D_e} across all datasets in Table 1, we firstly consider the rating prediction task. This reveals that the performance of the model trained on the weakly annotated dataset ($\Phi_{D_i^*}$) is comparable in terms of RMSE to the *same model* trained on the explicit feedback dataset (i.e. Φ_{D_e}). This answers research question RQ1(a).

On the other hand, for the ranking task, $\Phi_{D_i^*}$ significantly outperforms Φ_{D_e} (paired t-test, $p < 0.05$) for 5 out of 6 datasets across the three ranking metrics. This shows that once the sparsity problem and the large number of missing values in the explicit feedback has been addressed through weak supervision, an explicit recommendation approach can be more effective at ranking. This answers research question RQ1(b), demonstrating that weak supervision can have a positive impact.

However, considering the significance tests for nDCG@20, we note that for two out of the six datasets (namely Children, Fantasy & Paranormal), the performance of the proposed model is lower than models that are directly optimized for the ranking task (i.e. MF_{bpr}). Note however that for two of the six datasets (namely Comics & Graphics and Douban) the performances of the proposed model $\Phi_{D_i^*}$ is higher than MF_{bpr}. Finally, for the other two datasets (Steam and Dianping), the performances of the proposed model is comparable to and statistically indistinguishable from MF_{bpr} ($p > 0.05$). Further analysis of Table 1 reveals that the performances of MF_{bpr} is very close to Item_{pop}. Indeed, the simple item popularity performs very well on most datasets. This can be explained that in most datasets, item engagements are driven by item popularity, as evidenced in the effectiveness of Item_{pop} on the ranking metrics. In the following, we further analyze the occurrence of popular items in the recommendations.

5.2 RQ 2. Popularity Bias

The main role of a recommendation system is to help users discover items they might otherwise not have found [19]. Recommending serendipitous items from the long-tail is generally considered to be a key function of any recommendation, as these are items that users are less likely to know about [24]. Figure 1 shows the proportion of top-20 most popular items (i.e. items with the highest number of implicit interactions) in the top 20 ranked recommendations of the proposed model

[6] https://github.com/MengtingWan/chainRec

Table 1: Results on six different datasets, evaluated based on user's most explicit feedback (D_e^{test}). The best performance for each metric are shown in bold. Arrows denote which direction represents improvement. For the ranking metrics, $^{1/2/3/4/5}$ denote a significant difference according to the paired t-test ($p < 0.05$) compared to the indicated approach. Since ChainRec [23] leverages a custom data split, statistical tests cannot be conducted.

DataSets	Algorithms	Rating Prediction	Ranking Prediction			Popularity Bias
		RMSE ↓	MRR ↑	nDCG@20 ↑	MAP ↑	ARP@20 ↓
GoodReads (Children)	^{1} Item$_{\text{pop}}$	-	$0.1122^{3,4,5}$	$0.1067^{3,4,5}$	$0.0605^{3,4,5}$	49,412
	^{2} ChainRec	10.7856	0.0625	0.0962	0.0625	31,616
	^{3} MF$_{\text{bpr}}$	4.0915	$\mathbf{0.1127}^{1,4,5}$	$\mathbf{0.1072}^{1,4,5}$	$\mathbf{0.0606}^{1,4,5}$	48,913
	^{4} Φ_{D_e}	**0.4063**	$0.0271^{1,3,5}$	$0.0346^{1,3,5}$	$0.0141^{1,3,5}$	13,481
	^{5} $\Phi_{D_i^*}$	0.4160	$0.0833^{1,3,4}$	$0.0726^{1,3,4}$	$0.0428^{1,3,4}$	**12,712**
GoodReads (Comics & Graphics)	^{1} Item$_{\text{pop}}$	-	$0.0554^{3,4,5}$	$0.0458^{3,4,5}$	$0.0233^{3,4,5}$	9,568
	^{2} ChainRec	9.8613	0.0290	0.0503	0.0290	6,673
	^{3} MF$_{\text{bpr}}$	3.6611	$0.0637^{1,4,5}$	$0.0522^{1,4,5}$	$0.0256^{1,4,5}$	8,527
	^{4} Φ_{D_e}	**0.3775**	$0.0381^{1,3,5}$	$0.0443^{1,3,5}$	$0.0189^{1,3,5}$	2,574
	^{5} $\Phi_{D_i^*}$	0.3903	$\mathbf{0.0773}^{1,3,4}$	$\mathbf{0.0663}^{1,3,4}$	$\mathbf{0.0386}^{1,3,4}$	**2,438**
GoodReads (Fantasy & Paranormal)	^{1} Item$_{\text{pop}}$	-	$0.1143^{3,4,5}$	$0.0981^{3,4,5}$	$0.0527^{3,4,5}$	227,074
	^{2} ChainRec	13.1535	0.0541	0.0854	0.0541	109,598
	^{3} MF$_{\text{bpr}}$	3.3754	$\mathbf{0.1145}^{1,4,5}$	$\mathbf{0.0982}^{1,4,5}$	$\mathbf{0.0528}^{1,4,5}$	226,123
	^{4} Φ_{D_e}	**0.4077**	$0.0258^{1,3,5}$	$0.0230^{1,3,5}$	$0.0106^{1,3,5}$	**23,894**
	^{5} $\Phi_{D_i^*}$	0.4141	$0.0247^{1,3,4}$	$0.0208^{1,3,4}$	$0.0101^{1,3,4}$	24,515
Steam	^{1} Item$_{\text{pop}}$	-	$\mathbf{0.1126}^{3,4,5}$	$\mathbf{0.1296}^{3,4,5}$	$\mathbf{0.0871}^{3,4,5}$	239
	^{2} ChainRec	6.9688	0.0854	0.1174	0.0854	160
	^{3} MF$_{\text{bpr}}$	7.5434	$0.0793^{1,4}$	$0.0936^{1,4}$	$0.0581^{1,4,5}$	183
	^{4} Φ_{D_e}	0.4613	$0.0391^{1,3,5}$	$0.0698^{1,3,5}$	$0.0343^{1,3,5}$	**112**
	^{5} $\Phi_{D_i^*}$	**0.4590**	$0.0460^{1,4}$	$0.0661^{1,4}$	$0.0382^{1,3,4}$	118
Douban	^{1} Item$_{\text{pop}}$	-	0.0215^{4}	$0.0253^{3,4,5}$	$0.0134^{3,4}$	91
	^{2} ChainRec	6.2433	0.0065	0.0146	0.0065	63
	^{3} MF$_{\text{bpr}}$	4.2549	0.0200^{4}	$0.0239^{1,4,5}$	$0.0119^{1,4}$	76
	^{4} Φ_{D_e}	**0.3104**	$0.0034^{1,3,5}$	$0.0059^{1,3,5}$	$0.0018^{1,3,5}$	**18**
	^{5} $\Phi_{D_i^*}$	0.3271	$\mathbf{0.0223}^{4}$	$\mathbf{0.0345}^{1,3,4}$	$\mathbf{0.0149}^{4}$	21
Dianping	^{1} Item$_{\text{pop}}$	-	$0.0458^{3,4,5}$	$0.0490^{3,4}$	$0.0268^{3,4}$	108
	^{2} ChainRec	3.8796	0.0207	0.0372	0.0207	87
	^{3} MF$_{\text{bpr}}$	4.0115	$\mathbf{0.0559}^{1,4,5}$	$0.0586^{1,4}$	$\mathbf{0.0337}^{1,4,5}$	83
	^{4} Φ_{D_e}	**0.4332**	$0.0102^{1,3,5}$	$0.0204^{1,3,5}$	$0.0067^{1,3,5}$	34
	^{5} $\Phi_{D_i^*}$	0.4446	$0.0309^{1,3,4}$	$\mathbf{0.0609}^{4}$	$0.0228^{3,4}$	**17**

($\Phi_{D_i^*}$), the baseline explicit model (Φ_{D_e}) and MF$_{\text{bpr}}$. Interestingly, we observe that, on average, more than 70% of the suggested items by MF$_{\text{bpr}}$ are in the top-20 most popular items. This explains why the overall performance of MF$_{\text{bpr}}$ is very close to the popularity model (Item$_{\text{pop}}$) in Table 1.

Comparing the performances of the proposed model, based on the Average Recommendation Popularity (ARP@20) metric across all datasets, reveals that the explicit models (Φ_{D_e} and $\Phi_{D_i^*}$) are far better (lower) than the implicit models (and ChainRec, which is based on pairwise preferences) in recommending less popular, long-tail items. The main reason is that implicit models are constructed based on pairwise preferences (u, i, j), treating each interacted item i as a positive item and a random item j as a negative item. If the user has been exposed with only popular items[7], eventually all the positive interacted items i are sampled from the popular group. As a result, the trained model will be biased to favour popular items over other items irrespective of the user's preferences. On the other hand, in explicit models, for each interacted item, we have a rating value r that *explicitly* indicates whether the user likes the item or not. Therefore, in this situation, each user has a chance to dislike the suggested items even if they are sampled from the popular group.

In conclusion, with respect to research question RQ2, from the results in Table 1 and Figure 1, we observe that our

[7] This is a prevalent assumption in real-world recommendation systems [2, 21, 24].

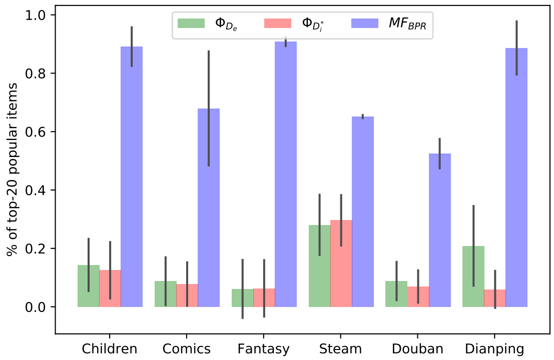

Figure 1: The percentage of top-20 most popular items in the recommended list of each model for six different datasets.

proposed model that exhibits less bias towards popular items than both the implicit and explicit baseline models, as well as ChainRec.

6 CONCLUSIONS

In this paper, we presented a novel weak supervision approach that bridges the gap between the rating prediction and ranking tasks in collaborative filtering recommendation systems while alleviating the bias against less popular long-tail items. In addition, the proposed approach is applied at the level of data pre-processing rather than the recommendation model, as a *weak supervision signal*.

Indeed, our method leveraged the baseline matrix factorization model trained on the explicit feedback dataset in order to provide a weak supervision signal that unifies both the explicit and implicit feedback into a combined dataset. Our experimental results on both rating prediction and ranking tasks using a wide range of datasets and evaluation metrics clearly demonstrated the usefulness of the weak supervision signal in predicting the missing values in the user-item rating matrix. In particular, our experiments revealed that with the aid of weak supervision, the performance of the classical explicit matrix factorization approach can be significantly improved on the ranking tasks, but ultimately it does not outperform the classical BPR approach, which itself is shown to be biased towards popular items, explaining BPR's high performance. As a future work, we will consider how to combine weak supervision into pairwise approaches such as BPR or more recent neural network-based recommenders such as [11]. Moreover, encapsulating an appropriate balance between good quality recommendations and popularity-bias requires more investigation.

ACKNOWLEDGMENTS

The authors acknowledge support from EPSRC grant EP/R018634/1 entitled Closed-Loop Data Science for Complex, Computationally- and Data-Intensive Analytics.

REFERENCES

[1] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2017. Statistical biases in Information Retrieval metrics for recommender systems. *Information Retrieval Journal* 20, 6 (2017), 606–634.

[2] Rocío Cañamares and Pablo . 2018. Should I Follow the Crowd?: A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems. In *Proceedings of SIGIR*. 415–424.

[3] Guy Hadash, Oren Sar Shalom, and Rita Osadchy. 2018. Rank and Rate: Multi-task Learning for Recommender Systems. In *Proceedings of RecSys*. 451–454.

[4] Y. Hu, Y. Koren, and C. Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of ICDM*. 263–272.

[5] Yehuda Koren. 2008. Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. In *Proceedings of KDD*. 426–434.

[6] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.

[7] Gai Li, Zhiqiang Zhang, Liyang Wang, Qiang Chen, and Jincai Pan. 2017. One-class Collaborative Filtering Based on Rating Prediction and Ranking Prediction. *Know.-Based Syst.* 124, C (2017), 46–54.

[8] Jian Liu, Chuan Shi, Binbin Hu, Shenghua Liu, and Philip S. Yu. 2017. Personalized Ranking Recommendation via Integrating Multiple Feedbacks. In *Proceedings of PAKDD*. 131–143.

[9] Nathan N. Liu, Evan W. Xiang, Min Zhao, and Qiang Yang. 2010. Unifying Explicit and Implicit Feedback for Collaborative Filtering. In *Proceedings of CIKM*. 1445–1448.

[10] Babak Loni, Roberto Pagano, Martha Larson, and Alan Hanjalic. 2016. Bayesian Personalized Ranking with Multi-Channel User Feedback. In *Proceedings of RecSys*. 361–364.

[11] Jarana Manotumruksa, Craig Macdonald, and Iadh Ounis. 2017. A Deep Recurrent Collaborative Filtering Framework for Venue Recommendation. In *Proceedings of CIKM*. 1429–1438.

[12] Jarana Manotumruksa, Craig Macdonald, and Iadh Ounis. 2017. A Personalised Ranking Framework with Multiple Sampling Criteria for Venue Recommendation. In *Proceedings of CIKM*. 1469–1478.

[13] Benjamin M. Marlin and Richard S. Zemel. 2009. Collaborative Prediction and Ranking with Non-random Missing Data. In *Proceedings of RecSys*. 5–12.

[14] Weike Pan, Nathan N. Liu, Evan W. Xiang, and Qiang Yang. 2011. Transfer Learning to Predict Missing Ratings via Heterogeneous User Feedbacks. In *Proceedings of IJCAI*. 2318–2323.

[15] Denis Parra, Alexandros Karatzoglou, Idil Yavuz, and Xavier Amatriain. 2011. Implicit feedback recommendation via implicit-to-explicit ordinal logistic regression mapping. In *Proceedings of CARS*.

[16] Apurva Pathak, Kshitiz Gupta, and Julian McAuley. 2017. Generating and Personalizing Bundle Recommendations on Steam. In *Proceedings of SIGIR*. 1073–1076.

[17] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of UAI*. 452–461.

[18] Yue Shi, Martha Larson, and Alan Hanjalic. 2013. Unifying Rating-oriented and Ranking-oriented Collaborative Filtering for Improved Recommendation. *Inf. Sci.* 229 (2013), 29–39.

[19] B. Smith and G. Linden. 2017. Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing* 21, 3 (2017), 12–18.

[20] Harald Steck. 2010. Training and Testing of Recommender Systems on Data Missing Not at Random. In *Proceedings of KDD*. 713–722.

[21] Harald Steck. 2011. Item Popularity and Recommendation Accuracy. In *Proceedings of RecSys*. 125–132.

[22] Harald Steck. 2013. Evaluation of Recommendations: Rating-prediction and Ranking. In *Proceedings of RecSys*. 213–220.

[23] Mengting Wan and Julian McAuley. 2018. Item Recommendation on Monotonic Behavior Chains. In *Proceedings of RecSys*. 86–94.

[24] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the Long Tail Recommendation. *Proc. VLDB Endow.* 5, 9 (2012), 896–907.

[25] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W. Bruce Croft. 2017. Joint Representation Learning for Top-N Recommendation with Heterogeneous Information Sources. In *Proceedings of CIKM*. 1449–1458.