Regularizing Matrix Factorization with User and Item Embeddings for Recommendation

Thanh Tran, Kyumin Lee
Worcester Polytechnic Institute, USA
{tdtran,kmlee}@wpi.edu

Yiming Liao, Dongwon Lee
Penn State University, USA
{yiming,dongwon}@psu.edu

ABSTRACT
Following recent successes in exploiting both latent factor and word embedding models in recommendation, we propose a novel Regularized Multi-Embedding (RME) based recommendation model that simultaneously encapsulates the following ideas via decomposition: (1) which items a user likes, (2) which two users co-like the same items, (3) which two items users often co-disliked, and (4) which two items users often co-disliked. In experimental validation, the RME outperforms competing state-of-the-art models in both explicit and implicit feedback datasets, significantly improving Recall@5 by 5.9–7.0%, NDCCG@20 by 4.3–5.6%, and MAP@10 by 7.9–8.9%. In addition, under the cold-start scenario for users with the lowest number of interactions, against the competing models, the RME outperforms NDCCG@5 by 20.2% and 29.4% in MovieLens-10M and MovieLens-20M datasets, respectively. Our datasets and source code are available at: https://github.com/thanhdtran/RME.git.

KEYWORDS
Recommendation; item embeddings; user embeddings; negative sampling; collaborative filtering.

ACM Reference Format:

1 INTRODUCTION
Among popular Collaborative Filtering (CF) methods in recommendation [14, 17, 29, 33], in recent years, latent factor models (LFM) using matrix factorization have been widely used. LFM are known to yield relatively high prediction accuracy, are language independent, and allow additional side information to be easily incorporated and decomposed together [1, 35]. However, most of conventional LFM only exploited positive feedback while neglected negative feedback and treated them as missing data [8, 14, 27, 34].

In movie recommender systems, it was observed that many users who enjoyed watching Thor: The Dark World, also enjoyed Thor: Ragnarok. In this case, Thor: The Dark World and Thor: Ragnarok can be seen as a pair of co-liked movies. So, if a user preferred Thor: The Dark World but never watch Thor: Ragnarok, the system can precisely recommend Thor: Ragnarok to her (first observation). Similarly, if two users A and B liked the same movies, we can assume A and B have the same movie interests. If user A likes a movie that B has never watched, the system can recommend the movie to B (second observation). In the same manner, we ask if co-occurred disliked movies can provide any meaningful information. We observed that most users, who rated Pledge This! poorly (0.8/5.0 on average), also gave a low rating to Run for Your Wife (1.3/5.0 on average). If the disliked co-occurrence pattern was exploited, Run for Your Wife would not be recommended to other users who did not enjoy Pledge This! (third observation). This will help reduce the false positive rate for recommender systems. The same phenomena would have also occurred in other recommendation domains.

The first two observations are similar to the basic assumptions of item CF and user CF where similar scores between items/users are used to infer the next recommended items for users. Unfortunately, only the first two observations have been exploited in conventional CF. While treating the negative-feedback items differently from missing data led to better results [13], to the best of our knowledge, no previous works exploited the third observation to enhance the recommender systems’ performance.

Therefore, in this paper, we attempt to exploit all three observations in one model to achieve better recommendation results. With the recent success of word embedding techniques in natural language processing, if we consider pairs of co-occurred liked/disliked items or pairs of co-occurred users as pairs of co-occurred words, we can apply word embedding to learn latent representations of items (e.g., item embeddings) and users (e.g. user embeddings). Based on this, we propose a Regularized Multi-Embedding based...
recommendation model (RME), which jointly decomposes (1) a user-item interaction matrix, (2) a user co-occurrence matrix, (3) a co-liked item co-occurrence matrix, and (4) a co-disliked item co-occurrence matrix. The RME model concurrently exploits the co-liked co-occurrence patterns and co-disliked co-occurrence patterns of items to enrich the items’ latent factors. It also augments users’ latent factors by incorporating user co-occurrence patterns on their preferred items. Figure 1 illustrates an overview of our RME model.

Both liked and disliked items can be explicitly measured by rating scores (e.g., a liked item is ≥ 4 star-rating and a disliked item is ≤ 2 star-rating) in explicit feedback datasets such as 5-star rating datasets (e.g., a Movie dataset and an Amazon dataset). However, in implicit feedback datasets (e.g., a music listening dataset and a browsing history dataset), users do not explicitly express their preferences. In implicit feedback datasets, the song plays and URL clicks could indicate how much users like the items (i.e., positive samples), but inferring the disliked items (i.e., negative samples) is a big challenge due to the nature of implicit feedback. In order to deal with this challenge, we propose an algorithm which infers a user’s disliked items in implicit feedback datasets, so that we can build an RME model and recommend items for both explicit and implicit feedback datasets. In this paper, we made the following contributions:

- We proposed a joint RME model, which combined weighted matrix factorization, co-liked item embedding, co-disliked item embedding, and user embedding, for both explicit and implicit feedback datasets.
- We designed a user-oriented EM-like algorithm to draw negative samples (i.e., disliked items) from implicit feedback datasets.
- We conducted comprehensive experiments and showed that the RME model substantially outperformed several baseline models in both explicit and implicit feedback datasets.

### 2 PRELIMINARIES

**Item.** Items are objects that users interact with or consume. They can be interpreted in various ways, depending on the context of a dataset. For example, an item is a movie in a movie dataset such as MovieLens, whereas it is a song in TasteProfile.

**Liked items and disliked items.** In explicit feedback datasets such as MovieLens (a 5-star rating dataset), an item ≥ 4 stars is classified to a liked item of the user, and an item ≤ 2 stars is classified to a disliked item of the user [5]. In implicit feedback datasets such as TasteProfile, the more a user consumes an item, the more he/she likes it (e.g., larger play count in TasteProfile indicates stronger preference). But, disliked items are not explicitly observable.

**Top-N recommendation.** In this paper, we focus on top-N recommendation scenario, in which a recommendation model suggests a list of top-N most appealing items to users. We represent the interactions between users and items by a matrix $M^{m \times n}$ where $m$ is the number of users and $n$ is the number of items. If a user $u$ likes an item $p$, $M_{up}$ will be set to 1. From $M$, we are interested in extracting co-occurrence patterns including liked item co-occurrences, disliked item co-occurrences, and user co-occurrences. Our goal is to exploit those co-occurrence information to learn the latent representations of users and items, then recommend top-N items to the users.

### Notations.** Table 1 shows key notations used in this paper. Note that all vectors in the paper are column vectors.

### 3 OUR RME MODEL

First, we review the Weighted Matrix Factorization (WMF), and co-liked item embedding. Then, we propose co-disliked item embedding and user embedding. Finally, we describe our RME model and present how to compute it.

#### 3.1 WMF, Embedding and RME model

**Weighted matrix factorization (WMF).** WMF is a widely-used collaborative filtering method in recommender systems [14]. Given a sparse user-item matrix $M^{m \times n}$, the basic idea of WMF is to decompose $M$ into a product of 2 low rank matrices $U^{m \times k}$ and $P^{n \times k}$ (i.e., $M = U \times P^T$), where $k$ is the number of dimensions and $k < \min(m, n)$. Here, $U$ is interpreted as a latent factor matrix of users, and $P$ is interpreted as a latent factor matrix of items.

We denote $U^T = (a_1, a_2, ..., a_m)$ where $a_i \in R^k$ $(u \in U, m)$ and $a_u$ represents the latent factor vector of user $u$. Similarly, we denote $P^T = (\beta_1, \beta_2, ..., \beta_n)$ where $\beta_p \in R^k$ $(p \in P, n)$ and $\beta_p$ represents the latent factor vector of item $p$. The objective of WMF is defined by:

$$L_{WMF} = \frac{1}{2} \sum_{u,p} w_{up} (M_{up} - a_u^T \beta_p)^2 + \frac{1}{2} \left[ \lambda_u \sum_u ||a_u||^2 + \lambda_p \sum_p ||\beta_p||^2 \right]$$

(1)

where $w_{up}$ is a hyperparameter to compensate the interaction between user $u$ and item $p$, and is used to balance between the number of non-zero and zero values in a sparse user-item matrix. The weight $w$ of the interaction between user $u$ and item $p$ (denoted as $w_{up}$) can be set as $w_{up} = I(1 + \phi M_{up})$ [14, 20] where $I$ is a relative scale and $\phi$ is a constant. $\lambda_u$ and $\lambda_p$ are used to adjust the importance of two quadratic regularization terms $\sum_u ||a_u||^2$ and $\sum_p ||\beta_p||^2$.

**Word embedding models.** Word embedding models have recently received a lot of attention from the research community. Given a
sequence of training words, the embedding models learn a latent representation for each word. For example, word2vec \cite{mikolov2013efficient} is one of popular word embedding methods. Especially, the skip-gram model in word2vec tries to predict surrounding words (i.e., word context) of a given word in the training set.

According to Levy et al. \cite{levy2014dependency}, skip-gram model with negative sampling (SGNS) is equivalent to implicitly factorize a word-context matrix, whose cells are the Pointwise Mutual Information (PMI) of the respective word and context pairs, shifted by a global constant. Let $D$ as a collection of observed word and context pairs, the PMI between a word $i$ and its word context $j$ is calculated as:

$$PMI(i, j) = \log \frac{P(i, j)}{P(i) \times P(j)}$$

where $P(i, j)$ is the joint probability that word $i$ and word $j$ appears together within a window size (e.g. $P(i, j) = \frac{#(i,j)}{|D|}$), where $|D|$ refers to the total number of word and word context pairs in $D$. Similarly, $P(i)$ is the probability the word $i$ appears in $D$, and $P(j)$ is the probability word $j$ appears in $D$ (e.g. $P(i) = \frac{#(i)}{|D|}$ and $P(j) = \frac{#(j)}{|D|}$). Obviously, $PMI(i, j)$ can be calculated as:

$$PMI(i, j) = \log \frac{#(i,j) \times |D|}{#(i) \times #(j)} \quad (2)$$

By calculating $PMI$ of all word-context pairs in $D$, we can form a squared $n \times n$ matrix $M^{PMI}$ where $n$ is the total number of distinct words in $D$. Next, a Shifted Positive Pointwise Mutual Information (SPPMI) of two words $i$ and $j$ is calculated as:

$$SPPMI(i, j) = \max(PMI(i, j) - \log(s), 0) \quad (3)$$

where $s$ is a hyperparameter to control the density of PMI matrix $M^{PMI}$ and $s$ can be interpreted equivalently as a hyperparameter that indicates the number of negative samples in SGNS. When $s$ is large, more values in the matrix $M^{PMI}$ are cleared, leading $M^{PMI}$ to become sparser. When $s$ is small, matrix $M^{PMI}$ becomes denser. Finally, factorizing matrix $M^{SPPMI}$, where each cell in $M^{SPPMI}$ is transformed by Formula (3), is equivalent to performing SGNS.

Co-disliked item embedding (DIE). As mentioned in the previous studies \cite{fernando2014fastcontext, devlin2015semantics, mikolov2013distributed}, when users liked/consumed items in a sequence, the items sorted by the ascending interaction time order can be inferred as a sequence. Thus, performing co-disliked item embeddings to learn latent representations of items is equivalent to perform word embeddings to learn latent representations of words. Therefore, we can apply word embedding methods to learn latent representations of items, and perform a joint learning between embedding models and traditional factorization methods (e.g. WMF).

Given each user’s liked item list, we generate co-disliked item-item co-occurrence pairs without considering liked time. Particularly, given a certain item in the item sequence, we consider all other items as its contexts. We call this method as a greedy context generation method which can be applied to other non-timestamped datasets. After generating item and item context pairs, we construct a item co-occurrence SPPMI matrix and perform SPPMI matrix factorization. In particular, given generated item-item co-occurrence pairs, we construct a SPPMI matrix of items by applying Equation (2) to calculate the pointwise mutual information of each pair, and then by measuring the shifted positive pointwise mutual information of the pair based on Equation (3). Once the SPPMI matrix of co-disliked items is constructed, we incorporate it to the traditional matrix factorization method to improve the item latent representations.

Co-disliked item embedding (DIE). As mentioned in the Introduction section, when many users disliked two items $p_1$ and $p_2$ together, the two items can form a pair of co-occurred disliked items. If the recommender systems learned this disliked co-occurrence pattern, it would not recommend item $p_2$ to a user, who disliked $p_1$. This will help reduce the false positive rate for the recommender systems. Therefore, similar to liked item embeddings, we applied the word embedding technique to exploit the disliked co-occurrence information to enhance the item’s latent factors.

User embedding (UE). When two users A and B preferred same items, we can assume the two users share similar interests. Therefore, if user $A$ enjoyed an item $p$ that has not been observed in user B’s transactions, we can recommend the item to user B. Similar to liked and disliked item embeddings, we applied the word embedding technique to learn user embeddings that explain the co-occurrence patterns among users.

From the user-item interaction matrix $M^{mn \times n}$, where each row represents consumed items of a user (e.g. a list of items that the user rated or backed), we only keep liked items per user in the matrix $M'$. Then, we construct a $n \times m$ reverse matrix $M'^T$ of $M'$, where each row represents users that liked a certain item. Then, users, who like the same item, form a sequence, and the sequence of users is interpreted as a sequence of words. From this point, word embedding techniques are applied to the user sequence to enhance latent representations of users.

Our RME model. It is a joint learning model combining WMF, co-liked item embedding, co-disliked item embedding, and user embedding. It minimizes the following objective function:

$$L = \frac{1}{2} \sum_{u,p} w_{up}(L_{up} - a_u^p \beta_p)^2 \quad (WMF)$$

$$L_2 = \frac{1}{2} \sum_{i,j, p, q} w_{p|i} (X_{pj} - \beta_p^j \gamma_j - b_p - c_i)^2 \quad (LIE)$$

$$L_3 = \frac{1}{2} \sum_{i,j, p, q} w_{p|i} (Y_{pq} - \beta_q^i \delta_i - d_p - e_j)^2 \quad (DIE)$$

$$L_4 = \frac{1}{2} \sum_{u, j} w_{u|j} (\alpha_u - a_u^j \theta_j - f_u - g_j)^2 \quad (UE)$$

$$+ \frac{1}{2} \sum_{u, j} |\alpha_u|^2 + \sum_{p} |\beta_p|^2 + \sum_{j} |\gamma_j|^2 + \sum_{i} |\delta_i|^2 + \sum_{j} |\theta_j|^2$$

where the item’s latent representation $\beta_p$ is shared among WMF, co-liked item embedding and co-disliked item embedding. The user’s latent representation $\alpha_u$ is shared between WMF and user embedding. $X$ and $Z$ are SPPMI matrices, constructed by co-liked item-item co-occurrence patterns and disliked item-item co-occurrence patterns, respectively. $\gamma$ and $\delta$ is $k \times 1$ latent representation vectors of co-liked item context and co-disliked item context, respectively. $Z$ is a SPPMI matrix constructed by user-user co-occurrence patterns. $\theta$ is a $k \times 1$ latent representation vector of a user context. $w_{p|i}$, $w_{p|i}$ and $w_{u|j}$ are hyperparameters to compensate for item/user co-occurrences in $X$, $Y$ and $Z$ when performing decomposition. $b$ is
liked item bias, and ε is co-liked item context bias. d is disliked item bias, and ε is co-disliked item context bias. f and g are user bias and user context bias, respectively. Fixing item latent vectors L\textsubscript{reg} with regard to each user’s latent representation α\textsubscript{u} we observe that only L\textsubscript{1}, L\textsubscript{2}, L\textsubscript{3} and item regularization \frac{λ}{2} \sum p |||p|||^2 \text{ contain α.} Therefore, we obtain:

\[
\frac{\partial L}{\partial u_p} = \frac{\partial L_1}{\partial u_p} + \frac{\partial L_2}{\partial u_p} + \frac{\partial L_3}{\partial u_p} + \frac{\partial \sum ||u||^2}{\partial u_p} = \sum u_p w_{up} (\lambda^{-1} u_p - α_u) \beta_p u_p - \sum j \sum u_p w_{uj} \beta_p (\lambda^{-1} α_j - β_j - f_u - g_j) β_j + \lambda α_u^2
\]

Fixing item latent vectors β, user context latent vectors θ, user bias d and user context bias ε, and solving \frac{\partial L}{\partial u_p} = 0, we obtain the updating rule of α\textsubscript{u} as follows:

\[
α_u = \left( \sum \sum w_{up} β_p β_j u_p + \sum w_{uj} β_j u_j \right)^{-1} \left( \sum \sum w_{up} M_u p u_p + \sum w_{uj} M_j u_j \right)
\]

Similarly, taking partial derivatives of L\textsubscript{reg} with respect to each item latent vector β\textsubscript{p} L\textsubscript{reg} needs to consider only L\textsubscript{1}, L\textsubscript{2}, L\textsubscript{3} and item regularization \frac{λ}{2} \sum p |||p|||^2. By fixing other parameters and solving \frac{\partial L}{\partial p} = 0, we obtain:

\[
β_p = \left[ \sum w_{up} α_u^2 u_p^2 + \sum w_{up} β_j u_j \right]^{-1} \left[ \sum w_{up} M_u p α_u u_p + \sum w_{up} β_j u_j \right] \lambda β_p
\]

\[
δ_j = \left[ \sum w_{ip} β_p u_p + \lambda β_j \right]^{-1} \left[ \sum w_{ip} M_j u_j (X_{ip} - β_p - ε_j) β_p \right]
\]

\[
eπ = \left[ \sum w_{ip} β_p u_p + \lambda ε \right]^{-1} \left[ \sum w_{ip} M_j u_j (Y_{ip} - ε - δ_j) β_p \right]
\]

\[
θ_j = \left[ \sum w_{uj} α_u^2 u_j^2 + \lambda θ_j \right]^{-1} \left[ \sum w_{uj} M_j u_j (Z_{ju} - α_u) α_u \right]
\]

The item biases and item context biases b, c, d, e, as well as the user and user context biases f, g are updated alternatively using the following update rules:

\[
b_p = \frac{1}{|i : X_{pi} ≠ 0|} \sum_{i, X_{pi} ≠ 0} (X_{pi} - β_p Y_{ri} - c_i)
\]

\[
c_i = \frac{1}{|p : X_{pi} ≠ 0|} \sum_{p, X_{pi} ≠ 0} (X_{ip} - β_p Y_{ri} - b_p)
\]

\[
d_p = \frac{1}{|i : Y_{ri} ≠ 0|} \sum_{i, Y_{ri} ≠ 0} (Y_{ri} - β_r δ_{iri} - e_{iri})
\]

\[
e_{iri} = \frac{1}{|p : Y_{rip} ≠ 0|} \sum_{p, Y_{rip} ≠ 0} (Y_{rip} - β_p δ_{iri} - d_p)
\]

\[
f_u = \frac{1}{|u : Z_{ju} ≠ 0|} \sum_{u, Z_{ju} ≠ 0} (Z_{ju} - α_u θ_j - f_u)
\]

\[
g_j = \frac{1}{|u : Z_{ju} ≠ 0|} \sum_{u, Z_{ju} ≠ 0} (Z_{ju} - α_u θ_j - f_u)
\]

3.2 Optimization

We can use the stochastic gradient descent to optimize the Equation (4). However, it is not stable and sensitive to parameters [37]. Therefore, we adopt vector-wise ALS algorithm [37, 39] that alternatively optimize each model’s parameter in parallel while fixing the other parameters until the model gets converged. Specifically, we calculate the partial derivatives of the model’s objective function with regard to the model parameters (i.e., {α\textsubscript{1:n}, β\textsubscript{1:n}, γ\textsubscript{1:n}, δ\textsubscript{1:n}, ε\textsubscript{1:n}, δ\textsubscript{1:m}, f\textsubscript{1:m}, g\textsubscript{1:m}}). Then we set them to zero and obtain updating rules. Details are given as follows:

From the objective function in Equation (4), while taking partial derivatives of L with regard to each user’s latent representation vector α\textsubscript{u}, we observe that only L\textsubscript{1}, L\textsubscript{4} and the L\textsubscript{2} user regularization \frac{λ}{2} \sum p |||p|||^2 contain α. Therefore, we obtain:

\[
\frac{\partial L}{\partial α_u} = \frac{\partial L_1}{\partial α_u} + \frac{\partial L_2}{\partial α_u} + \frac{\partial L_3}{\partial α_u} + \frac{\partial \sum ||α||^2}{\partial α_u} = \sum u_p w_{up} (β_p^2 - β_j^2) - \sum j \sum u_p w_{uj} (β_p^2 - ε_j α_u) β_j + \lambda α_u^2
\]

3.3 Complexity Analysis

In this section, we briefly provide time complexity analysis of our model. Let Ω\textsubscript{X} = {|u, p | M\textsubscript{up} ≠ 0}, Ω\textsubscript{Y} = {|p, i | X\textsubscript{pi} ≠ 0}, Ω\textsubscript{Y} = {|p, i' | Y\textsubscript{rip} ≠ 0}, Ω\textsubscript{Z} = {|u, j | Z\textsubscript{ju} ≠ 0}. Constructing SPMI matrices X, Y and Z take O(|Ω\textsubscript{X}|\textsuperscript{2}), O(|Ω\textsubscript{Y}|\textsuperscript{2}) and O(|Ω\textsubscript{Z}|\textsuperscript{2}), respectively. However, the SPMI matrices are calculated once and are constructed in parallel using batch processing, so they are not costly. For learning RME model, computing α takes O(((|Ω\textsubscript{X}| + |Ω\textsubscript{Y}|)^2 + k^2) time, and computing β takes O(((|Ω\textsubscript{X}| + |Ω\textsubscript{Y}|)^2 + k^2) time. Also, it takes O(|Ω\textsubscript{X}| k^2 + k^3) for computing co-liked item context γ, and so do other latent contexts δ, θ. It takes O(|Ω\textsubscript{Y}| k^2) time to compute all user bias f and so do the other biases. Thus, the time complexity for RME is O(n(2(|Ω\textsubscript{X}| + |Ω\textsubscript{Y}|) + |Ω\textsubscript{Y}| + |Ω\textsubscript{Z}|)k^2 + (2m + 3n)k^3), where n is the number of iterations. Since k << min(m, n) and M, X, Y, Z are often sparse, which mean (|Ω\textsubscript{M}| + |Ω\textsubscript{X}| + |Ω\textsubscript{Y}| + |Ω\textsubscript{Z}|) is small,
the time complexity of RME is shortened as $O(n(m + \frac{3}{2}n)k^3)$, which scales linearly to the conventional ALS algorithm for collaborative filtering [37].

4 INFERRING DISLIKED ITEMS IN IMPLICIT FEEDBACK DATASETS

Unlike explicit feedback datasets, there is a lack of substantial evidence, on which items the users disliked in implicit feedback datasets. Since our model exploits co-disliked item co-occurrences patterns among items, the implicit feedback datasets challenge our model. To deal with it, we can simply assume that missing values are equally likely to be negative feedback, then sample some negative instances from missing values with uniform weights [12, 27, 32, 34]. However, assigning uniform weight is suboptimal because the missing values are a mixture of negative and unknown feedbacks. A recent work suggests to sample negative instances by assigning non-uniform weights based on item popularity [13]. The idea is that popular items are highly aware by users, so if they are not observed in a user’s transactions, it assumes that the user dislikes them. However, this sampling method is also not optimal because same unobserved popular items can be sampled across multiple users. This approach does not reflect user’s personalized interests.

Instead, we follow the previous works [22, 25, 38], and propose a user-oriented EM-like algorithm to draw negative samples (i.e., inferred disliked items) for users in implicit feedback datasets. Our approach is described as follows:

First, we assume that an item with a low ranking score of being liked will have a higher probability to be drawn as a negative sample of a user. Given $r_u$, the ranked list of all items of the user $u$, the prior probabilities of items to be drawn as negative samples are calculated by using a softmax function as follows:

$$p_{\tau_i} = \frac{\exp(-r_u[i])}{\sum_{j=1}^{n}\exp(-r_u[j])} \quad (9)$$

After negative samples are drawn for each user, we built the RME model by using Algorithm 1. The pseudocode of the RME model for implicit feedback datasets is presented in Algorithm 2.

In Algorithm 2, since each user may prefer a different number of items, we define a hyper-parameter $\tau$ as a negative sample drawing ratio to control how many negative samples we will sample for each user. In line 6, $count(u)$ returns the number of observed items of a user $u$. Then, the number of drawn negative samples for the user $u$ is calculated and assigned to $ns$. If a user prefers 10 items and $\tau = 0.8$, the algorithm will sample 8 disliked items. We note that sampling with replacement is used such that different items are drawn independently. The value of $\tau$ is selected using the validation data. In line 8, we set the ranking of observed items to $+\infty$ to avoid drawing the observed items as negative samples. In line 12, we build the RME model based on the negative samples drawn in the Expectation step, and temporally store newly learned user latent matrix, item latent matrix and corresponding NDCG to $U_{\text{tmp}}, P_{\text{tmp}}, \text{ndcg}$ variables, respectively (NDCG is a measure to evaluate recommender systems, which will be mentioned in Experiment section). If we obtain a better ndcg comparing with the previous NDCG prev_ndcg (line 13), we will update $U, P, \text{prev}_n\text{dcg}$ with new values (line 14). Overall, at the end of the Expectation step, we obtain the disliked items

for each user. Then, in the Maximization step, we build our RME model to re-learn user and item latent representations $U$ and $P$. The process is repeated until getting converged or the early stopping condition (line 13 to 17) is satisfied.

Time Complexity: In order to construct RME model for implicit feedback datasets, we need to re-learn RME model, which includes re-building 3 SPMI matrices in the maximization step in $\eta$ iterations to get converged. Thus, it takes $O(n\eta(|\Omega_X|^2 + |\Omega_Y|^2 + |\Omega_Z|^2) + \eta(m + \frac{3}{2}n)k^3))$ time where $\eta$ is small.

5 EXPERIMENTS

5.1 Experimental Settings

Datasets: To measure the performance of our RME model, we evaluate the model on 3 real-world datasets:

- **MovieLens-10M** [29]: is an explicit feedback dataset. It consists of 69,878 users and 10,677 movies with 10m ratings. Following the k-cores preprocessing [11, 12], we only kept users, who rated at least 5 movies, and movies, which were rated by at least 5 users. This led to 58,057 users and 7,223 items (density = 0.978%).
- **MovieLens-20M**: is an explicit feedback dataset. It consists of 138,000 users, 27,000 movies, and 20 millions of ratings. We filtered with the same condition as for MovieLens-10M. This led to 111,146 users and 9,888 items (density= 0.745%).
- **TasteProfile**: is an implicit feedback dataset containing a song’s play count by a user. The play counts are user’s implicit preference and are binarized. Similar to [20], we first

\[\text{http://the.echonest.com/}\]
We note that we do not compare our models with user collaborative filtering methods. Table 2 presents the performance of the baselines, our RME model, and its two variants. The improvement of our model over the baselines was significant with p-value < 0.05 in the three datasets under the non-directional two-sample t-test.

<table>
<thead>
<tr>
<th>Method</th>
<th>MovieLens-10M</th>
<th>MovieLens-20M</th>
<th>TasteProfile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall@5</td>
<td>NDCG@20</td>
<td>MAP@10</td>
</tr>
<tr>
<td>Item-KNN</td>
<td>0.0137</td>
<td>0.0338</td>
<td>0.0397</td>
</tr>
<tr>
<td>Item2vec</td>
<td>0.1020</td>
<td>0.1001</td>
<td>0.0502</td>
</tr>
<tr>
<td>WMF</td>
<td>0.1280</td>
<td>0.1245</td>
<td>0.0655</td>
</tr>
<tr>
<td>Cofactor</td>
<td>0.1460</td>
<td>0.1381</td>
<td>0.0772</td>
</tr>
<tr>
<td>U_RME</td>
<td>0.1516</td>
<td>0.1412</td>
<td>0.0818</td>
</tr>
<tr>
<td>I_RME</td>
<td>0.1511</td>
<td>0.1422</td>
<td>0.0817</td>
</tr>
<tr>
<td>RME</td>
<td>0.1562</td>
<td>0.1458</td>
<td>0.0841</td>
</tr>
</tbody>
</table>

We subsampled the dataset to 250k users and 25k items. Then we kept only users, who listened to at least 20 songs, and songs, which were listened by at least 50 users. As a result, 221,011 users and 22,713 songs were remained (density = 0.291%).

**Baselines:** To illustrate the effectiveness of our RME model, we compare it with the following baselines:

- **WMF [14]:** It is a weighted matrix factorization with l2-norm regularization.
- **Item-KNN [7]:** This is an item neighborhood-based collaborative filtering method.
- **Item2vec [3]:** This method used Skip-gram with negative sampling [24] to learn item embeddings, then adopted a similarity score between item embeddings to generate user’s recommendation lists.
- **Cofactor [20]:** This is a method that combines WMF and co-liked item embedding.

We note that we do not compare our models with user collaborative filtering method (i.e. User-KNN) because it is not applicable to run the method on the large datasets. However, [31] reported that User-KNN had worse performance than Item-KNN, especially when there are many items but few ratings in a dataset.

**Our models:** We not only compare the baselines with our RME, but also two variants of our model such as U_RME and I_RME to show the effectiveness of incorporating all of the user embeddings, liked-item embeddings and disliked-item embeddings:

- **U_RME (i.e., RME - DIE):** This is a variant of our model, considering only WMF, user embeddings, and liked-item embeddings.
- **I_RME (i.e., RME - UE):** This is another variant of our model, considering only WMF, liked-item embeddings, and disliked-item embeddings.
- **RME:** This is our proposed RME model.

**Evaluation metrics:** We used three well-known ranking-based metrics – Recall@N, normalized discounted cumulative gain (NDCG@N), and mean average precision (MAP@N). Recall@N considers all items in top N items equally, whereas NDCG@N and MAP@N apply an increasing discount of log2 to items at lower ranks.

**Training, validation and test sets:** We follow 70/10/20 proportion for splitting the original dataset into training/validation/test sets [21]. MovieLens-10M and MovieLens-20M datasets contain timestamp values of user-movie interactions. To create training, validation and testing sets for these datasets, we sorted all user-item interaction pairs in the ascending interaction time order in each of MovieLens-10M and MovieLens-20M datasets. The first 80% was used for training and validation, and the rest 20% data was used as a test set. Out of 80% data extracted for training and validation, we randomly took 10% for the validation set. To measure the statistical significance of RME over the baselines, we repeated the splitting process five times (i.e., generating five pairs of training and validation sets). Since TasteProfile dataset did not contain timestamp information of user-song interactions, we randomly split the TasteProfile dataset into training/validation/test sets five times with 70/10/20 proportions. Averaged results are reported in the following subsection.

**Stopping criteria and Hyperparameters.** To decide when to stop training a model, we measured the model’s NDCG@100 by using the validation set. We stopped training the model when there was no further improvement. Then, we applied the best model to the test set to evaluate its performance. This method was applied to the baselines and RME.

All hyper-parameters were tuned on the validation set by a grid search. We used the same hyper-parameter setting in all models. The grid search of the regularization weight λ was performed in [0.001, 0.005, 0.01, 0.05, ..., 10]. The size of latent dimensions was in a range of [30, 40, 50, ..., 100]. We set weights $w^{(p)} = w^{(-p)} = w^{(u)} = w$ for all user-user and item-item co-occurrence pairs. When building our RME model for TasteProfile dataset, we do a grid search for the negative sample drawing ratio $r$ in [0.2, 0.4, 0.6, 0.8, 1.0].

**5.2 Experimental Results**

**RQ1: Performance of the baselines and RME.** Table 2 presents recommendation results of RME and compared models at Recall@5, NDCG@20, and MAP@10. First, we compared RME with the baselines. We observed that RME outperformed all baselines in the three datasets, improving the Recall by 6.3%, NDCG by 5.1%, and MAP by 8.3% on average over the best baseline (p-value < 0.001). Second, we compared two variants of RME model with the baselines. We see that both U_RME and I_RME performed better than the baselines. Adding user embeddings improved the Recall by 3.0–3.5%, NDCG by 1.4–2.2%, and MAP by 4.2–5.8% (p-value < 0.001), while adding disliked item embeddings improved the Recall by 3.1–3.8%, NDCG by 2.2–3.0%, and MAP by 4.9–6.0%. Third, we compare RME with its two variants. RME also achieved the best result, improving Recall by 2.6–3.0%, NDCG by 2.0–2.5%, and MAP by 2.6–2.8% (p-value < 0.05). We further evaluated NDCG@N of our model when varying...
Figure 3: Performance of models when varying the latent dimension size $k$ with fixing the value of $\lambda$.

Figure 4: Performance of models when varying $\lambda$ with fixing the latent dimension size $k$. Item2Vec did not contain regularization, so we excluded it.

disliking information) at 0.3%, 0.7% and 1.3% on average in Recall, NDCG, and MAP, respectively. The experimental results confirmed the effectiveness of Algorithm 2. We note that Algorithm 2 got converged in up to 4 iterations for all three datasets by the early stopping condition. Due to the space limitation, we do not include figures which show the loss over iterations.

**RQ2:** Parameter sensitivity analysis: We analyze the effects of the parameters in RME model in order to answer the following research questions: *(RQ2-1)* How does RME work when varying the latent dimension size $k$? *(RQ2-2)* How does RME model change with varying $\lambda$? *(RQ2-3)* How sensitive is the RME model on an implicit feedback dataset (e.g. TasteProfile) when varying negative sample drawing ratio $r$? and *(RQ2-4)* Can RME achieve better performance with a dynamic setting of regularization hyper-parameters?

Regarding *RQ2-1*, Figure 3 shows the sensitivity of all compared models when fixing $\lambda$ and varying the latent dimension size $k$ in $\{30, 40, 50, 60, 70, 80, 90, 100\}$. It is clearly observed that our
model outperforms the baselines in all datasets. In MovieLens-10M and MovieLens-20M datasets, all six models downgrade the performance when the latent dimension size $k$ is over 60. In the TasteProfile dataset, when increasing $k$, although all models gain a higher performance, our model tends to achieve much higher performance.

In a RQ2-2 experiment, we exclude Item2Vec because this model does not contain the regularization term. We fix $k=40$ in MovieLens-10M and MovieLens-20M. In TasteProfile dataset, we fix $k=100$, $	au=0.2$. We vary lambda in range $[0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10]$. Then, we report the average results of Recall@5, NDCG@5, and MAP@5. As shown in Figure 4, the performance of our model is better than the baselines. In MovieLens-10M and MovieLens-20M dataset, RME increases its performance when increasing $\lambda$ up to 1, then its performance goes down when $\lambda$ is increasing more. In TasteProfile, RME tends to gain a higher performance and more outperformed the baselines when $\lambda$ is increasing.

To understand the sensitivity of our model when varying negative sample drawing ratio $\tau$ in the implicit feedback dataset – TasteProfile (RQ2-3), we vary $\tau$ in $[0.2, 0.4, 0.6, 0.8, 1.0]$, and fix $k=100$ and $\lambda=10$. Figure 5 shows that when $\tau$ increases, our model degrades with a small amount (e.g. around -0.3% in Recall@5 and NDCG@5, and -0.4% in MAP@5). In NDCG@5, our model gains the best result when $\tau = 0.4$. We note that our worst case (when $\tau = 1.0$) is still better than the best baseline presented in Table 2. This shows that the sensitivity of our model with regard to the negative sample drawing ratio $\tau$ is small/limited.

In our previous experiments, we used a static setting of regularization hyper-parameters by setting $\lambda_{\alpha} = \lambda_{\beta} = \lambda_{\gamma} = \lambda_{\delta} = \lambda_{\gamma} = \lambda$. To explore if a dynamic setting of those regularization hyper-parameters could lead to better results for RME model (RQ2-4), we set $\lambda_{\alpha} = \lambda_{\beta} = \lambda_{1}$, $\lambda_{\gamma} = \lambda_{\delta} = \lambda_{\gamma} = \lambda_{d}$. Then we both vary $\lambda_{1}$ and $\lambda_{2}$ in $[100, 50, 10, 5, 1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001]$ while fixing the latent dimension size $k$. Next, we report the NDCG@5 for all 3 datasets. As shown in Figure 6, our model even get a higher performance with the dynamic setting. For example, it gains NDCG@5 = 0.1613 when $\lambda_{1} = 100$ and $\lambda_{2} = 0.005$ in MovieLens-10M dataset. Similarly, NDCG@5 = 0.1639 when $\lambda_{1} = 0.5$, $\lambda_{2} = 1$ in MovieLens-20M dataset. NDCG@5 = 0.204 when $\lambda_{1} = 100$, $\lambda_{2} = 10$ in TasteProfile dataset. The dynamic setting produced 0.3~2% higher results than the static setting presented in Table 2.

So far, we compared the performance of our model and the baselines while varying values of hyper-parameters. We showed that our model outperformed the baselines in all cases, indicating that our model was less sensitive with regard to the hyper-parameters. We also showed that our model produced better results under the dynamic setting.

**RQ3: Performance of models for different types of users.** We sorted users by the ascending order of their activity level in terms of the number of liked items. Then we categorized them into three groups: (1) cold-start users who were in the first 20% of the sorted user list (i.e., their activity level is the lowest); (2) warm-start users who were in between 20% and 80% of the sorted user list; (3) highly active users who were in the last 20% of the sorted user list (i.e., the most active users). Then, we measured the performance of all the compared models for each of the user groups.

Figure 7 shows the performance of all the compared models in MovieLens-10M, MovieLens-20M and TasteProfile datasets. In MovieLens-10M (Figure 7(a)), our model significantly outperformed the baselines and the two variants in all three user groups, improving Recall@5 by 4.7~6.7%, NDCG@5 by 6.8~8.8%, and MAP@5 by 9.2~11.0% over the best compared method. In MovieLens-20M dataset (Figure 7(b)), our model significantly outperformed the baselines and its variants in 2 groups: cold-start users and warm-start users. It improved Recall@5 by 16.1%, 4.0%, 0.8%, NDCG@5 by 15.3%, 4.4%, 0.9%, MAP@5 by 17.3%, 5.1%, 1.1% in cold-start users, warm-start users and highly-active users, respectively. Specially, in both MovieLens-10M and MovieLens-20M datasets, our model on average much improved the baselines in cold-start users with Recall@5, NDCG@5 and MAP@5 by 27.9%, 24.8% and 23.3%, respectively. It shows the benefit of incorporating disliked item embeddings and user embeddings. In TasteProfile dataset (Figure 7(c)), our model significantly improved baselines in highly-active users group, improving Recall@5 by 6.8%, NDCG@5 by 7.4%, and MAP@5 by 10.0% comparing to the best state-of-the-art method, while improving Recall@5 by 5.0%, NDCG@5 by 4.9%, and MAP@5 by 5.7% comparing to its best variant. However, in cold-start users and warm-start users group, RME got an equal performance comparing with the baselines (i.e., the difference between our model and other methods are not significant).

**RQ4: Joint learning vs separate learning.** What if we conduct learning separately for each part of our model? Will the separate learning model perform better than our joint learning model? To answer the questions, we built a separate learning model as follows: first, we learned latent representations of items by jointly
decomposing two SPPMI matrices $X^{(+)}$ and $X^{(-)}$ of liked item-item co-occurrences and disliked item-item co-occurrences, respectively. Then, we learned user’s latent representations by minimizing the objective function in Equation (4), where the latent representations of items and item contexts were already learned and fixed. Next, we compared our joint learning model (i.e., RME) with the separate learning model in MovieLens-10M, MovieLens-20M, and TasteProfile datasets. Our experimental results show that our joint learning model outperformed the separate learning model by significantly improving Recall@5, NDCG@5 and MAP@5 at least 12.1%, 13.5% and 17.1%, respectively (p-value < 0.05).

## 6 RELATED WORK

**Latent factor models (LFM):** Some of the first works in recommendation focused on explicit feedback datasets (name some: [15, 30, 31]). Our proposed method worked well for both explicit and implicit feedback settings with almost equal performances.

In implicit feedback datasets, which have been trending recently due to the difficulty of collecting users’ explicit feedback, properly treating/modeling missing data is a difficult problem [4, 21, 27]. Even though missing values are a mixture of negative feedback and unknown feedback, many works treated all missing data as negative instances [8, 14, 27, 34], or sampled missing data as negative instances with uniform weights [28]. This is suboptimal because treating negative instances and missing data differently can further improve recommenders’ performance [13]. [25] proposed a bagging of ALS learners [14] to sample negative instances. He et al. [13] assumed that unobserved popular items have a higher chance of being negative instances. In our work, we attempted to non-uniformly sample negative instances in implicit feedback datasets and treated them as additional information to enhance our model performance. Specifically, we (i) designed an EM-like algorithm with a softmax function to draw personalized negative instances for each user; (ii) employed a word embedding technique to exploit the co-occurrence patterns among disliked items, further enriching their latent representations.

**LFM with auxiliary information:** In latent factor models, additional sources of information were incorporated to improve collaborative filter-based recommender systems (e.g., user reviews, item categories, and article information [2, 10, 23, 35]). However, we only used an user-item-preference matrix without requiring additional side information. Adding the side information into our model would potentially further improve its performance. But, it is not a scope of our work in this paper.

**LFM with item embeddings:** [36] incorporated message embedding for retweet prediction. Cao et al. [6] co-factorized the user-item interaction matrix, user-list interaction matrix, and item-list co-occurrences to recommend songs and lists of songs for users. [20] learned liked item embeddings with an equivalent matrix factorization method of skip-gram negative sampling (SGNS), and performed joint learning with matrix factorization. [3] exploited item embeddings using the SGNS method for item collaborative
filtering. So far, the closest techniques to ours [3, 20] only considered liked item embeddings, but we proposed a joint learning model that not only considered LFM using matrix factorization with liked item embeddings, but also user embeddings and disliked item embeddings. Since integrating co-disliked item embedding is non-trivial for implicit feedback datasets, we also proposed an EM-like algorithm for extracting personalized negative instances for each user.

Word embeddings: Word embedding models [24, 26] represent worked on different types of users in terms of their interaction activity. We proposed a joint model combining WMF, co-liked embedding, co-disliked embedding and user embedding, following the recent success of word embedding techniques. Through comprehensive experiments, we successfully demonstrated that our model outperformed all baselines, significantly improving NDCG@20 by 5.6% in MovieLens-10M dataset, by 5.3% in MovieLens-20M dataset, and by 4.3% in TasteProfile dataset. We also analyzed how our model worked on different types of users in terms of their interaction activity levels. We observed that our model significantly improved NDCG@5 by 20.2% in MovieLens-10M, by 29.4% in MovieLens-20M for the cold-start users group. In the future extension of our model, we are interested in selecting contexts for users/items by setting a timestamp-based window size for timestamped datasets. In addition, we are also interested in incorporating co-disliked patterns among users (i.e., co-disliked user embeddings) into our model.

7 CONCLUSION

In this paper, we proposed to exploit different co-occurrence information: co-liked item-item co-occurrences and user-user co-occurrences, which were extracted from the user-item interaction matrix. We proposed a joint model combining WMF, co-liked embedding, co-disliked embedding and user embedding, following the recent success of word embedding techniques. Through comprehensive experiments, we successfully demonstrated that our model outperformed all baselines, significantly improving NDCG@20 by 5.6% in MovieLens-10M dataset, by 5.3% in MovieLens-20M dataset, and by 4.3% in TasteProfile dataset. We also analyzed how our model worked on different types of users in terms of their interaction activity levels. We observed that our model significantly improved NDCG@5 by 20.2% in MovieLens-10M, by 29.4% in MovieLens-20M for the cold-start users group. In the future extension of our model, we are interested in selecting contexts for users/items by setting a timestamp-based window size for timestamped datasets. In addition, we are also interested in incorporating co-disliked patterns among users (i.e., co-disliked user embeddings) into our model.

8 ACKNOWLEDGMENT

This work was supported in part by NSF grants CNS-1755536, CNS-1422215, DGE-1663343, CNS-1742702, DGE-1820609, Google Faculty Research Award, Microsoft Azure Research Award, and Nvidia GPU grant. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

REFERENCES


