Global-Local Item Embedding for Temporal Set Prediction

SEUNGJAE JUNG, Naver R&D Center NAVER CLOVA, Korea, Republic of

YOUNG-JIN PARK, Naver R&D Center NAVER CLOVA NAVER AI LAB, Korea, Republic of

JISU JEONG and KYUNG-MIN KIM, NAVER CLOVA NAVER AI LAB, Korea, Republic of

HIUN KIM, MINKYU KIM, and HANOCK KWAK, NAVER CLOVA, Korea, Republic of

Temporal set prediction is becoming increasingly important as many companies employ recommender systems in their online businesses, e.g., personalized purchase prediction of shopping baskets. While most previous techniques have focused on leveraging a user’s history, the study of combining it with others’ histories remains untapped potential. This paper proposes Global-Local Item Embedding (GLOIE) that learns to utilize the temporal properties of sets across whole users as well as within a user by coining the names as global and local information to distinguish the two temporal patterns. GLOIE uses Variational Autoencoder (VAE) and dynamic graph-based model to capture global and local information and then applies attention to integrate resulting item embeddings. Additionally, we propose to use Tweedie output for the decoder of VAE as it can easily model zero-inflated and long-tailed distribution, which is more suitable for several real-world data distributions than Gaussian or multinomial counterparts. When evaluated on three public benchmarks, our algorithm consistently outperforms previous state-of-the-art methods in most ranking metrics.

CCS Concepts:
- Information systems → Information systems applications; Recommender systems.

Additional Key Words and Phrases: Temporal Sets, Set Prediction, Tweedie Distribution, Variational Autoencoder

ACM Reference Format:
Seungjae Jung, Young-Jin Park, Jisu Jeong, Kyung-Min Kim, Hiun Kim, Minkyu Kim, and Hanock Kwak. 2021. Global-Local Item Embedding for Temporal Set Prediction. In Fifteenth ACM Conference on Recommender Systems (RecSys ’21), September 27-October 1, 2021, Amsterdam, Netherlands. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3460231.3478844

1 INTRODUCTION

Many recommendation tasks can be viewed as a problem of predicting the next set given the sequence of sets, e.g., predicting the next basket in online markets and the next playlist in streaming services. Previous works mainly focused on the local information (a given user’s history), applying RNNs [2, 4, 20] or self-attention [21] to learn the temporal tendency of a sequence of sets within a user. From the recommender system point of view, temporal set prediction can have a sparsity problem as many users interact with only a small number of items. Throughout the recommender system literature, the sparsity problem has been dealt with low-rank approximation or collaborative filtering [11, 15–17]. However, such attempts have been less explored in temporal set prediction literature.

In this paper, we propose Global-Local Item Embedding (GLOIE) that integrates global and local information for temporal set prediction. To capture the global information, we utilize the Variational Autoencoders (VAEs) [10, 12] which are effective on noisy and sparse data. GLOIE then integrates the local embeddings, that are learned through dynamic graphs [21], with global embeddings by using an attention method. In addition, we enhance the performance...
Fig. 1. Overview of GLOIE. We first make sum of time decayed vector $x_i$ (left part of the figure). VAE maximizes the ELBO of each $x_i$. The weighted sum of $\hat{x}_{ij} \cdot w_i$ and $z_{ij}$ becomes final embedding for interacted items. The weight $\alpha$ is determined by the attention mechanism. For the items that the user never interacted, we just use $\hat{x}_{ij}$.

of GLOIE by using Tweedie distribution for the likelihood of VAE instead of Gaussian or multinomial distributions. We show that purchase logs follow zero-inflated and long-tail distributions similar to the Tweedie distribution.

We conduct experiments on three public benchmarks, i.e. Dunnhumby Carbo, TaoBao, and Tags-Math-Sx. Empirical results demonstrate that the proposed method outperforms state-of-the-art methods on most metrics. The main contributions of the paper are summarized as follows:

- We propose GLOIE which differentiate global and local information and integrates them.
- We claim that using Tweedie output for VAE decoder is beneficial as it naturally models two properties of data distributions from temporal set prediction problems: zero-inflated and long-tailed.
- We achieve state-of-the-art performance on three public benchmarks.

2 PRELIMINARIES

2.1 Problem Definition

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ be set of users and $\mathcal{E} = \{e_1, e_2, \ldots, e_M\}$ be set of items. Given a user $u_i$’s sequence of sets $\mathcal{S}_i = \{S^k_i | 1 \leq k \leq T_i\}$, our goal is to predict next set $S^T_{i+1}$. Each set $S^k_i$ can also be represented as a binary vector form $v^k_i$, where each element $v^k_{ij} = 1(e_j \in S^k_i)$. $1(x)$ is a indicator function which returns 1 if $x$ is true and 0 otherwise. We will use the set notation $S^k_i$ and $v^k_i$ interchangeably to denote user $u_i$’s $k$-th set.

2.2 Variational Autoencoders

Variational Autoencoders (VAEs) [10, 13] are a class of deep generative models. VAEs provide latent structures that can nicely explain the observed data (e.g., customer’s purchase history). Formally speaking, VAEs find the latent variables ($z$) that maximize the evidence lower bound (ELBO), a surrogate objective function for the maximum likelihood estimation of the given observation $x$:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(x|z)} [\log p_\theta(x|z)] - D_{KL}(q_\phi(z \mid x) || p(z))$$

where $\theta$ and $\phi$ are model parameters of a decoder and an encoder, respectively.

Encoders are commonly structured by multilayer perceptrons (MLPs) that produce Gaussian distributions over the latent variable $z$. On the other hand, decoders are designed to have different probability distributions of output layer depending on the characteristics of datasets. For example, Gaussian distribution and multinomial distribution are often used to represent the real-valued continuous and binary data, respectively.
When training VAE, using gaussian output on We then update our model to maximize the ELBO of each $x_i$. We are interested in modeling other users’ history as well as the given user’s history. We proceed with this under the $D$, where $v_i$ represents the index and $T_i - k$ in $z_{T_i-k}$ represents the power. We then update our model to maximize the ELBO of each $x_i$:

$$
\mathcal{L}_{vae} = \mathbb{E}_{x_i \sim D_u} \left[ -\mathbb{E}_{z_i \sim q(z|x_i)} \log p(x_i | z_i) + D_{KL}(q(z | x_i) || p(z)) \right]
$$

where $D_u$ is a set of sum of time decayed sequence of sets $x_i$. Equation (3) is merely an expectation of negative of Equation (1) over $D_u$. Equation (3) can be understood as the low-rank approximation with the constraint $D_{KL}(q(z | x_i) || p(z))$.

As most people interact with a few items, $v_i$ contains many 0s and so does $x_i$. However, the reconstructed vector $\hat{x}_i$ generated from the following process:

$$
z_i \sim q(z | x_i) \quad \hat{x}_i \sim p(x_i | z_i)
$$

is a dense vector and contains the information of expected preference of user $u_i$ to item $e_j$ in $\hat{x}_{ij}$ even though the user $u_i$ never interacted with item $e_j$. Consider two users $u_{i1}$ and $u_{i2}$ with similar histories. $z_{i1}$ and $z_{i2}$ should be close as $x_{i1}$ and $x_{i2}$ are close. In turn, the distance between two reconstructed vectors $\hat{x}_{i1}$ and $\hat{x}_{i2}$ is small. Hence, even if user $u_{i1}$ never interacted with item $e_j$, $\hat{x}_{ij}$ should be high if $x_{ij}$ is high.

Note that Hu et al. [5]’s Personalized Item Frequency (PIF) is similar to our sum of time decayed vectors. However, since their work is based on K-Nearest Neighbor, two shortcomings arise: 1) the inference time grows cubic with the number of users and 2) they cannot use features unlike traditional deep learning approaches.

### 3.2 Tweedie Output on Decoder

When training VAE, using gaussian output on $p(x | z)$ is a straightforward option. However, the distributions of the data generated from temporal set prediction problems are zero-inflated and long-tailed as shown in Figure 2.

 funation $V(\mu) = \mu^p$ [6]. Tweedie distribution with $1 < p < 2$ corresponds to a class of compound Poisson distributions [7]. Consider two step sampling process $N \sim Poisson(\lambda)$ and $X_i \sim Gamma(\alpha, \beta)$ for $\lambda, \alpha, \beta > 0$ and $i = 1, \ldots, N$. Now we define a random variable $Z$ as follows:

$$
Z = \begin{cases} 
0 & N = 0 \\
X_1 + \cdots + X_N & N > 0
\end{cases}
$$

It is straightforward that the distribution defined by Equation (5) is zero-inflated for small enough $\lambda$ and long-tailed as $Z$ is addition of $X_i \sim Gamma(\alpha, \beta)$ for $N > 0$. Hence using Tweedie output on VAE’s decoder for temporal set prediction is beneficial as Tweedie distribution easily captures the properties of distributions that are shown in Figure 2.
Learning the mean parameter \( \mu \) and the power parameter \( p \) of Tweedie distribution via maximum likelihood is easy. Minimizing

\[
\mathcal{L}_{\text{Tweedie}}(z, \mu, p) = -z \left( \frac{\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p} \right)
\]

maximizes log-likelihood. Here \( z \) is the target. See Yang et al. [19] for details.

A line of works chose distributions other than Gaussian or Bernoulli on matrix factorization and VAE [3, 12, 18]. Poisson distribution or multinomial distribution are usual choices. This paper is the first attempt to apply Tweedie distribution for VAE’s decoder output to the best of our knowledge.

### 3.3 Integrating Global-Local Information

Though VAE with Tweedie output is already competent, it tends to underestimate a user’s preference for the frequently interacted items. This tendency owes to the learning objective of VAE as the model has to maximize the likelihood of 0 for never interacted items. This sometimes sacrifices the ability to maximize the likelihood of values of interacted items. Hence, we integrate item embeddings of frequently interacted items which are learned by state-of-the-art models to our VAE. We use DNNTSP [21] as it is the state-of-the-art method. DNNTSP makes user-dependent embeddings of items that are interacted at least once via dynamic graph neural networks. We denote \( z_{ij} \) as the embedding of user \( u_i \) for item \( e_j \).

When it comes to combining VAE with embeddings \( z_{ij} \), a problem arises: the reconstructed value \( \hat{x}_{ij} \) of Equation (4) is a scalar while learned embedding \( z_{ij} \) is a vector. Hence, to match the size between \( \hat{x}_{ij} \) and \( z_{ij} \), we simply multiply a vector \( w^* \), which is of same size as \( z_{ij} \), to \( \hat{x}_{ij} \). We defer the discussion on the selection of \( w^* \) to latter part of this section.

Now we combine updated embedding with \( \hat{x}_{ij} \cdot w^* \)’s and \( z_{ij} \)’s. The updated embedding \( \tilde{z}_{ij} \) is defined as
Table 1. Statistics of public benchmarks

| Dataset | # of users | # of sets | # of elements | #E/S | #S/U | #E/U |
|---------|------------|-----------|---------------|------|------|------|
| DC      | 9,010      | 42,905    | 217           | 1.52 | 4.76 | 5.44 |
| TaoBao  | 113,347    | 628,618   | 689           | 1.10 | 5.55 | 4.96 |
| TMS     | 15,726     | 243,394   | 1,565         | 2.19 | 15.48| 18.05|

\[
\tilde{x}_{ij} = \begin{cases} 
\text{Att}(\tilde{x}_{ij} \cdot w^*, z_{ij}) & e_j \in \bigcup_k S_i^k \\
\hat{x}_{ij} \cdot w^* & \text{otherwise} 
\end{cases}
\] (7)

where \( \tilde{x}_{ij} = \frac{x_{ij}}{\max_j x_{ij}} - 0.5 \) which normalizes all values of \( \tilde{x}_i \) to \([-0.5, 0.5]\) and \( \text{Att}(q, k) \) is defined as follows:

\[
\text{Att}(q, k) = \alpha \cdot q + (1 - \alpha) \cdot k, \text{ where } \alpha = \sigma \left( (W_q \cdot q)^T (W_k \cdot k) \right).
\] (8)

Lastly, we calculate the affinity of a user \( u_i \) to item \( e_j \) by

\[
\hat{y}_{ij} = \sigma \left( \tilde{x}_{ij}^T w_0 + b_0 \right)
\] (9)

where \( \sigma(x) = \frac{1}{1 + e^{-x}} \). We choose multi-label soft margin loss for training:

\[
L_{TSP} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}).
\] (10)

We now going back to the selection of \( w^* \) in Equation (7). We set \( w^* = w_0 \) in Equation (9) though we can set \( w^* \) as a learnable parameter. If we set \( w^* = w_0 \), Equation (9) becomes \( \hat{y}_{ij} = \sigma \left( \tilde{x}_{ij}^T w_0 + b_0 \right) \) for non-interacted item \( e_j \notin \bigcup_k S_i^k \), hence it preserves the order of affinity that is learned by VAE. We also run experiments with learnable parameter \( w^* \) but the performance was on par with \( w^* = w_0 \).

4 EXPERIMENTS

4.1 Benchmarks

We evaluate our method on three public benchmarks: Dunnhumby Carbo (DC), TaoBao, and Tags-Math-Sx (TMS). We partitioned each dataset into train, validation and test into 70%, 10% and 20% respectively following Yu et al. [21]. See Table 1 for statistics of benchmarks.

4.2 Compared Methods

We compare four methods: Toppop, PersonalToppop, Sets2Sets and DNNTP.

Toppop simply serves the items that are interacted the most across all users. PersonalToppop serves the items that the given user interacted with the most. Sets2Sets uses encoder-decoder framework to predict the next set [4]. Set embeddings are made by pooling operation and set-based attention is used to model temporal correlation relation. This method also models repeated elements. DNNTP is composed of three components: Element Relationship Learning (ERL), Temporal Dependency Learning (TDL) and Gated Information Fusing. ERL is simply a dynamic weighted graph neural networks. TDL captures temporal dependency. By Gated Information Fusing each user shares the embeddings of
uninteracted items. Hence DNNTPS can be seen as an ensemble of model which learns local information and Toppop model.

### 4.3 Results & Analyses

The results of our evaluation on three public benchmarks are shown in Table 2. We consider three metrics: Recall, Normalized Discounted Cumulative Gain (NDCG) and Personal Hit Ratio (PHR). PHR@K is calculated as 

\[
P_{PHR@K} = \frac{\sum_{i=1}^{N'} \mathbb{1}(\hat{S}_i \cap S_i > 0)}{N'}
\]

where \(N'\) is the number of test users, \(\hat{S}_i\) is the predicted top-K elements, and \(S_i\) is the ground truth set.

We trained VAE for 30 epochs and then trained DNNTPS for 30 epochs. We used only one layer for the VAEs across all benchmarks. For the decay factor in Equation (2), we set \(\tau = 0.6\). As illustrated in Figure 3, \(\tau = 0.6\) shows the best performance on TaoBao dataset. We empirically observe that \(\tau = 0.6\) could provide decent results across all metrics on the other datasets as well. The dimension of latent space is 128 for DC and TaoBao, and 512 for TMS. We used Adam optimizer [9] with learning rate 0.001.

Across all benchmarks and metrics, GLOIE with Tweedie output outperforms or is on par with all compared methods. Especially, GLOIE with Tweedie output outperforms the other methods on all metrics on DC and TaoBao datasets. We can see that GLOIE with Tweedie outperforms DNNTPS on every metric when \(K = 10\) which means that the embeddings learned by VAE are richly used as well as the embeddings learned by DNNTPS.

One thing to remark is that VAE with Tweedie output shows comparable performance to all compared methods on most metrics. Given that the number of items a user interacted with is 5.44, 4.96, and 18.05 respectively in DC, TaoBao, and TMS, this shows that VAE with Tweedie output captures the preference of users to non-interacted items.

To investigate the effectiveness of the attention-based integration method illustrated in Equation (8), we compared GLOIE with attention to the one with itemwise learnable weight similar to the method proposed in Yu et al. [21]. For overall datasets and metrics, we could observe performance gains: 3.09%, 0.45%, 0.92% improvement of NDCG@10 on DC, Taobao, and TMS, respectively.

Lastly, we note that the selection of output distribution of decoder on VAE largely affects the performance. We empirically show that both Gaussian and multinomial outputs do not fit temporal set prediction problems even though Gaussian is a popular choice for VAE and multinomial is a common choice in recommender system literature after the advent of VAECF [12].

### 5 CONCLUSION

This paper proposes Global-Local Item Embedding (GLOIE) that learns to utilize the temporal properties of sets across whole users as well as within a user. The proposed model learns global-local information by maximizing ELBO of the sum of time decayed vectors under the VAE framework and integrates local embeddings learned by dynamic graph
neural networks. As users with similar histories are reconstructed to close vectors, we could model the preference of the given user for a non-interacted item if other users with similar histories frequently interacted with the item. Data analysis and empirical results show that using Tweedie output for VAE’s decoder is effective for modeling temporal set prediction. The proposed method achieves state-of-the-art results by considering global information which is less explored in temporal set prediction literature.

Though the proposed VAE is powerful in itself, there is some room for improvement. Instead of using the sum of time decayed vector, we can model change of sets in continuous time by Neural ODE [1, 14]. Finding an appropriate form of prior that captures long-tailed distribution for the temporal set prediction can also be a future direction. Graph modality can also be used [8]. Last but not least, searching for better ways to integrating embeddings learned by VAE and embeddings learned by other algorithms which focus on local information is also an important future work.
