Temporal Heterogeneous Interaction Graph Embedding For Next-Item Recommendation

Yugang Ji[1], MingYang Yin[2], Yuan Fang[3], Hongxia Yang[2], Xiangwei Wang[2], Tianrui Jia[1], and Chuan Shi[1,4]✉

[1] Beijing University of Posts and Telecommunications
[2] Alibaba Group
[3] Singapore Management University
[4] Peng Cheng Laboratory
jiyugang@bupt.edu.cn, hengyang.yin@alibaba-inc.com, yfang@smu.edu.sg
{yang.yhx, florian.wxw}@alibaba-inc.com
{jiatianrui, shichuan}@bupt.edu.cn

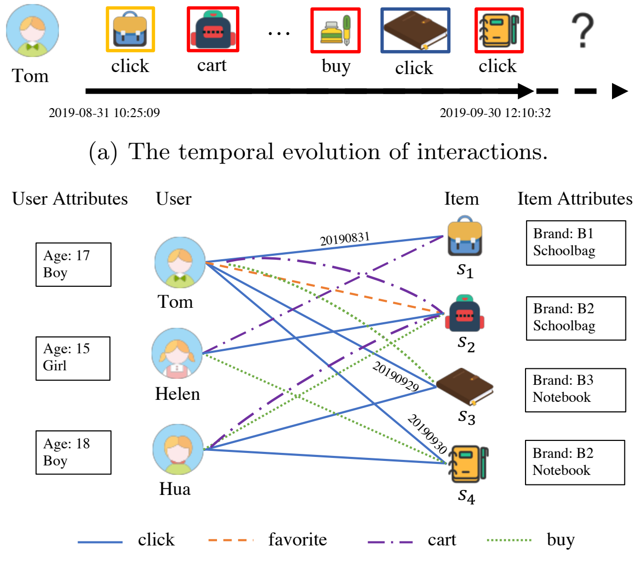
Abstract. In the scenario of next-item recommendation, previous methods attempt to model user preferences by capturing the evolution of sequential interactions. However, their sequential expression is often limited, without modeling complex dynamics that short-term demands can often be influenced by long-term habits. Moreover, few of them take into account the heterogeneous types of interaction between users and items. In this paper, we model such complex data as a *Temporal Heterogeneous Interaction Graph* (THIG) and learn both user and item embeddings on THIGs to address next-item recommendation. The main challenges involve two aspects: the *complex dynamics* and *rich heterogeneity* of interactions. We propose *THIG Embedding* (THIGE) which models the complex dynamics so that evolving short-term demands are guided by long-term historical habits, and leverages the rich heterogeneity to express the latent relevance of different-typed preferences. Extensive experiments on real-world datasets demonstrate that THIGE consistently outperforms the state-of-the-art methods.

Keywords: Temporal heterogeneous interaction graph · Next-item recommendation · Short-term demands · Long-term habits.

1 Introduction

With the prevalence of e-commerce, our shopping styles are revolutionized in recent years. By modeling historical user-item interactions, recommender systems play a fundamental role in e-commerce [9, 18]. In particular, the task of next-item recommendation—to predict the item that a user will interact with at the next time instance—not only caters to the business requirement of e-commerce platforms, but also enhances the user experience.

Earlier methods mainly exploit collaborative filtering (CF) [4, 14], which models interactions without any temporal dynamics. However, temporal evolution often contributes significantly to the next-item recommendation. As shown



(a) The temporal evolution of interactions.

(b) An example of THIG.

Fig. 1. Toy example of next-item recommendation, from (a) a temporal sequence of interactions, to (b) a Temporal Heterogeneous Interaction Graph (THIG).

in Fig. 1(a), Tom's current demand is more likely to be notebooks, rather than the general preference of bags which one would have concluded by analyzing his entire interaction history without temporal consideration. To capture such evolving demands of users, recurrent neural networks (RNN) [25, 8] have been widely used by considering a sequence of interactions. While RNNs are only capable of modelling short-term preferences (e.g., demands of notebooks) from relatively recent interactions, capturing long-term preferences (e.g., preferred brands) from historical habits is also an important element of temporal dynamics [18]. However, existing methods usually model short- and long-term preferences independently, ignoring the role of habits in driving the current, evolving demands. Taking Fig. 1(a) as an example, when browsing similar items (e.g., two schoolbags), users prefer to click those with attributes they habitually care (e.g., the brands).

This presents the first research challenge: *How to effectively model the complex temporal dynamics, coupling both historical habits and evolving demands?* In this paper, we model historical habits as the long-term preferences, and the current, evolving demands as the short-term preferences. More importantly, we propose to guide demand learning with historical habits, and develop a habit-guided attention to tightly couple both long- and short-term preferences.

Another dimension overlooked by existing sequential models is the abundant heterogeneous structural information. Taking Fig. 1(b) as an example, there exists large-scale inter-linking between various users and items, e.g., Hua bought the carted s_2, which cannot be explicitly modeled by separate interaction sequences. More importantly, there are also rich user-item interactions of heterogeneous types, such as "click", "favorite", "cart" and "buy". As shown in Fig. 1(b), Hua prefers items of brand B2 for the interactions of "buy" and "cart", while s_2 is more popular to boys for the various interactions rather than a single click of Helen. While heterogeneous graphs [15, 13, 16, 2] have been a *de facto* standard to model rich structural information and their representation learning have been studied extensively in heterogeneous network embedding and graph neural networks (GNN) [1, 21, 24], they ignore the complex temporal dynamics, treating the graph as a static snapshot. Furthermore, most of them treat the heterogeneous types of interaction independently, but in reality different types (e.g., clicks and buys) often express varying latent relevance w.r.t. each other. On the other hand, temporal graphs have been studied in some recent works in homogeneous settings [19, 10] without modeling the rich heterogeneity.

This leads to the second research challenge: *How to make full use of the temporal heterogeneous interactions to model the preferences of different types?* Here we propose a Temporal Heterogeneous Interaction Graph (THIG) to model the heterogeneous interactions and the temporal dimension jointly. Compared with static graphs, THIGs can express the evolving preference of users and the changing popularity of items; compared with temporal graphs, THIGs can exploit rich heterogeneous factors that simultaneously contribute to preference learning. Particularly, we design a novel heterogeneous self-attention mechanism to distill the latent relevance and multifaceted preferences from multiple interactions.

Hinged on the above insights, we propose THIGE, a novel model for **T**emporal **H**eterogeneous **I**nteraction **G**raph **E**mbedding, to effectively learn user and item embeddings on THIGs for next-item recommendation. In THIGE, we first encode heterogeneous interactions with temporal information. Building upon the temporal encoding, we take into account the influence of long-term habits on short-term demands, and design a habit-guided attention mechanism to couple short- and long-term preferences. To fully exploit the rich heterogeneous interactions to enhance multifaceted preferences, we futher capture the latent relevance of varying types of interaction via heterogeneous self-attention mechanisms.

We summarize the main contributions of this paper as follows.

- To our best knowledge, this work formulates and illustrates the first use of temporal heterogeneous interaction graphs for the problem of next-item recommendation. Different from previous sequential models which mainly focus on homogeneous sequences, here we fully utilize the structure information of multiple behaviors for item recommendation.
- We propose a novel model THIGE to couple long- and short-term preferences of heterogeneous nature, which fully exploits the temporal and heterogeneous interactions via habit-guided attention and heterogeneous self-attention. Both the dependence of heterogeneous preferences and impact from historical habits to recent demands are effectively modeled.

- We perform extensive experiments on the public datasets Yelp and CloudTheme, and the industrial dataset UserBehavior. We compare THIGE against various state of the arts and obtain promising results.

2 Related Work

We introduce the related work in two main domains, one of which is graph embedding and the other is next-item recommendation.

Graph neural networks. Graph neural networks (GNN) are widely used for node representation on real-world graphs including GCN [7], GraphSAGE [6] and GAT [20] to construct node embedding via neighborhood aggregation. Recently, taking the dynamic of links into considerations, Dyrep [19] and M^{2}DNE [10] split the graph into several snapshots to capture global evolution of local interests. Inspired by position embedding proposed in [13], it is promising to design the continuous-time function to generate time span-based temporal embedding. However, all these algorithms are proposed for homogeneous networks, which is not suitable to deal with graphs like THIGs in which nodes/edges are of multiple types. HetGCN [22] aggregates neighborhood information with meta-path guide random walks. HAN [21] proposes the hierarchical attention to guide aggregating heterogeneous neighborhood information. Furthermore, GATNE [1] combines heterogeneous information aggregated from different-typed neighborhoods via heterogeneous attention mechanisms. Furthermore, to fuse sequential information, MEIRec [5] captures evolution of users' same-typed interactions by LSTM. Unfortunately, this model is limited to the sequence length. How to model temporal heterogeneous graphs is challenging and meaningful [1].

Next-item recommendation. For dealing with next-item recommendation, sequence-based recommender systems are to understand the temporal dynamics between users and items [17]. The related works mainly focus on short-term interest learning. STAMP [9] captures the general and current interests by an attentive memory priority model. DIEN [25] respectively utilizes the classical and attentional GRUs on the interest extractor and evolving layers to capture short-term interest. Taking long-term preference into consideration, SHAN [23] designs the hierarchical attention to combine habits and demands modeling on hierarchical layers. M3R [18] mixes long-term, short-term and current interest modeling together to obtain the final interest. However, all of these sequential models cannot model the heterogeneous interactions. Le et al. [8] learn user preference from different-typed interactions by respectively modeling different-typed interactions with GRUs. Lv et al. [11] focus on capturing long-term preferences in different latent aspects like brands and categories. However, few of these models pay attention to the types of interactions within THIGs while different-typed interactions indicates various semantics.

3 Problem Formulation

Here we introduce the definition of THIGs and the problem of next-item recommendation on THIGs.

Definition 1 ***Temporal Heterogeneous Interaction Graph (THIG)***. *A THIG is $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathcal{A}, \mathcal{O}, \mathcal{R}, \phi, \psi)$, where $\mathcal{V}$ is a set of nodes with types $\mathcal{O}$, $\mathcal{E}$ is the set of edges with types $\mathcal{R}$, $\mathcal{T}$ is the set of timestamps on edges, and $\mathcal{A}$ is the set of attributes on nodes. Moreover, $\phi : \mathcal{V} \to \mathcal{O}$ is the node type mapping function and $\psi : \mathcal{E} \to \mathcal{R}$ is the edge type mapping function. In THIGs, $|\mathcal{O}| \geq 2$ and $|\mathcal{R}| \geq 2$.*

As shown in Figure 1(b), there are four types of interactions (edges), i.e., $\mathcal{R} = \{$click, favorite, cart, buy$\}$, between two types of nodes $\mathcal{O} = \{$user, item$\}$. A user may interact with the same item under multiple interactions like "click" and "buy" at different timestamps. Moreover, users and items contain their own features like age or brand. By modeling such heterogeneous interaction data with THIGs, richer semantics of dynamic interactions can be preserved for effective next-item recommendation.

Definition 2 ***Next-item recommendation on THIGs***. *On a THIG, a user u is associated with his/her historical interactions $\{(v_i, t_i, r_i) \mid 1 \leq i \leq n\}$ where the triple (v_i, t_i, r_i) denotes that item v_i is interacted under type r_i at time t_i, for some $t_n < T$ such that T is the current time. Similarly, an item v is associated with its historical interactions $\{(u_j, t_j, r_j) \mid 1 \leq j \leq n\}$ where the triple (u_j, t_j, r_j) denotes that item u_j interact with item v under type r_j at time t_j for some $t_n < T$. This task is to predict whether u will interact with v at the next time instance.*

For instance, as shown in Figure 1(b), given that Tom has clicked or bought S_1, S_2, S_3 and S_4 before, our goal is to predict the next item he will interact with, based on each candidate item's interaction history. This problem is fundamental and meaningful in e-commerce platforms to understand both user and items simultaneously.

4 Proposed Approach: THIGE

In this section, we propose our model called THIGE. We begin with an overview, before zooming into the details.

4.1 Overview

The overall framework is shown in Fig. 2. Specifically, we divide the historical interactions of a user into long and short term based on their timestamps. For short-term preferences, we model users' sequences of recent interactions with

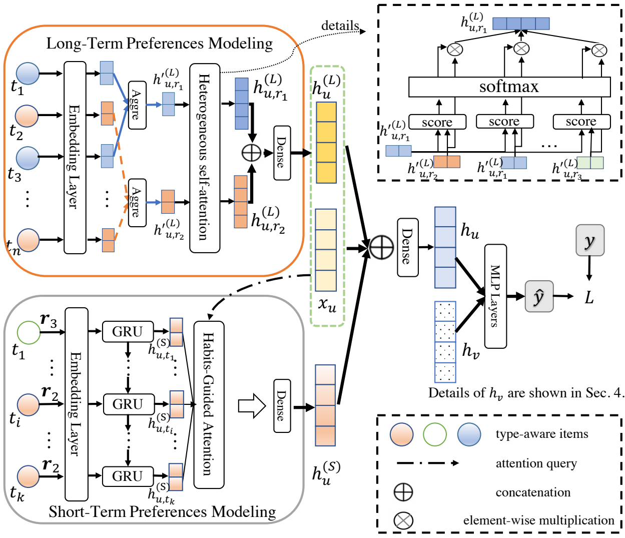


Fig. 2. Overall framework of user embedding in THIGE for next-item recommendation.

gated recurrent units (GRU), to embed users' current demands $\boldsymbol{h}_u^{(S)}$. For long-term preference, we model users' long-term interactions with a heterogeneous self-attention mechanism, to embed users' historical habits $\boldsymbol{h}_u^{(L)}$. Different from the decoupled combination (e.g., simple concatenation) of long- and short-term embeddings in previous methods [8, 23], we propose to exploit the long-term historical habits to guide the learning of short-term demands using the habit-guided attention, which effectively captures the impact of habits on recent behaviors.

Note that Fig. 2 only shows the learning of user representations. For items, we do not distinguish their long- and short-term interactions, and only adopt a long-term model similar to that of users. The reason lies in the fact that there may be numerous users interacting with an item around a short period of time, and these users have no significant short-term sequential dependency.

4.2 Embedding layer with temporal information

Each interacted item of a user is associated with not only attributes but also a timestamp. As shown in Fig. 2, the timestamps is in the form of $[t_1, t_2, \cdots, t_n]$.

Thus, the temporal embedding of an item v consists of both a static and a temporal component. The static component $\boldsymbol{x}_v = \boldsymbol{W}_{\phi(v)}\boldsymbol{a}_v$, where the input vector $\boldsymbol{a}_v \in \mathbb{R}^{d_\phi(v)}$ encodes the attributes of v, $\boldsymbol{W}_{\phi(v)} \in \mathbb{R}^{d \times d_\phi(v)}$ denotes the latent projection, $d_\phi(v)$ and d are the dimension of attributes and latent representation of v. Moreover, at time t, denoting Δt as the time span before the current time T and dividing the overall time span into B buckets, the temporal component of v is defined as $\boldsymbol{W}\xi(\Delta t)$, where $\xi(\Delta t) \in \mathbb{R}^B$ denotes the one-hot bucket representation of Δt, and $\boldsymbol{W} \in \mathbb{R}^{d_\mathcal{T} \times B}$ denotes the projection matrix and $d_{\mathcal{T}}$ is the output dimension. Thus, the temporal embedding of an item v at time t is

$$\boldsymbol{x}_{v,t} = [\boldsymbol{W}\xi(\mathit{\Delta} t) \oplus \boldsymbol{x}_v], \tag{1}$$

where $\oplus$ denotes concatenation. Similarly, we generate the static representation of a user u as $\boldsymbol{x}_u \in \mathbb{R}^d$, and temporal representation of u at time t as $\boldsymbol{x}_{u,t}$.

To further consider the sequential evolution of heterogeneous interactions, we generate the i^{th} interacted item embedding $\boldsymbol{x}_{v_i,t_i,r_i} = [\boldsymbol{x}_{v_i,t_i} \oplus \boldsymbol{r}_i]$ as the combination of the temporal embedding $\boldsymbol{x}_{v_i,t_i}$ and the corresponding type embedding $\boldsymbol{r}_i = \boldsymbol{W}_{\mathcal{R}}\boldsymbol{I}(r_i)$ where $\boldsymbol{I}(r_i)$ denotes the one-hot vector of r_i with dimension $|\mathcal{R}|$, $\boldsymbol{W}_{\mathcal{R}} \in \mathbb{R}^{d_{\mathcal{R}} \times |\mathcal{R}|}$ is the projection matrix and $d_{\mathcal{R}}$ is the latent dimension. For long-term preference modeling, we input the temporal embedding into type-aware aggregators to distinguish preferences of different types.

4.3 Short-term preference with habit-guidance

Recent interactions of users usually indicate the evolving current demands. For instance, as shown in Fig. 1(a), Tom's current demand has been evolved from bags to notebooks. In order to model the short-term and evolving preferences, we adopt gated recurrent units (GRU) [3], which can capture the dependency of recent interactions. Consider a user u here. Let his/her k recent interactions be $\{(v_i, t_i, r_i) \mid 1 \leq i \leq k\}$, where t_k is the most recent timestamp before the current time T. Subsequently, we encode the user preference at time t_i as $\boldsymbol{h}^{(S)}_{u,t_i}$, using a GRU based on the embedding of interaction (v_i, t_i, r_i), namely, $\boldsymbol{x}_{v_i,t_i,r_i}$, and his/her preference at t_{i-1}, as follows.

$$\boldsymbol{h}^{(S)}_{u,t_i} = \mathrm{GRU}(\boldsymbol{x}_{v_i,t_i,r_i}, \boldsymbol{h}^{(S)}_{u,t_{i-1}}), \quad \forall 1 < i \leq k, \tag{2}$$

where $\boldsymbol{h}^{(S)}_{u,t_i} \in \mathbb{R}^d$. The time-dependent user embeddings $\{\boldsymbol{h}^{(S)}_{u,t_i} | 1 \leq \ i \leq k\}$ can be further aggregated to encode the current demand of user u.

However, the current and evolving demands of user are not only influenced by their recent transactions. Their long-term preferences, i.e., historical habits such as brands and lifestyle inclinations, often play a subtle but important role. Thus, we enhance the encoding of short-term preferences under the guidance of historical habits, in order to discover more fine-grained and personalized preferences. Specially, we propose a habit-guided attention mechanism to aggregate short-term user preferences, as follows.

$$\boldsymbol{h}^{(S)}_{u} = \sigma\left(W^{(S)} \cdot \sum_i a_{u,i}\boldsymbol{h}^{(S)}_{u,t_i} + b_s\right), \quad \forall 1 \leq i \leq k, \tag{3}$$

where $\boldsymbol{h}^{(S)}_u \in \mathbb{R}^d$ denotes the overall short-term preference of u, $W^{(S)} \in \mathbb{R}^{d\times d}$ denotes the projection matrix, σ is the activation function and we adopt RELU here to ensure the non-linearity , b_s is the bias, and $a_{u,i}$ is the habit-guided weight:

$$a_{u,i} = \frac{\exp\left([\boldsymbol{h}_u^{(L)} \oplus \boldsymbol{x}_u]^T \boldsymbol{W}_a \boldsymbol{h}_{u,t_i}^{(S)}\right)}{\sum_{j=1}^{k} \exp\left([\boldsymbol{h}_u^{(L)} \oplus \boldsymbol{x}_u]^T \boldsymbol{W}_a \boldsymbol{h}_{u,t_j}^{(S)}\right)}, \tag{4}$$

where $\boldsymbol{h}_u^{(L)} \in \mathbb{R}^d$ is the long-term preference of u which would have encoded the habits of u, and $\boldsymbol{W}_a \in \mathbb{R}^{2d \times d}$ is a mapping to quantify the fine-grained relevance between the short-term preference of u at different times, and the long-term habits of u. Therefore, how to encode the long-term habits $\boldsymbol{h}_u^{(L)}$ in the context of heterogeneous interactions is the second key thesis of this work, as we will introduce next.

4.4 Long-term preference with heterogeneous interactions

Besides short-term preferences to encode current and evolving demands, users also exhibit long-term preferences to express personal and historical habits. In particular, there exist multiple types of heterogeneous interactions which have different relevance w.r.t. each other. For example, a "click" is more relevant to a "cart" or "buy" on the same item or similar items; "favorite" could be less relevant to "cart" or "buy", but is closely tied to the user's brand or lifestyle choices in the long run. Thus, different types of interactions entail both latent relevance and multifaceted preferences.

Thus, our goal is to fully encode the latent, fine-grained relevance of multifaceted long-term preferences.

Consider a user u, and his/her long-term interactions $\{(v_i, t_i, r_i) \mid 1 \leq i \leq n\}$ where $n \gg k$ (k is the count of recent interactions in short-term modeling). To differentiate the explicit interaction types, we first aggregate the embeddings of items which the user have interacted with under a specific type r:

$$\boldsymbol{h}'^{(L)}_{u,r} = \sigma\left(\boldsymbol{W}_r \cdot \text{aggre}(\{\boldsymbol{x}_{v_i,t_i} \mid 1 \leq i \leq n, r_i = r\})\right), \tag{5}$$

where $\boldsymbol{h}'^{(L)}_{u,r} \in \mathbb{R}^d$ is the type-r long-term preferences of user u, $\boldsymbol{W}_r \in \mathbb{R}^{d \times (d_{\mathcal{T}} + d)}$ is the type-r learnable mapping, aggre$(\cdot)$ is an aggregator, and we utilize mean-pooling here.

While we can simply sum or concatenate the type-specific long-term preferences into an overall representation, there exists latent relevance among the types (e.g., "click" and "buy"), and latent multifaceted preferences (e.g., brands and lifestyles). In this paper, we design a heterogeneous self-attention mechanism to express the latent relevance of different-typed interactions and long-term multifaceted preferences. By concatenating all long-term preferences of different types as $\boldsymbol{H}_u^{(L)} = \oplus_{r \in \mathcal{R}} \boldsymbol{h}'^{(L)}_{u,r}$ with size d-by-$|\mathcal{R}|$, we first formulate the self-attention to capture the latent relevant of heterogeneous types in $\mathcal{R}$ w.r.t. each other:

$$\boldsymbol{h}_{u,r}^{(L)} = \sum_{r' \in \mathcal{R}} \left(\frac{\exp\left(\boldsymbol{Q}_{u,r}^T \boldsymbol{K}_{u,r'} / \sqrt{d_a}\right)}{\sum_{r'' \in \mathcal{R}} \exp\left(\boldsymbol{Q}_{u,r}^T \boldsymbol{K}_{u,r''} / \sqrt{d_a}\right)} \boldsymbol{V}_{u,r'} \right), \tag{6}$$

where $\boldsymbol{Q}_u = \boldsymbol{W}_Q \boldsymbol{H}_u^{(L)}, \boldsymbol{K}_u = \boldsymbol{W}_K \boldsymbol{H}_u^{(L)}, \boldsymbol{V}_u = \boldsymbol{W}_V \boldsymbol{H}_u^{(L)}$, $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_a \times d}$ and $\boldsymbol{W}_V \in \mathbb{R}^{d \times d}$ are the projection matrices, and d_a is the dimension of keys and queries.

Next, to express multifaceted preferences, we adopt a multi-head approach to model latent, fine-grained facets. Specifically, the original embeddings of preferences are split into multi-heads and we adopt the self-attention for each head. The type-r long-term preference is concatenated from the h heads:

$$\boldsymbol{h}_{u,r}^{(L)} = \oplus_{m=1}^{m=h} \boldsymbol{h}_{u,r,m}^{(L)}, \tag{7}$$

where $\boldsymbol{h}_{u,r,m}^{(L)}$ denotes the m^{th} head based preference and there are h heads. The overall long-term preference can also be derived by fusing different types in $\mathcal{R}$:

$$\boldsymbol{h}_u^{(L)} = \sigma\left(\boldsymbol{W}^{(L)}(\oplus_{r \in \mathcal{R}} \boldsymbol{h}_{u,r}^{(L)}) + b_l\right), \tag{8}$$

where $\boldsymbol{W}^{(L)} \in \mathbb{R}^{d \times |\mathcal{R}| d}$ and b_l are the projection parameters. By now, both short- and long-term preferences haven been modeled. Taking the inherent attributes of users into consideration, the final representation of user u is calculated by

$$\boldsymbol{h}_u = \sigma(\boldsymbol{W}_{\boldsymbol{u}}[\boldsymbol{x}_u \oplus \boldsymbol{h}_u^{(S)} \oplus \boldsymbol{h}_u^{(L)}] + b_u), \tag{9}$$

where $\boldsymbol{h}_u \in \mathbb{R}^d$ will be used for next-item prediction, and $\boldsymbol{W}_{\boldsymbol{u}} \in \mathbb{R}^{d \times 3d}$ and b_u are learnable parameters.

4.5 Preference modeling of items

The temporal interactions of an item is significantly different from those of a user. In practice, on a mass e-commerce platform, it is typical that many users interact with the same item around the same time constantly, without a meaningful sequential effect among different users. In other words, it is more reasonable to only model the general, long-term popularity of items. Thus, we model item representation $\boldsymbol{h}_v^{(L)}$ similar to the long-term preference modeling of users in Eq. (8) with heterogeneous multi-head self-attention, and encode the item representation as follows:

$$\boldsymbol{h}_v = \sigma(\boldsymbol{W}_v[\boldsymbol{x}_v \oplus \boldsymbol{h}_v^{(L)}] + b_v), \tag{10}$$

where $\boldsymbol{h}_v \in \mathbb{R}^d$ is the final representation of item v for next-item prediction, and $\boldsymbol{W}_v$ and b_v are learnable parameters and $\boldsymbol{x}_v$ is the attribute vector of item v.

4.6 Optimization objective

To deal with next-item recommendation, we predict $\hat{y}_{u,v}$ between user u and item v, indicating whether u will interact with v (under a given type) at the next time. Here we utilize a Multi-Layer Perception (MLP) [12]:

$$\hat{y}_{u,v} = \text{sigmoid}(\text{MLP}(\boldsymbol{h}_u \oplus \boldsymbol{h}_v)), \tag{11}$$

Table 1. Description of datasets.

Dataset	**Yelp**	**CloudTheme**	**UserBehavior**
# User	103,569	144,197	533,974
# Item/Business	133,502	272,334	4,152,242
# Interaction	1,889,132	1,143,567	122,451,055
# Interaction type	2	2	4
# Training instance	611,568	865,182	3,203,844
(Training time span)	5 years	2 weeks	1 weeks
# Test instance	108,408	216,295	800,961
(Test time span)	next one quarter	next day	next day

where $\boldsymbol{h}_u$ and $\boldsymbol{h}_v$ are the final representation of user u and item v, respectively. Model parameters can be optimized with the following cross-entropy loss:

$$L = -\sum_{\langle u,v \rangle} (1 - y_{u,v}) \log(1 - \hat{y}_{u,v}) + y_{u,v} \log(\hat{y}_{u,v}), \tag{12}$$

where $\langle u, v \rangle$ is a sample of user u and item v, and $y_{u,v} \in \{0, 1\}$ is the ground truth of the sample. We also optimize the L2 regularization of latent parameters to ensure the robustness.

5 Experiments

In this section, we showcase the performances of our proposed THIGE[5] for next item-recommendation, and discuss the effectiveness of our design choices and key factors.

5.1 Datasets

We evaluate the empirical performance of THIGE for next-item recommendation on three real-world datasets including Yelp, CloudTheme and UserBehavior. The statistics are summarized in Table 1 and the details are introduced as follows.

- **Yelp**[6]: A public business dataset with two types of temporal interactions, namely, "review" and "tip" between users and businesses. Both users and businesses contain continuous and discrete features. We select interactions that happened before 14 Aug. 2019 as training data and the remaining as test data. For both training and test data, the last interacted business is labeled as positive instance, while five never interacted businesses of the same category are randomly sampled as negative instances.
- **CloudTheme**[7]: A public e-commerce dataset that records the click and purchase logs. Users and items are associated with embedding vectors given

[5] The source is available at https://github.com/yuduo93/THIGE
[6] https://www.yelp.com/dataset/documentation/main
[7] https://tianchi.aliyun.com/dataset/dataDetail?dataId=9716

by the dataset. We select interactions happened before the last day as training data and the remaining as test data. For both training and test data, we treat the last interacted item as the positive instance and randomly sample five other items of the same theme as negative instances.

- **UserBehavior**: An industrial dataset extracted from Taobao website, consisting of “click”, “favorite”, “cart” and “buy” interactions between users and items. For both training and test data, we utilize the actual feedback of users as labels—among the candidates displayed to users, we take the clicked item as the positive instance, and sample five other items from the remaining candidates as negative instances.

5.2 Baselines and experimental settings

We compare THIGE with six representative models and showcase the effectiveness of next-item recommendation and evaluate the effectiveness of our design choices. The baselines are listed as follows:

- **DIEN** [25] and **STAMP** [9] are two sequential models where the former is a hierarchical GRU to encode evolving interests and the latter is a short-term memory priority model to extract session-based interests;
- **SHAN** [23] and **M3R** [18] focus on modeling long-term interactions to enhance preference learning. SHAN adopts hierarchical attention mechanisms to fuse historical and recent interactions, while M3R models long- and short-term interests with GRUs and attention mechanisms respectively. The two methods treat short- and long-term preferences independently and combine their embeddings naïvely via concatenation or addition.
- **MEIRec** [5] and **GATNE** [1] are two heterogeneous GNN-based models. MEIRec focuses on aggregating information based on different meta-paths without paying attention to the relevance of meta-paths, while GATNE integrates multiple types of interactions with the attention mechanism but fails to model the dynamic.

For all baselines and our method, we set embedding size $d = 128$, $d_a = 128$, $d_{\mathcal{T}} = 16$, heads $h = 8$, the maximum iterations as 100, batch size as 128, learning rate as 0.001 and weight of regularization as 0.001 on all three datasets. The number of temporal buckets B is set as 60, 14, and 7 on the three datasets, respectively. For DIEN, MEIRec and our THIGE, we set three-layers MLP with dimensions 64, 32 and 1. For our THIGE and all baselines learning long- or short-term preferences, we consider the last 10, 10 and 50 interactions as the short term, and sample up to 50, 50 and 200 historical interactions as the long term for Yelp, CloudTheme and UserBehavior respectively. We will further analyze the impact of the length of the short term in Sect. 5.5.

We evaluate the performance of next-item recommendation with the metrics of F1, PR-AUC and ROC-AUC.

Table 2. Performance of next-item recommendation (with standard deviation). The best result is in bold while the second best is underlined. PR. denotes PR-AUC and ROC. denotes ROC-AUC.

Dataset	Metric	DIEN	STAMP	SHAN	M3R	MEIRec	GATNE	THIGE
Yelp	F1	39.52 (1.31)	40.37 (0.94)	40.17 (1.10)	33.49 (1.04)	42.86 (0.44)	42.21 (0.96)	**43.77** (0.66)
	PR.	30.04 (0.37)	31.36 (1.23)	32.35 (1.10)	26.40 (0.92)	32.69 (0.54)	33.39 (1.42)	**36.45** (1.66)
	ROC.	74.69 (0.57)	73.74 (1.15)	70.91 (1.14)	72.03 (1.33)	74.65 (0.23)	76.15 (0.64)	**79.23** (0.80)
CT.	F1	25.70 (1.25)	21.42 (0.91)	26.25 (1.09)	33.54 (1.67)	25.02 (0.98)	27.33 (0.50)	**37.17** (1.36)
	PR.	41.16 (0.22)	25.65 (0.44)	40.92 (1.09)	34.23 (0.95)	43.86 (0.42)	44.74 (0.20)	**51.94** (0.43)
	ROC.	68.41 (0.34)	52.97 (0.52)	67.48 (1.06)	62.92 (0.89)	69.98 (0.35)	71.22 (0.11)	**75.38** (0.33)
UB.	F1	67.32 (3.45)	63.06 (1.51)	58.84 (7.83)	61.37 (2.20)	66.48 (1.16)	**67.81** (1.14)	67.19 (0.98)
	PR.	63.38 (0.19)	59.09 (0.22)	63.86 (4.76)	57.68 (0.03)	64.94 (0.15)	65.42 (0.05)	**65.71** (0.09)
	ROC.	62.90 (0.23)	58.29 (0.40)	55.45 (3.98)	57.82 (0.09)	64.82 (0.16)	65.06 (0.08)	**65.39** (0.06)

5.3 Comparison with baselines

We report the results of different methods for next-item recommendation in Table 2. In general, THIGE achieves the best performance on the three datasets, outperform the second best method by 4.04% on Yelp, 5.84% on CloudTheme and 0.51% on UserBehavior.

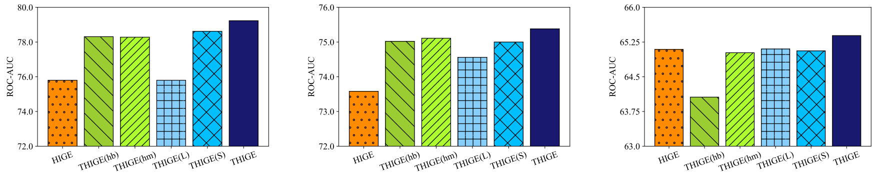
Compared with sequential models (DIEN, STAMP, SHAN and M3R), the reason that THIGE is superior is twofold. First, THIGE designs a more effective way to integrate long- and short-term preferences, such that the current demands are explicitly guided by historical habits. Second, it also considers different types of interactions between users and items, leading to better performance.

Compared with GNN-based models (MEIRec and GATNE), the main improvement of THIGE comes from jointly modeling historical habits and evolving demands. Moreover, MEIRec models heterogeneous interactions in an entirely decoupled manner, whereas GATNE and THIGE achieves better performance by modeling their latent relevance. It is also not surprising that heterogeneous GNN-based methods typically outperform sequential models, as the former accounts for multi-typed interactions whereas the latter only models single-typed interactions.

5.4 Comparison of model variants

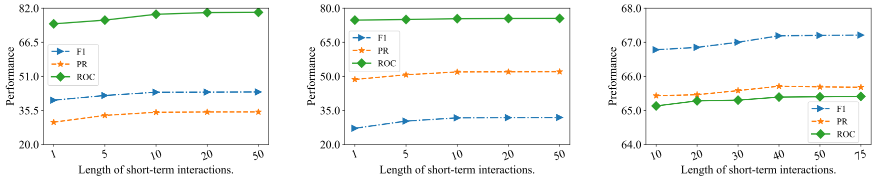
In this section, we analyze three categories of THIGE variants to evaluate the effectiveness of our design choices, as follows.

- Attention effect: **THIGE(hm)** only uses the heterogeneous multi-head self-attention, without the habit-guided attention; **THIGE(hb)** is the opposite, using only the habit-guided attention.
- Range of preferences: **THIGE(L)** models long range only while **THIGE(S)** models short rane only;
- Temporal effect: **HIGE** removes the temporal dimension from THIGE, treating the graph as static.



(a) Yelp (b) CloudTheme (c) UserBehavior

Fig. 3. Performance comparison of THIGE and its variants.



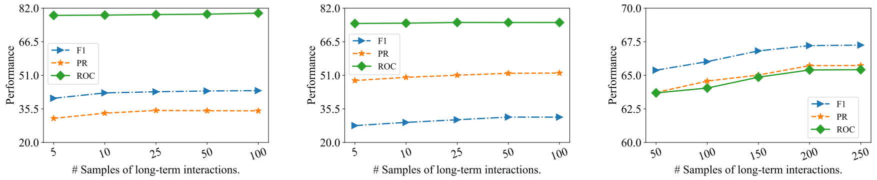
(a) Yelp (b) CloudTheme (c) UserBehavior

Fig. 4. Analyzing the length of short-term interactions in THIGE.

5.5 Analysis of key factors in THIGE

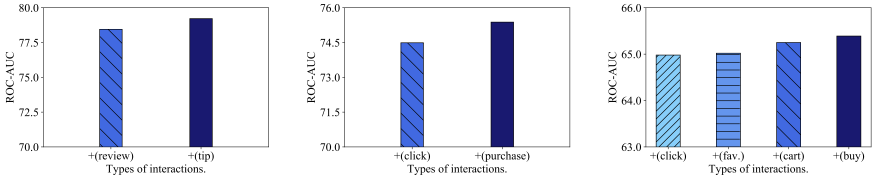
As shown in Fig. 3, our THIGE outperforms all three categories of variants. We make the following observations. (1) Compared with THIGE(hb) and THIGE(hm), THIGE models the latent relevance of heterogeneous preferences in a fine-grained manner, and model the impact of historical habits on current demands, leading to better performance. (2) Compared with THIGE(S) and THIGE(L), the joint modeling of short- and long-term preferences can improve performance, which also validates the assumption that the immediate decision of users is guided by their historical habits. (3) Compared with HIGE, the improvement in THIGE demonstrates the effectiveness of temporal embedding.

In THIGE, there are four key factors that may significantly affect the model performance: the length of short-term interactions, the samples of long-term interactions, the types of interactions and the number of latent preferences (i.e., the number of heads). In Fig. 4, we investigate how the length of the short term would impact the model. We respectively fix the samples of long-term interactions as 50, 50 and 200 for the three datasets, and then adjust the length of short-term interactions. Taking Fig. 4(c) as an example, we vary the length between 10 and 75 (i.e., treat the last 10–75 actions as the short term). When the length initially increases, the performance of THIGE is continuously improved, until reaching a saturation at about 40 or 50. Further treating more interactions as short-term has no additional benefit, which is expected as they can no longer be considered as current demands. This also justifies that the long-term, historical actions must be modeled differently to capture the user habit; simply extending the length of the short term does not work.



(a) Yelp (b) CloudTheme (c) UserBehavior

Fig. 5. Analyzing the samples of long-term interactions in THIGE.



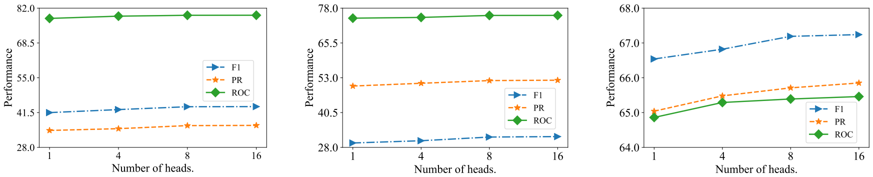
(a) Yelp (b) CloudTheme (c) UserBehavior

Fig. 6. Analyzing the types of interactions in THIGE.

Next, in Fig. 5, we focus on detecting the influence of the length of long-term interactions. We respectively fix the length of short-term interactions as 10, 10 and 50 for the three datasets, and then vary the corresponding samples of long-term interactions. It is obvious that all improvements in performance are continuous but gradually weakened. There are two main reasons resulting in such phenomenons. On the one hand, with the length increases, the whole historical interactions of more and more users are captured and modeled. On the other hand, users who contain too many interactions may be abnormal and introduce noise that limits performance.

Furthermore, in Fig. 6, we demonstrate the contribution from different types of interactions. Taking UserBehavior as an example, we progressively include the interactions of “click”, “favorite”, “cart” and “buy”, one at a time. As an example, the performance in Fig 6(c) gradually improves, implying the effectiveness to integrate heterogeneous interactions. Moreover, comparing “favorite” and “cart”, the “favorite” action has a smaller marginal return than “cart”, which is not surprising given that “favorite” only have a weak tendency to induce future “cart”, whereas “cart” actions are more likely to lead to purchases. That also means different types of interactions cannot be treated independently. Thus, modeling the relevance of different-typed interactions plays an important role.

Moreover, since the number of heads h in heterogeneous multi-head self-attention mechanism reflects the number of latent preferences like categories and brands, we also vary the number of heads from 1 to 16 on the three datasets to analyze the influence of h in THIGE. The experimental results in Fig 7 indicate that $h = 8$ is a generally suitable and robust choice.



(a) Yelp (b) CloudTheme (c) UserBehavior

Fig. 7. Analyzing the number of heads in THIGE.

6 Conclusion

In this paper, we study the problem of representation learning on THIGs for next-item recommendation. To make full use of dynamic and heterogeneous information, and propose the THIGE to model short- and long-term preferences through habit-guided and heterogeneous self-attention mechanisms. The extensive experimental results on three real-world datasets demonstrate the effectiveness of our proposed model.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (No. 61772082, 61806020, 61702296), the National Key Research and Development Program of China (2018YFB1402600), and Alibaba Group under its Alibaba Innovative Research (AIR) programme. This research is also supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This work was done when Yugang Ji was visiting Alibaba as a research intern.

References

1. Cen, Y., Zou, X., Zhang, J., et al.: Representation learning for attributed multiplex heterogeneous network. In: Proceedings of SIGKDD 2019, pp. 1358–1368 (2019)
2. Chang, S., Han, W., Tang, J., et al.: Heterogeneous network embedding via deep architectures. In: Proceedings of SIGKDD 2015, pp. 119–128 (2015)
3. Cho, K., Van Merriënboer, B., Gulcehre, C., et al.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: Proceedings of EMNLP 2014, pp. 1724–1734 (2014)
4. Ekstrand, M.D., Riedl, J.T., Konstan, J.A., et al.: Collaborative filtering recommender systems. Now Publishers Inc **4**(2), 81–173 (2011)
5. Fan, S., Zhu, J., Han, X., et al.: Metapath-guided heterogeneous graph neural network for intent recommendation. In: Proceedings of SIGKDD 2019, pp. 2478–2486 (2019)

6. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Proceedings of NeurIPS 2017, pp. 1024–1034 (2017)
7. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of ICLR 2017 (2017)
8. LE, D.T., LAUW, H.W., Fang, Y.: Modeling contemporaneous basket sequences with twin networks for next-item recommendation. In: IJCAI 2018, pp. 3414–3420 (2018)
9. Liu, Q., Zeng, Y., Mokhosi, R., Zhang, H.: Stamp: short-term attention/memory priority model for session-based recommendation. In: Proceedings of SIGKDD 2018, pp. 1831–1839 (2018)
10. Lu, Y., Wang, X., Shi, C., et al.: Temporal network embedding with micro-and macro-dynamics. In: Proceedings of CIKM 2019, pp. 469–478 (2019)
11. Lv, F., Jin, T., Yu, C., et al.: Sdm: Sequential deep matching model for online large-scale recommender system. In: Proceedings of CIKM 2019, pp. 2635–2643 (2019)
12. Pal, S.K., Mitra, S.: Multilayer perceptron, fuzzy sets, and classification. IEEE Transactions on neural networks **3**(5), 683–697 (1992)
13. Qu, M., Tang, J., Shang, J., et al.: An attention-based collaboration framework for multi-view network representation learning. In: Proceedings of CIKM 2017, pp. 1767–1776 (2017)
14. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of WWW 2001, pp. 285–295 (2001)
15. Shi, C., Li, Y., Zhang, J., Sun, Y., Philip, S.Y.: A survey of heterogeneous information network analysis. IEEE TKDE 2016 **29**(1), 17–37 (2016)
16. Shi, Y., Han, F., He, X., et al.: mvn2vec: Preservation and collaboration in multi-view network embedding. arXiv preprint arXiv:1801.06597 (2018)
17. Song, Y., Elkahky, A.M., He, X.: Multi-rate deep learning for temporal recommendation. In: Proceedings of SIGIR 2016, pp. 909–912 (2016)
18. Tang, J., Belletti, F., Jain, S., et al.: Towards neural mixture recommender for long range dependent user sequences. In: Proceedings of WWW 2019, pp. 1782–1793 (2019)
19. Trivedi, R., Farajtabar, M., Biswal, P., Zha, H.: Representation learning over dynamic graphs. In: Proceedings of ICLR 2019 (2019)
20. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: Proceedings of ICLR (2018)
21. Wang, X., Ji, H., Shi, C., et al.: Heterogeneous graph attention network. In: Proceedings of WWW 2019, pp. 2022–2032 (2019)
22. Wang, Y., Duan, Z., Liao, B., et al.: Heterogeneous attributed network embedding with graph convolutional networks. Methods **25**(50), 75 (2019)
23. Ying, H., Zhuang, F., Zhang, F., et al.: Sequential recommender system based on hierarchical attention networks. In: Proceedings of AAAI 2018, pp. 3926–3932 (2018)
24. Zheng, V.W., Sha, M., Li, Y., et al.: Heterogeneous embedding propagation for large-scale e-commerce user alignment. In: Proceedings of ICDM 2018, pp. 1434–1439 (2018)
25. Zhou, G., Mou, N., Fan, Y., et al.: Deep interest evolution network for click-through rate prediction. In: Proceedings of AAAI 2019, pp. 5941–5948 (2019)