Target Interest Distillation for Multi-Interest Recommendation

Chenyang Wang
DCST, BNRist, Tsinghua University
Beijing 100084, China
wangcy18@mails.tsinghua.edu.cn

Zhefan Wang
DCST, BNRist, Tsinghua University
Beijing 100084, China
wzf19@mails.tsinghua.edu.cn

Yankai Liu
China Mobile Research Institute &
THU-CMCC Joint Institute
Beijing 100084, China
liuyankai@chinamobile.com

Yang Ge
China Mobile Research Institute
Beijing 100084, China
geyang100299@163.com

Weizhi Ma
AIR, Tsinghua University
Beijing 100084, China
mawz@tsinghua.edu.cn

Min Zhang*
DCST, BNRist, Tsinghua University &
THU-CMCC Joint Institute
Beijing 100084, China
z-m@tsinghua.edu.cn

Yiqun Liu
DCST, BNRist, Tsinghua University
Beijing 100084, China
yiqunliu@tsinghua.edu.cn

Junlan Feng
China Mobile Research Institute
Beijing 100084, China
fengjunlan@chinamobile.com

Chao Deng
China Mobile Research Institute
Beijing 100084, China
dengchao@chinamobile.com

Shaoping Ma
DCST, BNRist, Tsinghua University
Beijing 100084, China
msp@tsinghua.edu.cn

——CIKM 2022

https://github.com/THUwangcy/ReChorus/tree/CIKM22

gesis Leibniz-Institut für Sozialwissenschaften

Forschungszentrum · Research Center L3S

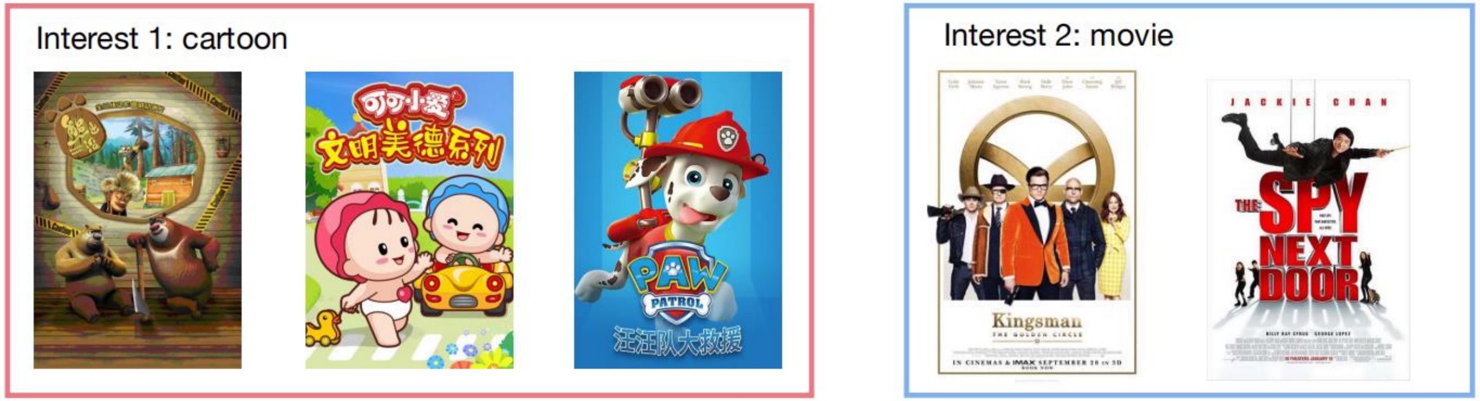
Reported by Yabo Yin

1.Introduction

2.Method

3.Experiments

Introduction



1. Previous work uses the best matching interest for each candidate item to calculate the ranking score, neglecting the target interest distribution in different contexts.

2. There is generally no labelling data for the actual user interest, which makes it hard to provide appropriate supervision signals to the target-interest predictor.

Method

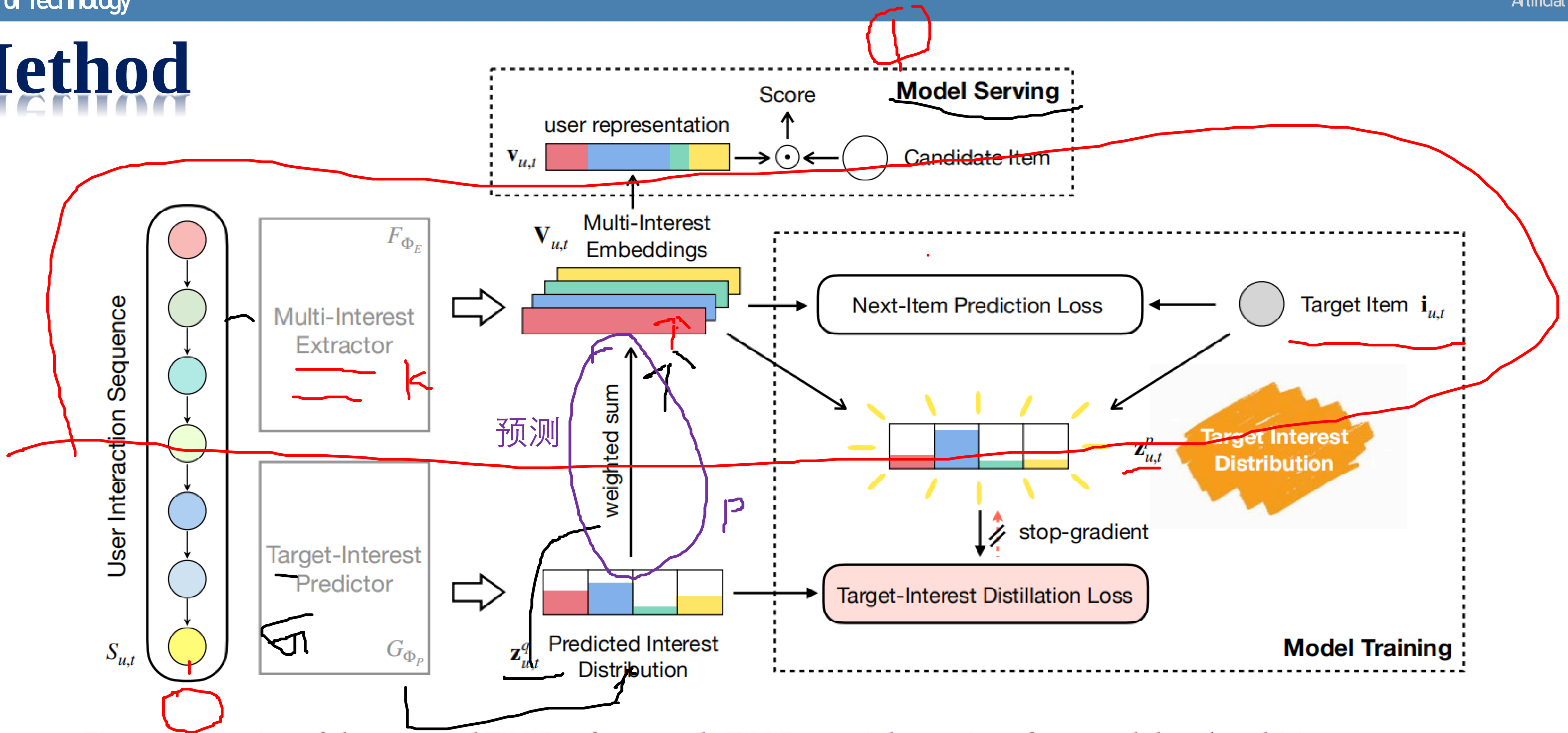


Figure 1: Overview of the proposed TiMiRec framework. TiMiRec mainly consists of two modules: 1) multi-interest extractor and 2) target-interest predictor. The former derives multiple interest embeddings from a user's interaction sequence, while the latter gives the predicted interest distribution in the current context. Then we use the predicted interest distribution to aggregate multi-interest embeddings and calculate the next-item prediction loss. Besides, a target-interest distillation loss is devised to instruct the target-interest predictor, where the soft label of the target interest is derived by the compatibility (cosine similarity) between the target item and multi-interest embeddings, serving as an additional supervision signal.

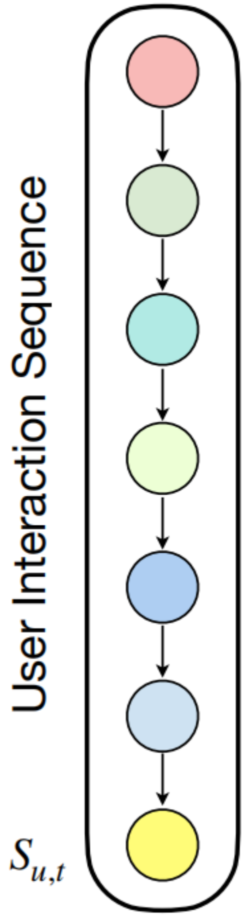
Method

Preliminaries

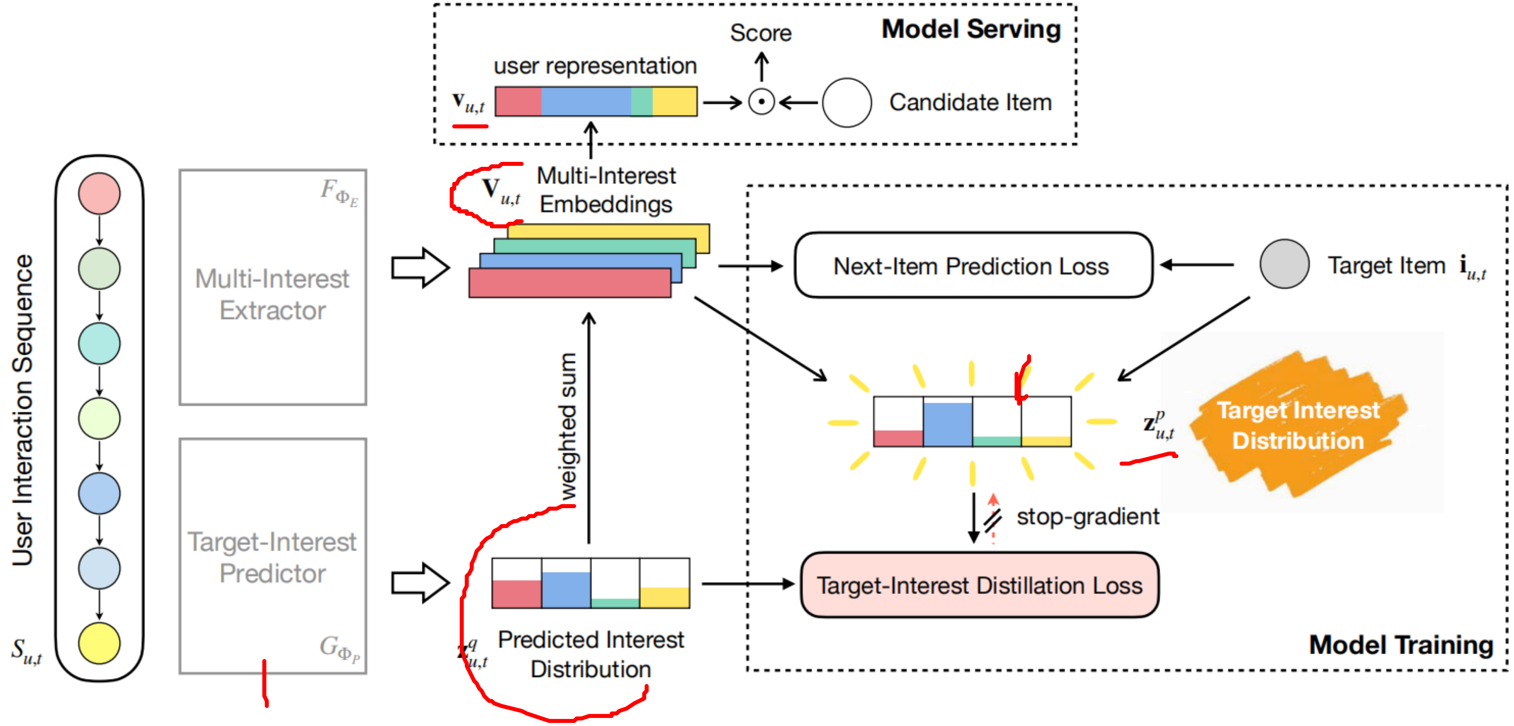
Let $\mathcal{U}$ and $\mathcal{I}$ denote the user and item set,

ordered list $[i_{u,1}, i_{u,2}, \cdots, i_{u,N_u}]$, where each element $i_{u,t} \in \mathcal{I}$

given the historical sequence before the target time step t, denoted as $S_{u,t}$



Method



Pretrain multi-interest extractor

$$\mathbf{V}_{u,t} = [\mathbf{v}^1_{u,t}, \cdots, \mathbf{v}^K_{u,t}] \in \mathbb{R}^{D\times K}, \tag{1}$$

$$\mathbf{v}_{u,t} = \mathbf{V}_{u,t}[:, \operatorname{argmax}\left(\mathbf{V}^T_{u,t}\mathbf{i}_{u,t}\right)], \tag{2}$$

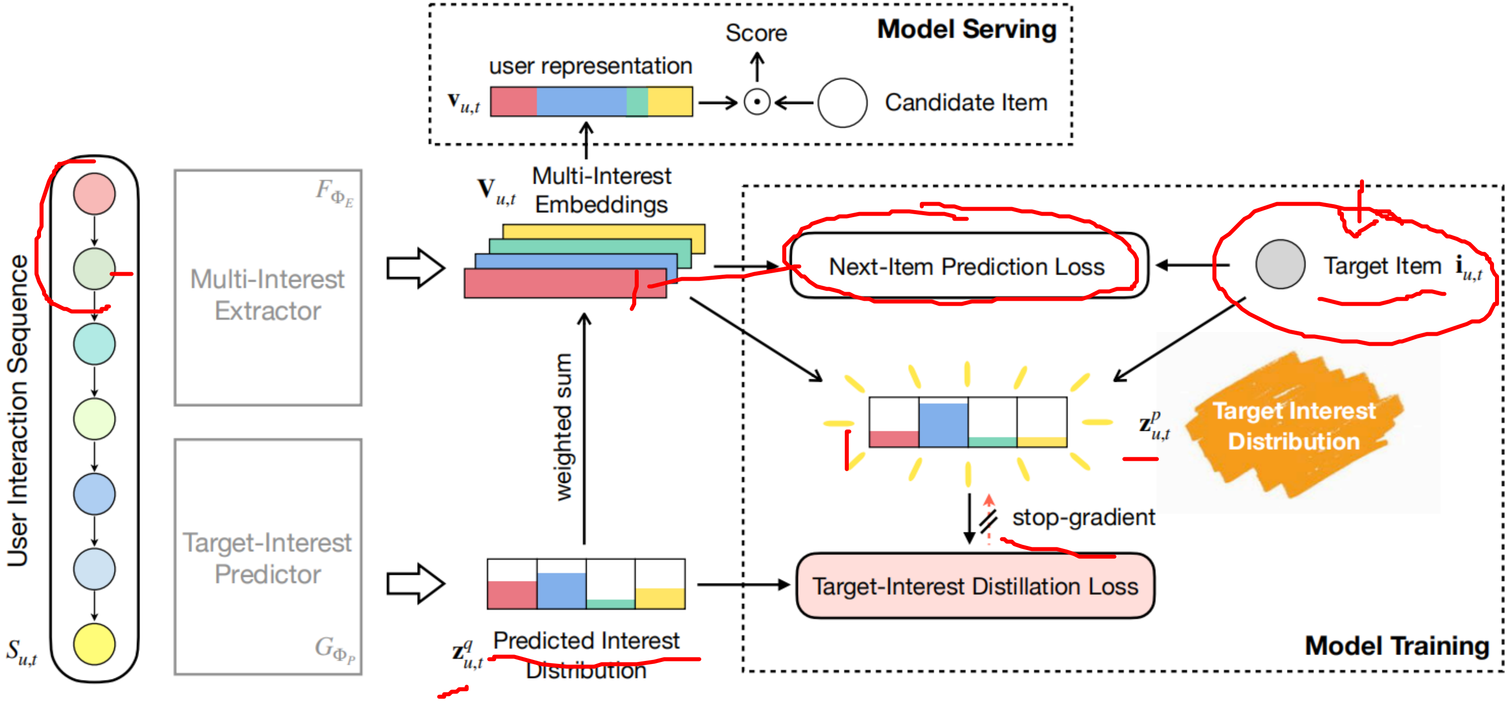
where $\mathbf{i}_{u,t}$ is the embedding of the candidate item.

$$\mathbf{z}^q_{u,t} = G_{\Phi_P}(S_{u,t}) \in \mathbb{R}^K, \tag{3}$$

$$\mathbf{v}_{u,t} = \mathbf{V}_{u,t}\ \mathrm{softmax}(\mathbf{z}^q_{u,t}). \tag{4}$$

$$\mathbf{z}^p_{u,t} = \mathrm{sim}(\mathbf{V}_{u,t}, \mathbf{i}_{u,t}) = \left[\frac{\mathbf{v}^1_{u,t}}{||\mathbf{v}^1_{u,t}||_2}\frac{\mathbf{i}_{u,t}}{||\mathbf{i}_{u,t}||_2}, \cdots, \frac{\mathbf{v}^K_{u,t}}{||\mathbf{v}^K_{u,t}||_2}\frac{\mathbf{i}_{u,t}}{||\mathbf{i}_{u,t}||_2}\right]. \tag{5}$$

Method



$$\mathbf{q}_{u,t} = \frac{\exp(\mathbf{z}^q_{u,t} \;/\; T)}{\sum_{k=1}^{K} \exp(\mathbf{z}^q_{u,t}[k] \;/\; T)}, \tag{6}$$

$$\mathbf{p}_{u,t} = \frac{\exp(\mathbf{z}^p_{u,t} \;/\; T)}{\sum_{k=1}^{K} \exp(\mathbf{z}^p_{u,t}[k] \;/\; T)}. \tag{7}$$

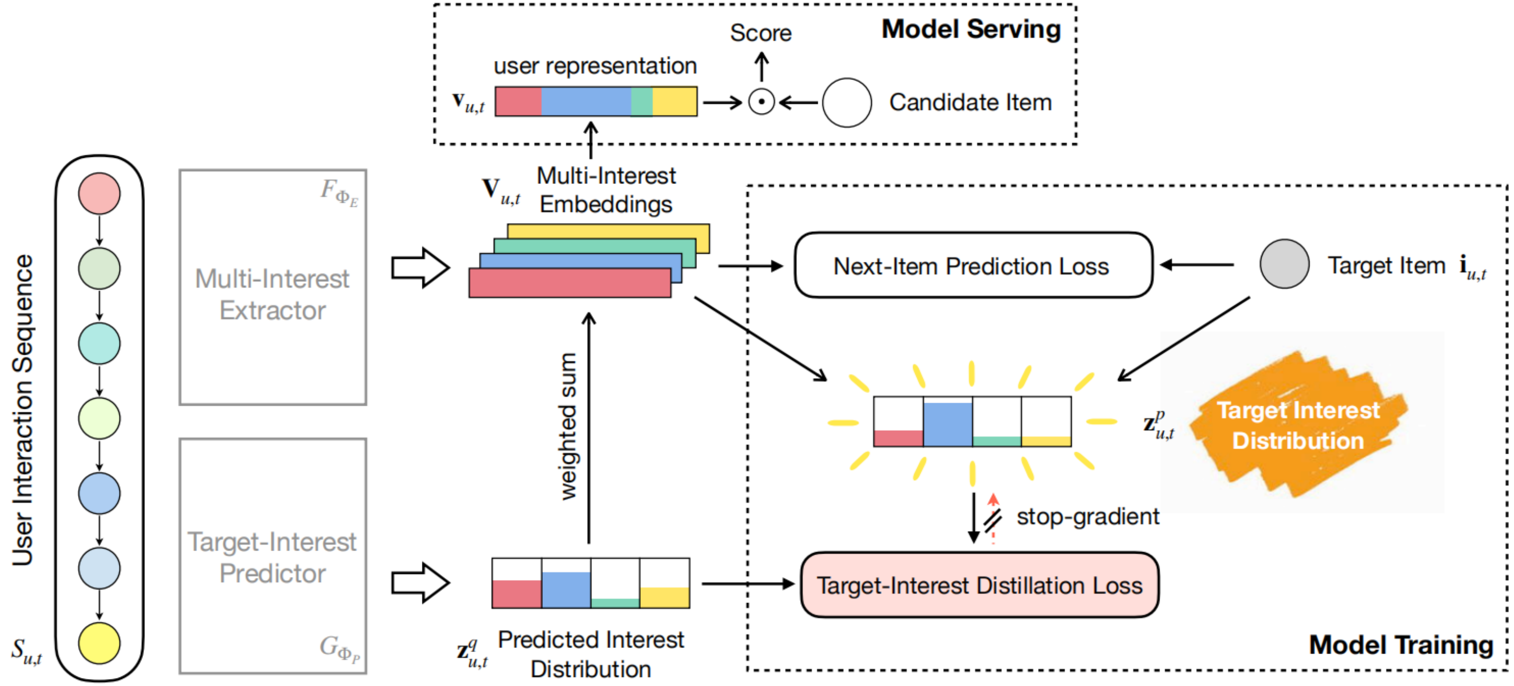
$$\mathcal{L}'_{\text{distill}} = -\sum_{u\in\mathcal{U}} \sum_{t=2}^{N_u} \mathbf{p}^T_{u,t} \log(\mathbf{q}_{u,t}). \tag{8}$$

$$\mathcal{L}_{\text{distill}} = -\sum_{u\in\mathcal{U}} \sum_{t=2}^{N_u} \text{stopgrad}(\mathbf{p}^T_{u,t}) \log(\mathbf{q}_{u,t}). \tag{9}$$

$$\mathcal{L}_{\text{rec}} = -\sum_{u\in\mathcal{U}} \sum_{t} \log \sigma\left(\mathbf{v}^T_{u,t}\mathbf{i}_{u,t} - \mathbf{v}^T_{u,t}\mathbf{i}^-_{u,t}\right), \tag{10}$$

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + T^2 \mathcal{L}_{\text{distill}}. \tag{11}$$

Method



Multi-Interest Extractor.

$S_{u,t}$ with length n is first transformed into embeddings $\mathbf{H} \in \mathbb{R}^{D\times n}$

$$\mathbf{A} = \mathrm{softmax}\left(\mathbf{W}_2^T \tanh(\mathbf{W}_1\mathbf{H})\right)^T, \tag{12}$$

$A \in \mathbb{R}^{n\times K}$ $\quad$ $\mathbf{W}_1$ $D_a \times D$ $\quad$ $\mathbf{W}_2$ $D_a \times K$.

$$\mathbf{V}_{u,t} = \mathbf{HA}. \tag{13}$$

Target-Interest Predictor.

$$\mathbf{s}_{u,t} = \mathrm{GRU}(\mathbf{H}'), \tag{14}$$

where $\mathbf{H}'$ is transformed from $S_{u,t}$ with another embedding layer.

$$\mathbf{z}^q_{u,t} = \mathbf{W}^q_2 \cdot \mathrm{ReLU}\left(\mathbf{W}^q_1 \cdot \mathbf{s}_{u,t} + \mathbf{b}_1\right) + \mathbf{b}_2, \tag{15}$$

where $\mathbf{W}^q_1 \in \mathbb{R}^{D\times D}, \mathbf{W}^q_2 \in \mathbb{R}^{K\times D}, \mathbf{b}_1 \in \mathbb{R}^D, \mathbf{b}_2 \in \mathbb{R}^K$

Experiments

Table 1: Statistics of datasets.

Dataset	#user ($\lvert\mathcal{U}\rvert$)	#item ($\lvert\mathcal{I}\rvert$)	#inter ($\sum_u N_u$)	density
Beauty	22,363	12,101	198,502	0.07%
MovieLens	6,040	3,706	1,000,209	4.47%
CMCC	49,847	29,074	1,300,351	0.09%

Experiments

Table 2: Top-*K* recommendation performance on the three datasets. TiMiRec and TiMiRec+ adopt ComiRec and ComiRec+ as the multi-interest extractor, respectively. The best results within the same set of methods are in bold face, and the overall best results are underlined. The superscripts * and ** indicate $p \leq 0.05$ and $p \leq 0.01$ for the paired t-test of TiMiRec/TiMiRec+ vs. the best baseline within the corresponding model set.

Setting		Models without Transformer Layer					Models with Transformer Layer			
Dataset	Metric	GRU4Rec	YouTube	MIND	ComiRec	TiMiRec	SASRec	TiSASRec	ComiRec+	TiMiRec+
Beauty	H@5	0.1072	0.1040	0.1193	0.1257	**0.1437****	0.1435	0.1529	0.1546	**0.1573***
	H@10	0.1552	0.1563	0.1727	0.1832	**0.2006****	0.2058	0.2084	0.2123	**0.2196***
	H@20	0.2107	0.2264	0.2492	0.2543	**0.2645****	0.2706	0.2760	0.2809	**0.2887****
	N@5	0.0719	0.0702	0.0809	0.0852	**0.1006****	0.1004	0.1087	0.1095	**0.1112***
	N@10	0.0873	0.0870	0.0981	0.1038	**0.1118****	0.1192	0.1266	0.1272	**0.1313***
	N@20	0.1013	0.1046	0.1173	0.1217	**0.1350****	0.1356	0.1436	0.1459	**0.1488***
MovieLens	H@5	0.2730	0.2336	0.1863	0.2513	**0.3091****	0.3124	0.3212	0.2745	**0.3333****
	H@10	0.3964	0.3406	0.2881	0.3659	**0.4310****	0.4407	0.4397	0.3906	**0.4556****
	H@20	0.5323	0.4719	0.4152	0.4937	**0.5625****	0.5674	0.5712	0.5091	**0.5843****
	N@5	0.1875	0.1597	0.1229	0.1708	**0.2136****	0.2177	0.2241	0.1875	**0.2346****
	N@10	0.2273	0.1942	0.1558	0.2078	**0.2529****	0.2593	0.2625	0.2249	**0.2741****
	N@20	0.2616	0.2274	0.1877	0.2400	**0.2861****	0.2910	0.2956	0.2549	**0.3067****
CMCC	H@5	0.3978	0.4170	0.4229	0.4547	**0.4812****	0.4681	0.4768	0.4831	**0.4886***
	H@10	0.5121	0.5328	0.5381	0.5716	**0.5934****	0.5828	0.5882	0.5960	**0.6020****
	H@20	0.6306	0.6453	0.6533	0.6845	**0.7018****	0.6853	0.6937	0.6997	**0.7091****
	N@5	0.2916	0.3064	0.3119	0.3356	**0.3636****	0.3533	0.3615	0.3662	**0.3690***
	N@10	0.3286	0.3438	0.3492	0.3735	**0.3999****	0.3905	0.3975	0.4027	**0.4057***
	N@20	0.3587	0.3723	0.3784	0.4020	**0.4273****	0.4164	0.4242	0.4290	**0.4329***

Experiments

Table 4: Performance of TiMiRec variants.

Method	Beauty		MovieLens	
	H@10	N@10	H@10	N@10
joint train	0.1787	0.1030	0.4267	0.2483
w/o stopgrad	0.1971	**0.1157**	0.4260	0.2468
TiMiRec	**0.2006**	0.1118	**0.4310**	**0.2529**

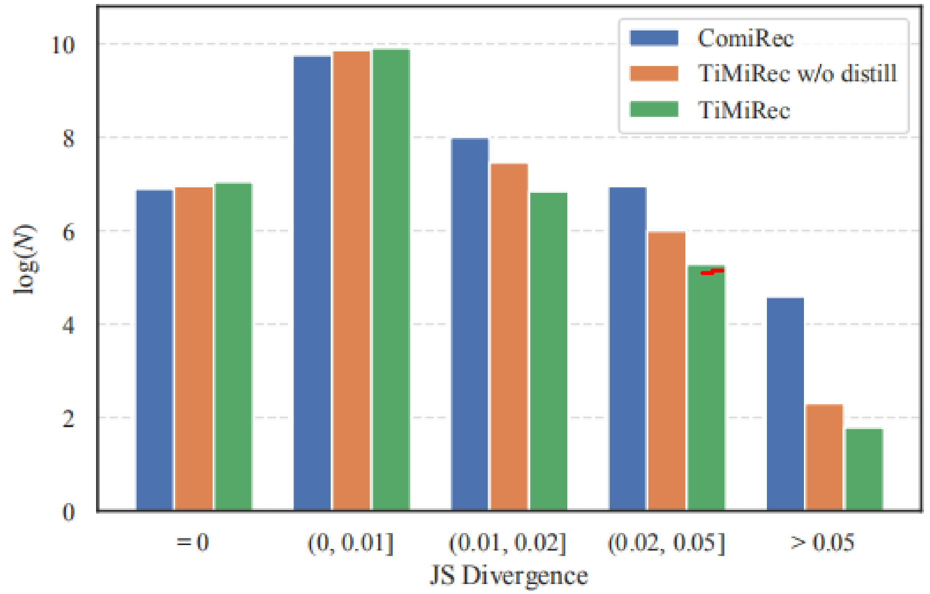


Figure 2: Distribution of Jensen–Shannon divergence between interest distributions of the target item and the top-1 recommended item for different methods on the test set.

Experiments

Table 4: Performance of TiMiRec variants.

Method	Beauty		MovieLens	
	H@10	N@10	H@10	N@10
joint train	0.1787	0.1030	0.4267	0.2483
w/o stopgrad	0.1971	**0.1157**	0.4260	0.2468
TiMiRec	**0.2006**	0.1118	**0.4310**	**0.2529**

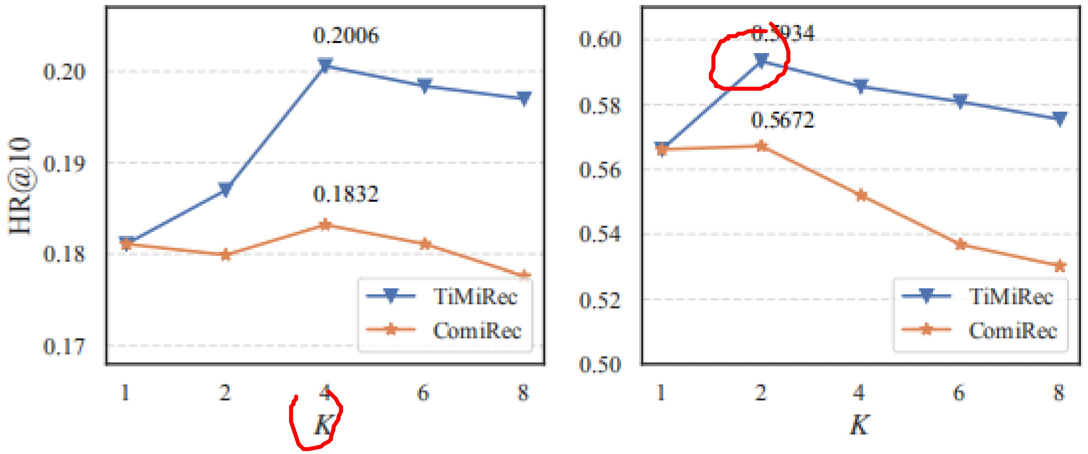


Figure 3: Parameter sensitivity analysis.

Experiments

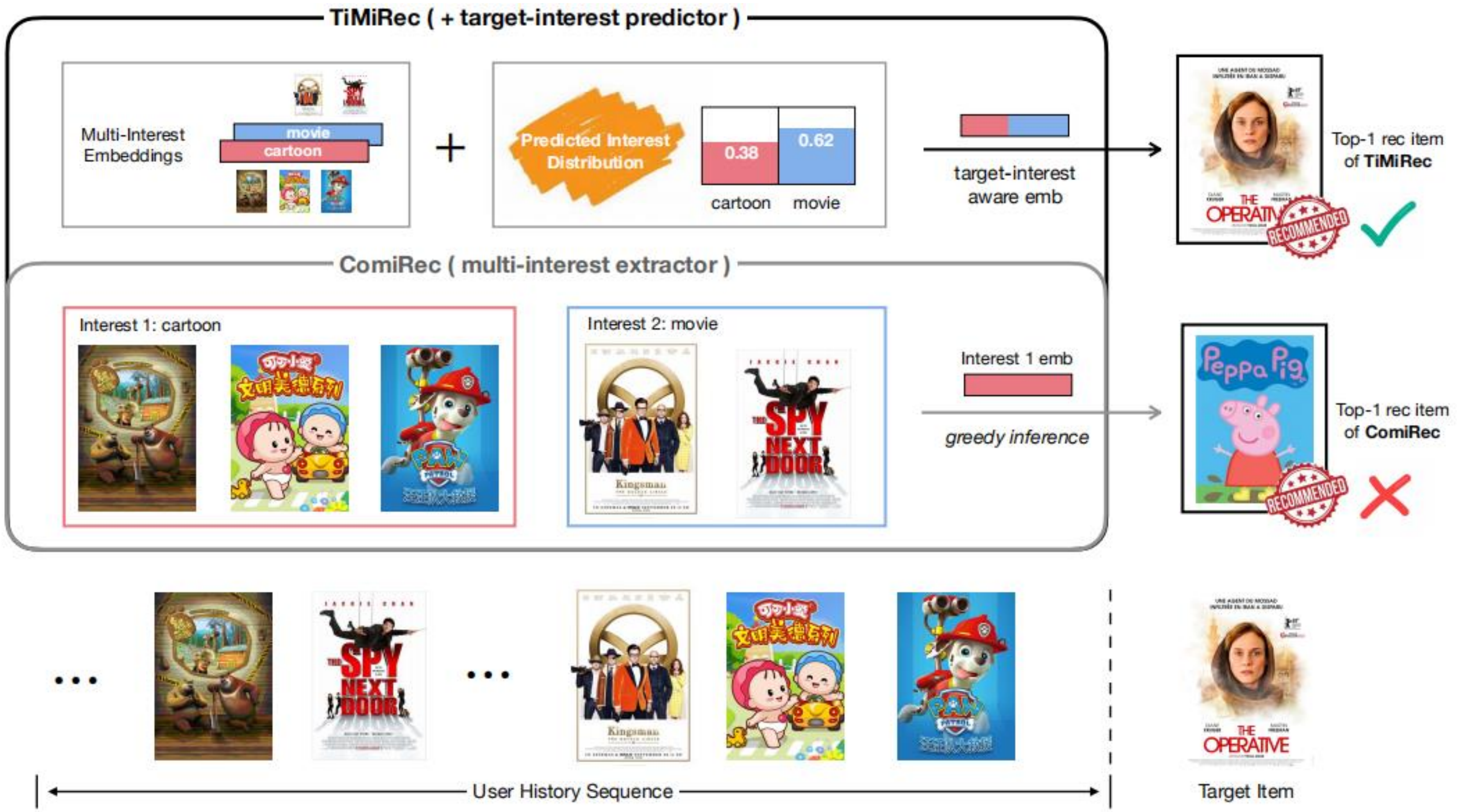


Figure 4: A case study on CMCC dataset. The multi-interest extractor generates two interests from the user history sequence: 1) cartoon and 2) movie. When making recommendations, ComiRec only considers the best matching interest for each candidate item, and hence wrongly recommends a cartoon because it gets the maximal matching score. However, after a series of cartoon watching, the interest for parents to watch movies may take advantage. TiMiRec can capture such dynamic intent with the target-interest predictor and gives the exactly correct recommendation.

Experiments

Algorithm 1 Learning algorithm of TiMiRec

Input: multi-interest extractor structure F_{Φ_E}, target-interest predictor structure G_{Φ_P}, interest number K

Output: parameters Φ_E, Φ_P

while not converged **do**
　$\mathbf{V}_{u,t} = F_{\Phi_E}(S_{u,t})$.
　$\mathbf{v}_{u,t} = \mathbf{V}_{u,t}[:, \operatorname{argmax}\left(\mathbf{V}_{u,t}^T \mathbf{i}_{u,t}\right)]$.
　Pretrain multi-interest extractor with $\mathcal{L}_{\text{rec}}$.
end while
while not converged **do**
　$\mathbf{V}_{u,t} = F_{\Phi_E}(S_{u,t})$.
　$\mathbf{z}_{u,t}^q = G_{\Phi_P}(S_{u,t})$.
　$\mathbf{z}_{u,t}^p = \operatorname{sim}(\mathbf{V}_{u,t}, \mathbf{i}_{u,t})$, i.e., Eq.(5).
　Calculate target-interest distillation loss $\mathcal{L}_{\text{distill}}$.
　$\mathbf{v}_{u,t} = \mathbf{V}_{u,t}\operatorname{softmax}(\mathbf{z}_{u,t}^q)$.
　Calculate next-item recommendation loss $\mathcal{L}_{\text{rec}}$.
　Finetune Φ_E and Φ_P with $\mathcal{L}$, i.e., Eq.(11).
end while
return Φ_E, Φ_P

Thank you!