CauseRec: Counterfactual User Sequence Synthesis for Sequential Recommendation

Shengyu Zhang*
Zhejiang University, China
sy_zhang@zju.edu.cn

Dong Yao*
Zhejiang University, China
yaodongai@zju.edu.cn

Zhou Zhao†
Zhejiang University, China
zhaozhou@zju.edu.cn

Tat-seng Chua
National University of Singapore
dcscts@nus.edu.sg

Fei Wu†
Zhejiang University, China
wufei@zju.edu.cn

Problem Formulation

- **Sequential Recommendation:** Given the historical behaviors $x_{u,t}$, where $x_{u,t} = \{y_{u,1:t}\}$ denotes a user's historical behaviors prior to the t-th behavior $y_{u,t}$ and arranged in a chronological order, the goal of sequential recommendation is to predict the next item $y_{u,t+1}$, which can be formulated as modeling the probability of all possible items:

$$p(y_{u,t+1)} = y|x_{u,t})$$

Challenges

- **Noise of implicit feedback**: Due to the ubiquitous distractions that may affect the users' first impressions, there are inconsistencies between users' interest and their clicking behaviors, known as the natural noise.

- **Data sparsity problem**: users in general only interact with a limited number of items compared with the item gallery which can easily reach 100 million in large live systems.

Methodology

- Modeling the counterfactual data distribution (besides the observation data distribution) to confront the data sparsity problem.

- Devising counterfactual transformations on both fine-grained item-level and abstract interest-level with various contrastive objectives to learn accurate and robust user representations.

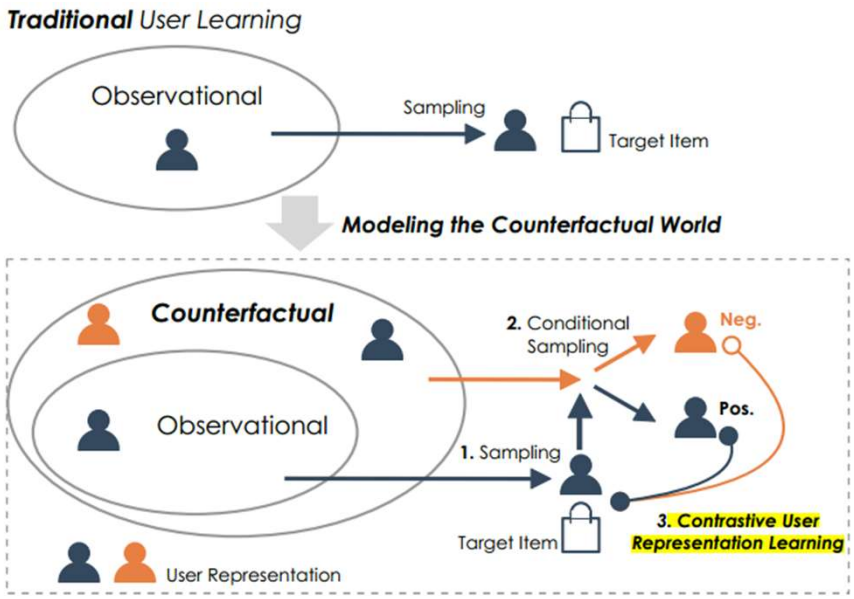

Figure 1: An illustration of the proposed contrastive user representation learning by modeling the counterfactual world (below), compared with most traditional approaches that solely model the observational user sequences (above).

Indispensable/Dispensable Concepts

- **Indispensable concept** indicates a subset of one behavior sequence that can jointly represent a meaningful aspect of the user's interest.

- **Dispensable concept** indicates a noisy subset that is less/meaningful/important in representing an aspect of interest.

- **Negative user representation:** Replacing indispensable concepts in the original user sequence.

- **Positive user representation:** Replacing dispensable concepts in the original user sequence.

CauserRec

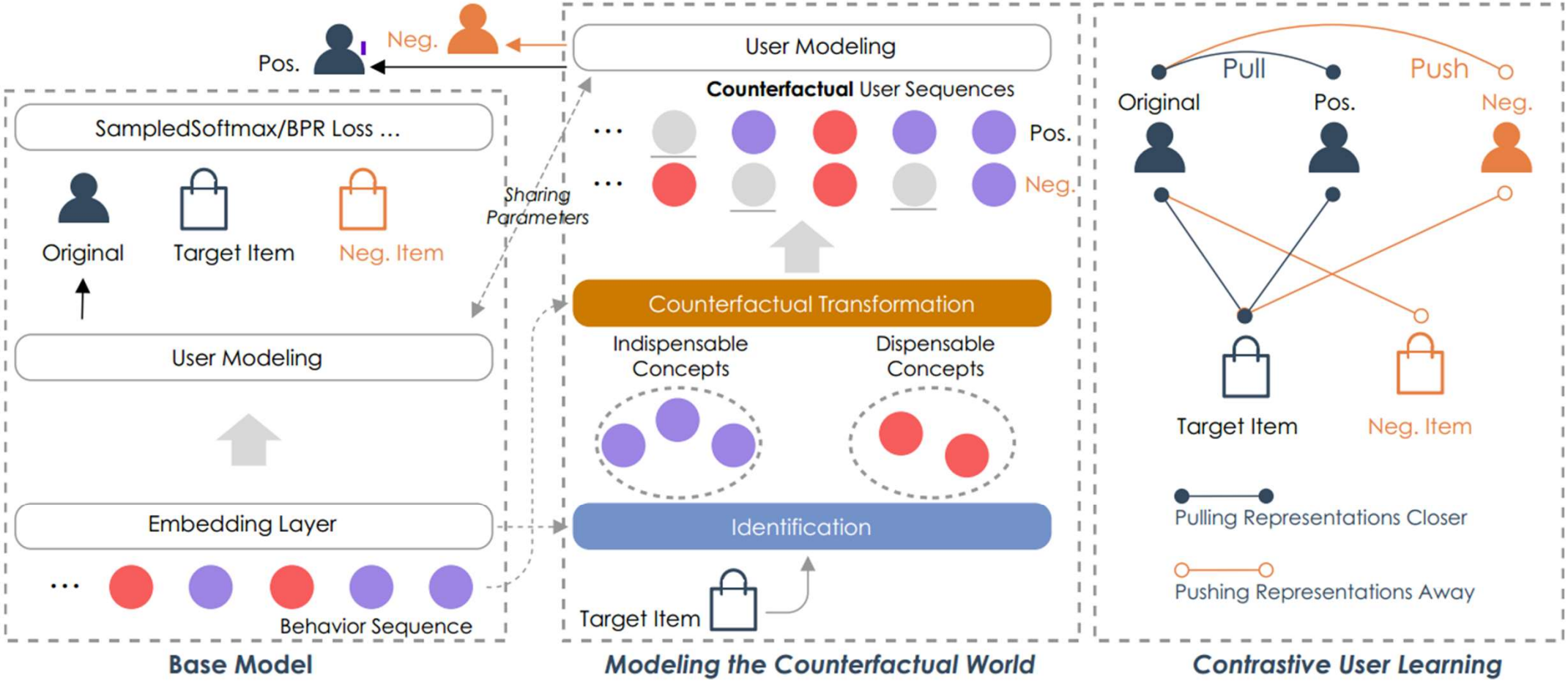

Figure 2: Schematic of the proposed CauseRec-Item framework.

Indispensable/Dispensable Concepts

- **Item-level concepts**

Denoting each item in the behavior sequence as an item-level concepts $\boldsymbol{C_{item}} = \mathbf{X} \in \mathbb{R}^{t \times d}$, where $\mathbf{X} = g_{\boldsymbol{\theta}}(\boldsymbol{x_{u,t+1}})$, and $g_{\boldsymbol{\theta}}$ is an item encoder. In this paper, $g_{\boldsymbol{\theta}}$ is a plain lookup embedding matrix where the n-th vector represents the item embedding with item id$n$. Denote $\boldsymbol{y}$ as the representation of the target item. Thus, the score for the i-th item-level concepts p_i^{item} can be calculated:

$$p_i^{item} = \phi_\theta(\boldsymbol{c_i}, \boldsymbol{y})$$

where $\phi_\theta(\cdot)$ is the similarity function. Here, dot product is adopted for the effectiveness in the experiment.

Indispensable/Dispensable Concepts

- **Interest-level concepts**

Due to some items may share similar semantics, and might deteriorate the capability of modeling higher-order relationships between items, interest-level concepts $\boldsymbol{C_{interest}}$ was proposed.

$$\boldsymbol{C_{interest}} = \boldsymbol{A}^T \boldsymbol{X}$$

$$\boldsymbol{A} = \text{softmax}(\boldsymbol{W}_2 \tanh(\boldsymbol{W}_1 \boldsymbol{X}^T))^T$$

$$p^{interest} = \boldsymbol{A}^T \phi_\theta(\boldsymbol{X}, \boldsymbol{y})$$

where $\boldsymbol{X}$ is the representations of the behavior sequence, $p^{interest}$ is the score of interest-level concepts. $\boldsymbol{W}_1 \in \mathbb{R}^{d_a \times d}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{K \times d_a}$ are trainable transformation matrices. K is thus the number of concepts that is pre-defined.

Counterfactual Transformation

- For both item-level concepts and interest-level concepts, the top half concepts with the highest scores can be denoted as **indispensable concepts** and the remaining half concepts as **dispensable concepts**.

- Replacing the identified **indispensable/dispensable concepts** at the rate of r_{rep} to construct counterfactually **negative/positive** user sequences, respectively.

- A first-in-first-out queue as a concept memory for each level and use dequeued concepts as substitutes.

User Encoders

- **CauseRec-Item/CauseRec-Interest**

CauseRec-Item/CauseRec-Inteset obtains counterfactually positive/negative user representations $\boldsymbol{x}^{+,m}/\boldsymbol{x}^{-,n}$ from counterfactual item-level concept sequences $\boldsymbol{C_{item}}/\boldsymbol{C_{interst}}$ using the original user encoder f_θ: $f_\theta = \mathrm{MLP}(\frac{1}{t}\sum_{i=1}^{t} g_\theta(y_i))$, y_i is the item embedding from $\boldsymbol{C_{item}}/\boldsymbol{C_{interst}}$.

- **CauseRec-H(ierarchical)**

CauseRecH further considers counterfactual transformations performed on item-level concepts. The counterfactually transformed item-level sequence will be forwarded to construct interest-level concept sequence.

Learning objectives

- **Contrast between Counterfactual and observation**

Using triplet margin loss to measure the relative similarity between samples:

$$\mathcal{L}_{co} = \sum_{m=1}^{M} \sum_{n=1}^{N} \max\{d(\boldsymbol{x}^q, \boldsymbol{x}^{+,m}) - d(\boldsymbol{x}^q, \boldsymbol{x}^{-,n}) + \Delta_{co}, 0\}$$

where $\boldsymbol{x}^q$ denotes the original user representation. The distance function $d(\cdot)$ is the L2 distance. The margin $\Delta_{co} = 1$.

- **Contrast between Interest and Items**

The target item y_t was further used to enhance the user representation learning.

$$\mathcal{L}_{ii} = \sum_{m=1}^{M} 1 - \tilde{\boldsymbol{x}}^{+,m} \cdot \tilde{\boldsymbol{y}} + \sum_{n=1}^{N} \max\{\tilde{\boldsymbol{x}}^{-,n} \cdot \tilde{\boldsymbol{y}} + \Delta_{ii}, 0\}$$

where $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}$ are the L2-normalized representation of $\boldsymbol{x}, \boldsymbol{y}$. The margin $\Delta_{ii} = 0.5$.

- **The loss function of the whole framework**

$$\mathcal{L}_{cause} = \mathcal{L}_{matching} + \lambda_1 \mathcal{L}_{co} + \lambda_2 \mathcal{L}_{ii}$$

$$\mathcal{L}_{matching} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} -log p_\theta(\boldsymbol{y}|\boldsymbol{x}), \qquad where\ \ p_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{exp\phi_\theta(\boldsymbol{x}, \boldsymbol{y})}{\sum_{y'\in\mathcal{Y}} \phi_\theta(\boldsymbol{x}, \boldsymbol{y}')}$$

Experiments

Table 2: Comparison results of three CauseRec architectures with SOTA sequential recommenders designed for the matching phase. CauseItem/CauseIn/CauseH stand for CauseRec -Item/-Interest/-Hierarchical, respectively. The symbol $*$ indicates the improvements over the strongest baseline (underlined) are statistically significant ($p < 0.05$) with one-sample t-tests.

Datasets	Metric	POP	Y-DNN	GRU4Rec	MIND	ComiSA	ComiDR	CauseItem	CauseIn	CauseH	Improv.
Books	R@20	0.0137	0.0457	0.0406	0.0486	0.0549	0.0531	0.0582	0.0593	**0.0623***	13.5%
	R@50	0.0240	0.0731	0.0650	0.0764	0.0847	0.0811	0.1001	0.0993	**0.1018***	20.2%
	NDCG@20	0.0226	0.0767	0.0680	0.0793	0.0899	0.0918	0.0985	0.1006	**0.1051***	14.5%
	NDCG@50	0.0394	0.1208	0.1037	0.1223	0.1356	0.1352	0.1628	0.1619	**0.1655***	22.1%
	HR@20	0.0302	0.1029	0.0894	0.1062	0.1140	0.1201	0.1280	0.1303	**0.1370***	14.1%
	HR@50	0.0523	0.1589	0.1370	0.1610	0.1720	0.1758	0.2078	0.2062	**0.2113***	20.2%
Yelp	R@20	0.0016	0.0506	0.0454	0.044	0.0534	0.0472	0.0570	0.0580	**0.0591***	10.7%
	R@50	0.003	0.1048	0.0937	0.0943	0.1101	0.0935	0.1163	0.1175	**0.1182***	7.36%
	NDCG@20	0.0065	0.1582	0.1447	0.1414	0.1728	0.1453	0.1812	0.1806	**0.1830***	5.90%
	NDCG@50	0.0129	0.2887	0.2673	0.2699	0.3025	0.2612	0.3179	0.3180	**0.3210***	6.12%
	HR@20	0.0152	0.3015	0.2826	0.2681	0.3249	0.2775	0.3416	0.3391	**0.3426***	5.45%
	HR@50	0.0268	0.5131	0.4853	0.4866	0.5324	0.4629	0.5583	0.5576	**0.5605***	5.28%
Gowalla	R@20	0.0028	0.1127	0.1273	0.1218	0.1277	0.1153	0.1315	0.1355	**0.1359***	6.42%
	R@50	0.0054	0.1926	0.2043	0.2049	0.2072	0.1831	0.2238	0.2204	**0.2251***	8.64%
	NDCG@20	0.0073	0.2378	0.2803	0.2565	0.2736	0.2534	0.2747	0.2825	**0.2842***	1.39%
	NDCG@50	0.0135	0.3638	0.4002	0.3888	0.4019	0.3621	0.4123	0.4113	**0.4221***	5.03%
	HR@20	0.0104	0.3443	0.3814	0.3627	0.3838	0.3429	0.3918	0.3995	**0.4042***	5.32%
	HR@50	0.0224	0.5010	0.5251	0.5301	0.5288	0.5355	0.5596	0.5553	**0.5697***	6.39%

Experiments

Table 3: Ablation studies by constructing different architectures. We progressively ablate key components in CauseRec-Item, which is a model-agnostic and non-intrusive design.

Model	Yelp						Gowalla					
	Metrics@20			Metrics@50			Metrics@20			Metrics@50		
	Recall	NDCG	Hit Rate	Recall	NDCG	Hit Rate	Recall	NDCG	Hit Rate	Recall	NDCG	Hit Rate
CauseRec-Item	0.0570	0.1812	0.3416	0.1163	0.3179	0.5583	0.1315	0.2747	0.3918	0.2238	0.4123	0.5596
w.o. $\mathcal{L}_{co}$	0.0563	0.1794	0.3350	0.1169	0.3174	0.5507	0.1286	0.2719	0.3865	0.2153	0.4055	0.5522
w.o. $\mathcal{L}_{ii}$	0.0518	0.1679	0.3113	0.1067	0.2983	0.5267	0.1256	0.2668	0.3758	0.2110	0.3968	0.5445
Pos Only	0.0518	0.1662	0.3101	0.1047	0.2911	0.5115	0.1256	0.2693	0.3851	0.2090	0.3945	0.5398
Neg Only	0.0494	0.1633	0.3044	0.1024	0.2909	0.5137	0.1246	0.2608	0.3754	0.2053	0.3925	0.5368
Base Model	0.0444	0.1444	0.2722	0.0934	0.2650	0.4692	0.1208	0.2569	0.3670	0.2000	0.3811	0.5218

Table 4: Performance analysis on the number of counterfactually positive/negative user representations in CauseRec-Item, denoted in M/N.

	Yelp			Gowalla		
Model	R@50	N@50	H@50	R@50	N@50	H@50
$N = 1, M = 1$	0.111	0.310	0.541	0.219	**0.413**	0.559
$N = 4, M = 4$	0.113	0.308	0.542	0.210	0.394	0.541
$N = 8, M = 8$	0.106	0.298	0.524	0.204	0.381	0.521
$N = 8, M = 1$	**0.116**	**0.318**	**0.558**	**0.224**	0.412	**0.560**
$N = 1, M = 8$	0.096	0.273	0.483	0.185	0.361	0.498

Table 5: Performance analysis on the replace rate r_{rep} in counterfactual transformation for CauseRec-Item.

	Yelp			Gowalla		
Model	R@50	N@50	H@50	R@50	N@50	H@50
$r_{rep} = 0.2$	0.115	**0.318**	0.552	0.209	0.389	0.531
$r_{rep} = 0.4$	**0.117**	**0.318**	0.556	0.209	0.401	0.543
$r_{rep} = 0.5$	0.116	**0.318**	**0.558**	**0.224**	**0.412**	**0.560**
$r_{rep} = 0.6$	0.113	0.307	0.537	0.210	0.397	0.537
$r_{rep} = 0.8$	0.106	0.296	0.520	0.211	0.399	0.544

Table 6: Analysis on the number of constructed interest concepts K for CauseRec-Interest.

	Yelp			Gowalla		
Model	R@50	N@50	H@50	R@50	N@50	H@50
$K = 4$	0.100	0.28	0.501	0.206	0.390	0.531
$K = 10$	0.111	0.305	0.536	0.215	0.403	0.546
$K = 20$	**0.118**	**0.318**	**0.558**	**0.220**	**0.411**	**0.555**
$K = 30$	0.117	0.311	0.547	0.219	0.406	0.547