Path-based Deep Network for Candidate Item Matching in Recommenders

Houyi Li[1], Zhihong Chen[1], Chenliang Li[3†], Rong Xiao[1], Hongbo Deng[1†], Peng Zhang[1], Yongchao Liu[2], Haihong Tang[1]
[1]Alibaba Group, Hangzhou, China, [2]Ant Group, Hangzhou, China
[3]School of Cyber Science and Engineering, Wuhan University, Wuhan, China
[1]{houyi.lhy, jhon.czh, xiaorong.xr, dhb167148, zhangpeng04, piaoxue}@alibaba-inc.com
[3]cllee@whu.edu.cn, [2]yongchao.ly@antgroup.com

ABSTRACT

The large-scale recommender system mainly consists of two stages: matching and ranking. The matching stage (also known as the retrieval step) identifies a small fraction of relevant items from billion-scale item corpus in low latency and computational cost. Item-to-item collaborative filtering (item-based CF) and embedding-based retrieval (EBR) have been long used in the industrial matching stage owing to its efficiency. However, item-based CF is hard to meet personalization, while EBR has difficulty in satisfying diversity. In this paper, we propose a novel matching architecture, Path-based Deep Network (named PDN), through incorporating both personalization and diversity to enhance matching performance. Specifically, PDN is comprised of two modules: *Trigger Net* and *Similarity Net*. PDN utilizes Trigger Net to capture the user's interest in each of his/her interacted item. Similarity Net is devised to evaluate the similarity between each interacted item and the target item based on these items' profile and CF information. The final relevance between the user and the target item is calculated by explicitly considering user's diverse interests, *i.e.,* aggregating the relevance weights of the related two-hop paths (one hop of a path corresponds to user-item interaction and the other to item-item relevance). Furthermore, we describe the architecture design of the proposed PDN in a leading real-world E-Commerce service (Mobile Taobao App). Based on offline evaluations and online A/B test, we show that PDN outperforms the existing solutions for the same task. The online results also demonstrate that PDN can retrieve more personalized and more diverse items to significantly improve user engagement. Currently, PDN system has been successfully deployed at Mobile Taobao App and handling major online traffic.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Deep Learning, Recommendation Systems

ACM Reference Format:
Houyi Li, Zhihong Chen, Chenliang Li, Rong Xiao, Hongbo Deng, Peng Zhang, Yongchao Liu, Haihong Tang. 2021. Path-based Deep Network for Candidate Item Matching in Recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462878

1 INTRODUCTION

Recommender systems are important in customer-oriented E- Commerce platforms (e.g., Taobao and Amazon). The purpose is to connect users to their preferred items and produce more profits. Due to the tremendous number of items available in the E-Commerce platforms, many industrial systems [6, 11] are devised with a matching stage and a ranking stage. Specifically, a matching stage is expected to retrieve a small fraction of relevant items in low latency and computational cost, and a ranking stage aims to refine the ranking of these relevant items in terms of the user's interest with more complex models. In this work, we focus on the matching stage since it is the fundamental part and also the bottleneck of the system.

Item-to-item based collaborative filtering, as known as item-based CF, is an information retrieval solution for item matching. It estimates the relevance between two items based on their co-occurrence patterns. Several outstanding merits make item-based CF a natural choice for the matching stage in industry [7, 24, 28]: (1) Since only the items previously interacted by the user are used for retrieval, an item-based CF could well match the efficiency requirement for many online services; (2) Furthermore, a user's interest could be very diverse. By taking each interacted item into consideration, the user's interest can be well covered to its maximum; (3) At last, the relevance between items is mainly derived based on massive user behaviors. These signals are effective to identify the relevant items w.r.t the interacted ones.

However, this kinds of methods also have some limitations. The traditional inverted indexes are hard to meet subtle personalization needs [33]. By considering only item co-occurrence patterns, item-based CF is inferior to accommodate the user's unique characteristics, *e.g.,* gender, consumption capacity and others. Similarly, given the volume of items available in a E-Commerce platform keeps growing, without considering auxiliary information, item-based CF could suffer a lot from the data sparsity problem [22].

† Chenliang Li and Hongbo Deng are the corresponding authors.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462878

To overcome the above limitations, embedding-based retrieval (EBR), especially the deep learning networks with a two-tower architecture, has drawn growing interests recently [17, 31, 32]. Briefly speaking, EBR aims to represent each user and item by embedding their profile respectively. In this sense, the matching process is transferred to perform nearest neighbor (NN) search in the embedding space. Although the representation learning could alleviate the data sparsity problem to some extent, these methods also have some limitations. The two-tower architecture is not easy to explicitly integrate the co-occurrence information between items. Moreover, a user is often represented as a single embedding vector, which is insufficient to encode the diversity of the user's interest [6, 18].

A normal user would interact with hundreds of items that belong to different categories every month, indicating the diversity of user interests. Actually, in real industrial systems, to simultaneously capture the diversity of user interests and ensure personalization, typically there are multiple strategies, e.g., various kinds of inverted indexes based on collaborative filtering models and EBR strategies with different network structures. These models are deployed in parallel for the matching task. Note that the candidates are usually generated with the relevance scores in different scales. It is not straightforward to fuse these incomparable values for promising performance. We argue that this multi-strategy solution may be suboptimal due to the high maintenance cost of the these strategies and lack of tailored joint optimization.

Deep Interest Network (DIN) [35] introduces the similarity between interacted and target items through the target attention for a better recommendation. However, this attention mechanism is simply used to fuse user interaction sequences, which ignores the user's interest in each interacted item. Also, DIN is difficult to be applied for the matching stage since it requires recalculating the user representation for each target item. Inspired by DIN, in this work, we propose a novel matching architecture called Path-based Deep Network (PDN), which decouples the target item based attention from user representation learning by building a marriage between item-based CF and EBR. In PDN, we use the thought of representation learning of EBR (*i.e.,* user profile, interacted item sequence and item profile) and item-based CF (*i.e.,* item co-occurrence) to accommodate both user personalization and diverse interest modeling for better performance. Specifically, PDN consists of two main subnetworks: Trigger Net and Similarity Net. Trigger Net (*TrigNet*) is introduced to encode the user's interest by considering each interacted item as a trigger[1]. That is, the generated user representation has a variable dimension such that each dimension describes the user's interest on the interacted item. Analogous to item-based CF, Similarity Net (*SimNet*) generates an item representation and each dimension describes the similarity between an interacted item and the target item. Note that, the dimensions of user representation and item representation extracted by EBR are constant, while the dimensions of the user and item representations extracted by PDN are variable which are equal to the number of the user's triggers. As shown in Figure 1, by connecting the user with the target item through her interacted items, we can form a series of 2-hop paths. Based on these 2-hop paths, PDN aggregate the relevance between the user and the target item by explicitly considering a user's diverse interests for better performance. Another merit is that the whole model is trained with an end-to-end fashion. Therefore, the relevance scores can be compared with each other in a uniform way.

It is worthwhile to highlight that the proposed PDN bears the advantages of both item-based CF and EBR for efficient online processing. On one hand, empowered by the feature embedding, we can utilize *TrigNet* to extract top-m most important triggers w.r.t. the user interests to satisfy the real-time requirement. On the other hand, *SimNet* works independently from *TrigNet*. We can apply parallel computing to support offline index construction with item-to-item relevance calculated by *SimNet* efficiently. In summary, the main contributions of this paper are as follows:

- **A novel matching model.** We propose Path-based deep network by incorporating advantages of item-based CF and EBR. PDN integrates all of profile information for user attention and co-occurrence patterns between items for target attention in the form of 2-hop path aggregation. Both the user personalization and the interest diversity are accommodated for better item matching.
- **Efficient online retrieval.** We construct an industrial-scale online matching system based on PDN. In particular, we describe how to leverage PDN for item retrieval in low latency and computational cost.
- **Offline and online experiments.** Extensive offline experiments on several real-world datasets demonstrates that the proposed PDN achieves much better performance than the existing alternatives. Besides, we evaluate PDN on the recommender system of Taobao with A/B test over a two-week period. The results suggest that a large performance gain is obtained on almost all metrics.

2 RELATED WORK

CF-based methods are successful in building recommender systems at the matching stage [29]. Among them, item-based CF [23, 26], which calculates the similarity matrix of items in advance, and recommend items similar to the user's clicked ones, has been widely employed in industrial settings due to its interpretability and efficiency. Early works utilize statistical measures such as cosine similarity and Pearson coefficient to estimate item similarities. In recent years, several approaches attempt to learn item similarities by optimizing a recommendation-aware objective function. Ning *et al.* [25] propose SLIM for item relevance learning by minimizing the loss between the original user-item interaction matrix and the reconstructed one. He *et al.* [14] propose NAIS with an attention mechanism to distinguish the different importance of historical items in a user profile, which shares a similar idea with DIN [35]. However, attention mechanism based methods are only applicable to the ranking stage due to the complexity of computation.

With the success of EBR [3], two-tower architectures based on deep neural networks have been widely adopted in industrial recommender systems to capture personalized information of a user by leveraging rich content features [17, 31, 32]. Note that, in the matching stage, to process billions or trillions of items in low latency and computational cost, the two towers can not interact with

[1]The terms *interacted item, trigger, trigger item* are exchangeable since they refer to the same meaning in this paper.

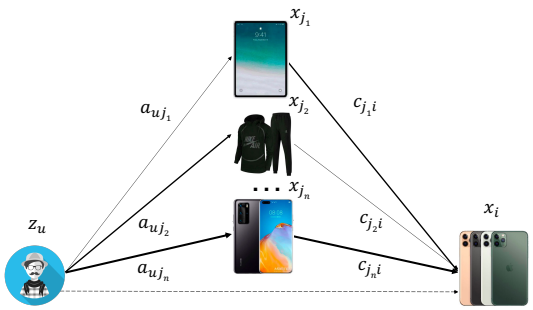


Figure 1: Modeling the preference of user u to target item i with a 2-hop graph where j_1 to j_n represent interacted items. The thickness of edge indicates the relevance.

each other to ensure the parallel feature extraction. Particularly, the DSSM-based model [18] learns the relevance based on the inner product between user features and item features. To extract more discriminative user features, Youtube DNN [6] extends user features by average pooling user behavior, while BST [4] utilizes the powerful Transformer model to capture the sequential signals underlying users' behavior sequences.

Different from the above methods, the PDN we propose combines the advantages of item-based CF and EBF based on deep neural networks, which constructs user-item subnetwork to ensure personalization similar to EMB, and constructs item-item subnetwork to capture multiple user interests similar to item-based CF.

3 PRELIMINARIES

Figure 1 summarizes the recommendation problem in the form of 2-hop paths between the user and the target item. Here, z_u represents the user information, *e.g.,* user id, age, gender, click times, purchase times of each category, for user u; $\boldsymbol{x}_i$ represents the information of target item, *e.g.,* item id, brand id, category id, monthly sales of target item i; $\{\boldsymbol{x}_{j_k}\}_{k=1}^n$ represent the auxiliary information of items $\{j_k\}_{k=1}^n$ interacted by u; $\{\boldsymbol{a}_{uj_k}\}_{k=1}^n$ represent the behavior information from user u on these interacted items, *e.g.,* stay time, purchase times; and $\{\boldsymbol{c}_{j_ki}\}_{k=1}^n$ represent the relevance information between interacted items and target item, which is obtained from an item-based CF algorithm or statistical correlation measure based on item co-occurrence patterns.

As shown in Figure 1, there is a direct link between the user and the target item (shown in dashed line), which indicates the user's intuitive interest on the target item. Also, we can further form n 2-hop paths by bridging through the interacted items. The first hop represents the user's interest in the interacted items and the second hop represents the relevance between the interacted items and the target item. Hence, the item matching with the above available information for recommendation can be formulated as:

$$\hat{y}_{ui} = f\big(z_u, \boldsymbol{x}_i, \{\boldsymbol{x}_j\}, \{\boldsymbol{a}_{uj}\}, \{\boldsymbol{c}_{ji}\}\big) \ \textit{with}\ j \in N(u) \tag{1}$$

where f is defined as the recommendation algorithm, $\hat{y}_{ui}$ is defined as the relevance score between user u and target item i, and $N(u)$ is defined as the items interacted by u.

Most of the existing work for recommender systems, including item-based CF and EBR, can be regarded as a special case of Eq. 1. For example, the regression form of item-based CF [27] can be formulated as:

$$\hat{y}_{ui} = f\big(\{\boldsymbol{a}_{uj}\}, \{\boldsymbol{c}_{ji}\}\big) = \sum_{j \in N(u)} f_r(\boldsymbol{a}_{uj}) c_{ji} \tag{2}$$

where $f_r : \mathcal{R}^m \rightarrow \mathcal{R}^1$ is a weighting function to capture user's interest for each trigger, $c_{ji} \in \mathcal{R}^1$ represents the relevance between interacted item j and target item i based on item co-occurrence information. Hence, the method can be seen as the sum of weights of all 2-hop paths based on $\{\boldsymbol{a}_{uj}\}$ and $\{\boldsymbol{c}_{ji}\}$, and each path weight can be calculated as $f_r(\boldsymbol{a}_{uj})c_{ji}$.

Besides, the methods based on EBR are also a special case of Eq. 1. For example, the matrix factorization (MF) [20] can be formulated as:

$$\hat{y}_{ui} = f\big(z_u, \boldsymbol{x}_i, \{\boldsymbol{x}_j\}\big) = \boldsymbol{q}_i\Big(\boldsymbol{p}_u + \frac{1}{\sqrt{|N(u)|}} \sum_{j \in N(u)} \boldsymbol{q}_j\Big)^T \tag{3}$$

where the $\boldsymbol{q}_i$, $\boldsymbol{p}_u$, $\boldsymbol{q}_j$ represent the embedding vector for target item information $\boldsymbol{x}_i$, user information z_u and interacted item $\{\boldsymbol{x}_j\}$, respectively. MF can be regarded as the sum of the weights of $n+1$ paths. Specifically, the weight of direct path is $\boldsymbol{q}_i\boldsymbol{p}_u$, and the weight of each 2-hop path is $1/\sqrt{|N(u)|} \cdot \boldsymbol{q}_i\boldsymbol{q}_j$. While YoutubeDNN [6] utilizes deep neural networks as a generalization of matrix factorization, which can be formulated as:

$$\hat{y}_{ui} = f\big(z_u, \boldsymbol{x}_i, \{\boldsymbol{x}_j\}\big) = \boldsymbol{q}_i\Big(MLP\big(\boldsymbol{p}_u, \frac{1}{|N(u)|} \sum_{j \in N(u)} \boldsymbol{q}_j\big)\Big)^T \tag{4}$$

Here *MLP* refers to the multilayer perception. DIN can also be formulated as [35]:

$$\begin{aligned}\hat{y}_{ui} &= f\big(z_u, \boldsymbol{x}_i, \{\boldsymbol{x}_j\}, \{\boldsymbol{a}_{uj}\}\big) \\ &= MLP\Big(\boldsymbol{p}_u, \boldsymbol{q}_i, \sum_{j \in N(u)} \big(MLP(\boldsymbol{q}_j, \boldsymbol{a}_{uj}, \boldsymbol{q}_i) \odot \boldsymbol{q}_j\big)\Big)\end{aligned} \tag{5}$$

where $\odot$ represents the element-wise product. Note that DIN [35] can be considered as a representation of a 2-hop path ($u \rightarrow j \rightarrow i$) in terms of relevance between the target item and each trigger. However, it requires re-calculation of path representation for each target item, making it only applicable for the ranking stage.

To ensure the efficiency of retrieval, item-based CF builds inverted index while EBR applies k-nearest neighbors (KNN) search for online serving. However, because both model architectures are constrained by efficiency, they cannot make use of all available information in Figure 1, resulting in suboptimal performance. For example, item-based CF lacks user and item profiles, while EBR lacks the explicit co-occurrence information between items. Hence, in this paper, we propose a novel architecture, named PDN, to support both personalized and diversity retrieval with low latency.

4 METHOD

In this section, we present the design of the Path-based Deep Network (PDN) for the matching stage of recommender systems. We first introduce the overall architecture of PDN, and then we elaborate on each module of PDN including Embedding Layer, Trigger Net (*TrigNet*), Similarity Net (*SimNet*), Direct Net, and Bias Net.

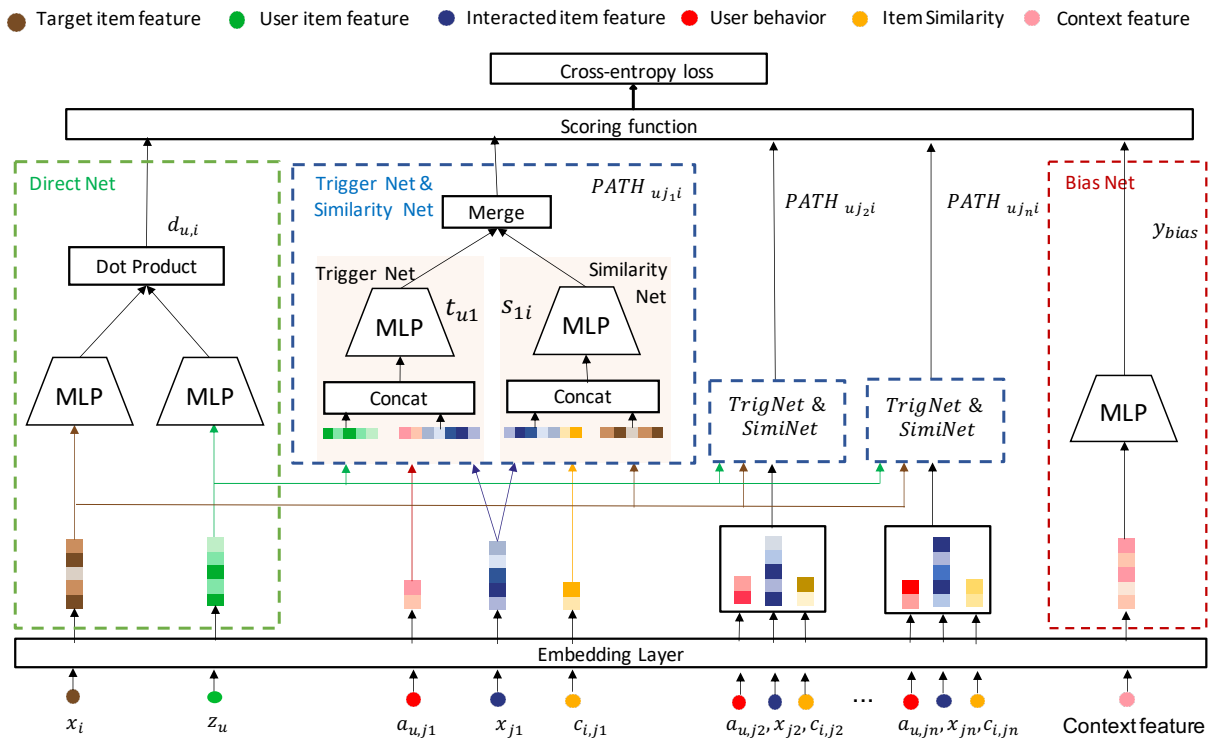


Figure 2: The network structure of PDN. Direct net is used to obtain the weight of the direct path to capture the user's intuitive interest in the target item, and *TrigNet* and *SimNet* obtain the first-hop weight and the second-hop weight of each 2-hop path respectively to capture the fine-grained user's personalized diversity interests, bias net is used to capture various types of selection bias for further unbiased serving. (Best viewed in color)

4.1 Overview of PDN

According to Eq 1, the basic workflow of PDN can be formulated as follows:

$$\hat{y}_{ui} = AGG\Big(f_d(\boldsymbol{z}_u, \boldsymbol{x}_i), \{PATH_{uji}\}\Big) \ with\ j \in N(u) \tag{6a}$$

$$PATH_{uji} = MEG\big(TrigNet(\boldsymbol{z}_u, \boldsymbol{a}_{uj}, \boldsymbol{x}_j), SimNet(\boldsymbol{x}_j, \boldsymbol{c}_{ji}, \boldsymbol{x}_i)\big) \tag{6b}$$

where f_d is a function to get the relevance weight of the direct path, $PATH_{uji}$ represents the relevance weight of the 2-hop path via trigger item j, AGG is a scoring function to obtain the final relevance score between the user and the target item by summing relevance weights of $n+1$ paths, MEG is a function to merge the relevance weights in each 2-hop path, and *TrigNet*, *SimNet* are two independent subnetworks as mentioned previously.

To enable PDN to perform fast online item retrieval, we first define MEG as the inner product or summation of the weights of the first hop and the second hop. Then f_d is defined also as the inner product between the user and the target item representations. Therefore, the final form of PDN is written as follows:

$$\hat{y}_{ui} = \boldsymbol{p}_u \boldsymbol{q}_i^T + \sum_{j \in N(u)} MEG\big(TrigNet(\boldsymbol{z}_u, \boldsymbol{a}_{uj}, \boldsymbol{x}_j), SimNet(\boldsymbol{x}_j, \boldsymbol{c}_{ji}, \boldsymbol{x}_i)\big) \tag{7}$$

The overall architecture of PDN is shown in Figure 2. Note that, in each 2-hop path, when the outputs of *TrigNet* and *SimNet* are vectors, they can be considered as representations of corresponding edges, which is called as vector-based PDN. This setting may have higher model capacity, and make online retrieval more complicated. On the other hand, when the outputs of *TrigNet* and *SimNet* are scalars, they can be regarded as the weights of corresponding edge, which is called as scalar-based PDN. Comparing with vector-based PDN, scalar-based PDN has lower degree of freedom, which can alleviate the complexity of online retrieval using greedy strategy based on path retrieval. Hence, we introduce each component of PDN with the setting of scalar-based PDN in the following.

4.2 Feature Composition & Embedding Layer

As shown in Figure 1, there are four feature fields in our recommender system: user field $\boldsymbol{z}_u$, user's behavior field $\{\boldsymbol{a}_{uj}\}$, item co-occurrence field $\{\boldsymbol{c}_{ji}\}$, item field $\boldsymbol{x}$ for both of interacted items $\{\boldsymbol{x}_j\}$ and target item $\boldsymbol{x}_i$. These fields include one-hot features, *e.g.*, user id, item id, age id, brand id, and continuous features, *e.g.*, monthly sales, stay time, the statistical correlation between items. At first, we transfer dense features into one-hot scheme through the discretization. Then, each one-hot feature is projected into a fixed-size dense representation. After embedding, we concatenate embedding vectors belonging to the same field as the representation of this field. Formally, field representations of user field, user's behavior field, item co-occurrence field and item field can be written as $\boldsymbol{E}(\boldsymbol{z}_u) \in \mathcal{R}^{1\times d_u}$, $\boldsymbol{E}(\boldsymbol{a}_{uj}) \in \mathcal{R}^{1\times d_a}$, $\boldsymbol{E}(\boldsymbol{c}_{ji}) \in \mathcal{R}^{1\times d_c}$, $\boldsymbol{E}(\boldsymbol{x}) \in \mathcal{R}^{1\times d_i}$, respectively, where d_u, d_a, d_c and d_i are the dimension size of corresponding field respectively.

4.3 Trigger Net & Similarity Net

After going through the embedding layer, we calculate the relevance weights for each 2-hop path between the user and the target item. For the first hop, we utilize *TrigNet* to capture the user's multi interests by calculating the preference for each of her triggers. Specifically, given user u and her trigger item j, the preference score is calculated as follows:

$$t_{uj} = TrigNet(\boldsymbol{z}_u, \boldsymbol{a}_{uj}, \boldsymbol{x}_j) = MLP\Big(CAT\big(\boldsymbol{E}(z_u), \boldsymbol{E}(\boldsymbol{a}_{uj}), \boldsymbol{E}(\boldsymbol{x}_j)\big)\Big) \tag{8}$$

where $CAT\big(\boldsymbol{E}(z_u), \boldsymbol{E}(\boldsymbol{a}_{uj}), \boldsymbol{E}(\boldsymbol{x}_j)\big) \in \mathcal{R}^{1\times(d_u+d_a+d_i)}$ is the concatenation of the user embedding, user behavior embedding and interacted item embedding, and t_{uj} represents the user's preference for an interacted item j. When there are n distinct interacted items, $\boldsymbol{T}_u = [t_{u1}, t_{u2}, ..., t_{un}]$ can be considered as a variable dimension representation for user u. EBR-based methods represent user interests by one fixed dimension representation vector, which can be a bottleneck for capturing diverse interests of users [22], because all information about diverse interests of one user is mixed together, causing inaccurate item retrieval for the matching stage. Compared with EBR-based solutions that encodes the user representation with a fixed dimension vector, $\boldsymbol{T}_u$ explicitly describes the user's preference for each interacted item, which can better represent the user's diverse interests and is more interpretable.

It is worth mentioning that the *TrigNet* can employ other more powerful neural networks, such as Recurrent Neural Network (RNN) and transformer-based models for user behavior [4, 30]. However, we would like to emphasize that a simple MLP is more cost-effective to our industrial system.

As to the second hop, we utilize *SimNet* to calculate the relevance between each interacted item and target item based on the item profile and co-occurrence information:

$$s_{ji} = SimNet(\boldsymbol{x}_j, \boldsymbol{c}_{ji}, \boldsymbol{x}_i) = MLP\Big(CAT\big(\boldsymbol{E}(\boldsymbol{x}_j), \boldsymbol{E}(\boldsymbol{c}_{ji}), \boldsymbol{E}(\boldsymbol{x}_i)\big)\Big) \tag{9}$$

where $CAT\big(\boldsymbol{E}(\boldsymbol{x}_j), \boldsymbol{E}(\boldsymbol{c}_{ji}), \boldsymbol{E}(\boldsymbol{x}_i)\big) \in \mathcal{R}^{1\times(2*d_i+d_c)}$ is the concatenation of interacted item embedding, co-occurrence embedding and target item embedding, and s_{ji} represents the relevance between item j and item i. $\boldsymbol{S}_i = [s_{1i}, s_{2i}, ..., s_{ni}]$ can be considered as a variable dimension representation of target item i. We emphasize that *SimNet* explicitly learns the relevance based on co-occurrence information and side information of items. In this sense, it can be deployed independently for item-to-item retrieval.

After obtaining $\{t_{uj}\}$ and $\{s_{ji}\}$, PDN merges them to get the relevance weight $PATH_{uji}$ of each two-hop path as follows:

$$PATH_{uji} = MEG(t_{uj}, s_{ji}) = ln(1 + e^{t_{uj}} e^{s_{ji}}) \tag{10}$$

4.4 Direct Net

We further model the user's general interests in a broader range with another set of user and item embeddings. For example, women are more interested in dresses, while men are more interested in belts. This can be consider as a 1-hop path directly connecting the user to the target item. Hence, we utilize a direct network composed of a user tower and an item tower. Specifically, these two towers go through MLP with Leaky Rectified Linear Units (LeakyReLU) based on user field (z_u) and target item field (x_i) to output a user representation $\boldsymbol{p}_u \in \mathcal{R}^{1\times k}$ and an item representation $\boldsymbol{q}_i \in \mathcal{R}^{1\times k}$, respectively. And then, the relevance weight of the direct path can be formulated as follows:

$$d_{u,i} = \boldsymbol{p}_u \boldsymbol{q}_i^T = MLP\big(\boldsymbol{E}(z_u)\big) MLP\big(\boldsymbol{E}(x_i)\big)^T \tag{11}$$

where $d_{u,i}$ is the direct relevance between user u and target item i.

4.5 Bias Net

Position bias, and many other types of selection biases, are studied and verified to be an important factor in recommender systems [1, 5, 34]. For example, it is common that users are inclined to click items displayed closer to the top of the list, even though it was not the most useful one of the entire corpus. To remove selection biases during model training, we train a shallow tower with features contributing to selection bias, such as position feature for position bias, and hour feature for temporal bias. The resultant bias logit y_{bias} is added to the final logit of the main model, as shown in Figure 2. Note that, at serving time, the bias net is removed to get an unbiased relevance score.

4.6 Loss Function

Whether user u would click target item i can be seen as a binary classification task. Therefore, we merge the relevance weights of $n+1$ paths and bias logit to get the final relevance score between u and i, and convert it into user click probability $p_{u,i}$:

$$\hat{y}_{u,i} = softplus(d_{u,i}) + \sum_{j=1}^{n} PATH_{uji} + softplus(y_{bias}) \tag{12}$$

$$p_{u,i} = 1 - exp(-\hat{y}_{u,i}) \tag{13}$$

Note that *softplus* function produces the relevance score $\hat{y}_{u,i}$ in the range of $(0, +\infty)$. Hence, we utilize Eq. 13 to convert it to a probability value between 0 and 1. To train the model, we apply the cross-entropy objective function as: $l_{u,i} = -\big(y_{u,i} log(p_{u,i}) + (1 - y_{u,i}) log(1 - p_{u,i})\big)$ where $y_{u,i}$ is ground-truth label indicating whether the user clicked on the item.

4.7 Discussion

To ensure that the training of PDN can converge to a better optimum, we carefully design the relevance weight for each path. As described above, we utilize $exp(\cdot)$ instead of other activation functions to constrain the output to be positive, i.e., $e^{s_{ji}}$ and $e^{t_{uj}}$, and further, through Eq. 10 to achieve the merge of the weights of each 2-hop path. This treatment of constraining the output to be positive is intuitive and fits real-world rationality. Note that the relevance weight would be negative in nature. However, this setting could allow PDN to search the local optimum in a much broader parameter space, which easily leads to overfitting. In Figure 3, We illustrate two bad examples by allowing the relevance weight to be negative. As shown in Figure 3a, when a negative target is connected by two totally irrelevant triggers, *SimNet* may generate a positive relevance weight for one path, and a negative one for the other path. After aggregation, the click probability is still quite low, *i.e.,* a perfect match with the ground truth. But it is clear that milk can not have a positive relevance with a mobile phone. Similarly, as shown in Figure 3b, *TrigNet* can also learn a negative preference towards a trigger, mainly to overfit the data.

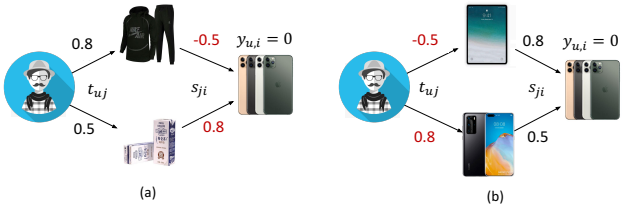


Figure 3: The bad cases in model training where (a) and (b) are negative sample. To optimize the loss of model, the similarly in (a) and the trigger weight in (b) have to be negtive.

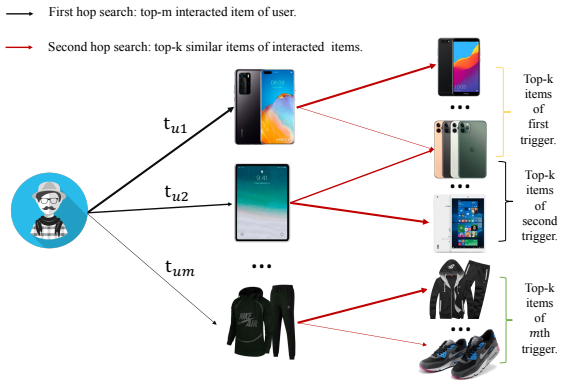


Figure 4: Online top-k path retrieval with greedy strategy.

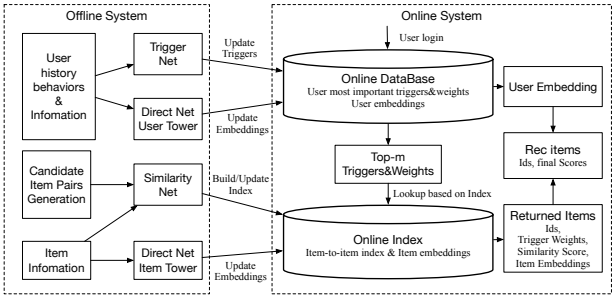


Figure 5: The framwork of online retrieval with PDN.

5 SYSTEM

In this section, we describe the implementation and deployment of PDN at Taobao for item recommendation in detail. As a user opens Mobile Taobao App, the recommender system firstly retrieves thousands of relevant items for this user from the corpus containing up to billions of items. Subsequently, every retrieved items are scored by a ranking model and the order list is displayed to the user as the recommendation.

5.1 Onling Retrieval

As aforementioned, billions or even trillions of items need to be processed in the matching stage. To satisfy the online real-time services, it is not possible to utilize PDN to score all available items. Based on architecture of PDN, it is intuitive that the larger path relevance weight, the more likely the user would be interested in the item. Therefore, the matching problem can be regarded as retrieving the item node with the larger path weights in the user's 2-hop neighborhood. As shown in Figure 4, we implement online real-time top-K nearest neighbor retrieval in the form of path retrieval via a greedy strategy. Specifically, we decompose the path retrieval into two parts: (a) top-m important trigger search (the first hop) using the *TrigNet*, (b) top-k item search for each top-m trigger (the second hop) based on the index generated by *SimNet*. For *TrigNet*, we deploy the model in a real-time online service for user trigger scoring. For *SimNet*, we use the model to compute and index item relevance in offline fashion. In detail, the online retrieval of PDN can be summarized as:

- **Index generation (Step 1):** Based on *SimNet*, the k most similar items of each item in the corpus are generated offline and stored with the relevance score s_{ji} in the database. The details can be seen in Section 5.2
- **Trigger extraction (Step 2):** When the user opens Mobile Taobao App, we take out all the items that the user has interacted with and utilize *TrigNet* to score all her triggers t_{uj} and return top-m triggers.
- **Top-K retrieval (Step 3):** We start with these top-m triggers to query the database and obtain $m \times k$ candidate items in total. Without considering the bias feature, the top-K items are returned for recommendation as follows.

$$\hat{s}_{u,i} = softplus(d_{u,i}) + \sum_{j=1}^{m} softolus(t_{uj} + s_{ji}) \tag{14}$$

Note that $\boldsymbol{p}_u$ and $\boldsymbol{q}_i$ are static representations. Hence, these two representations can also be inferred offline and stored in the database. When performing online service, they can be queried directly based on user or item id.

5.2 Index Generation

For industrial systems with large item corpus, we need to compress the dense matrix of item relevance, *i.e.*, $\mathcal{R}^{N\times N} \rightarrow \mathcal{R}^{N\times k}$, to reduce index construction time and storage cost. Even this process can be finished offline, it is too cost to calculate the relevance of all $N \times N$ item pairs, so we first reduce $N \times N$ to $N \times \hat{k}$ based on candidate generation, where $\hat{k}$ is an order of magnitude greater than k. The compression operation mainly consists of three steps:

- **Candidate item pair generation.** We mainly generate item pairs from two strategies, one is based on co-occurrence information, such as items that are clicked in the same session, the other is based on the profile information of items, such as items of the same brand. In this sense, the item pairs which have been not co-occurred before but have some similar property can also be considered as candidates.
- **Candidate item pair ranking.** We extract the relevance score s_{ji} of each item pair by using *SimNet*.
- **Index building.** For each item, we take the top-k similar items to form $N \times k$ item pairs, and store them with s_{ji} in the database.

It is worth mentioning that *SimNet* can be used independently for indexing, so it can be used also for item-to-item retrieval.

6 EXPERIEMENTS

In this section, we evaluate PDN against the existing state-of-the-art solutions with both offline and online settings, including ablation study, model analysis and case study.

Table 1: Statistics of experimental datasets.

Dataset	User#	Item#	Interaction#
MovieLens	6,040	3,706	1,000,209
Pinterest	55,187	9,916	1,500,809
Amazon Books	351,356	393,801	6,271,511

6.1 Offline Evaluation

6.1.1 Datasets and Evaluation Protocol. We experiment with three publicly real-world datasets: MovieLens [2], Pinterest [10], and Amazon books [12] for performance evaluation. The statistics of the three datasets are summarized in Table 1. Following the settings of [15], we filter these datasets in the same way that retained only users with at least 20 interactions. More details on this preprocessing for the three datasets have been elaborated in [15], so we do not restate here.

To evaluate the performance of PDN, we adopt the leave-one-out evaluation protocol [2, 16]. For each user, the last interaction is used as the target item, while the previous interaction items are collected as the user behaviors. Specifically, we followed the common strategy in [9, 19]. That is, all negative items are utilized in each test case, and the target items are ranked among these items. Hit Ratio (HR) [8] and Normalized Discounted Cumulative Gain (NDCG) [13] are adopted as the performance metrics. Here HR can be interpreted as a recall-based measure and NDCG is a ranking-based measure.

6.1.2 Comparing Methods. We compare PDN with the following two-tower methods and conventional item-based CF methods. The two-tower methods are introduced as follows:

- **DSSM [18].** DSSM employs embeddings to represent users and items. The relevance score is calculated based on the inner product between the user representation and the item representation.
- **Youtube DNN [6].** Youtube DNN utilizes the user's interaction sequence to derive user representations for item recommendation. It treats each item in the user's historical behaviors equally and adopts average pooling to extract a user's interest. To ensure fair comparison, hyperparameter tuning is conducted by a grid search, and each method is tested with the best hyperparameters.
- **BST [4].** BST extends the Youtube DNN, by utilizing the transformer layers to capture the user's short-term interest over the behavior sequence. We use the inner poduct of user and item representations instead of MLP.

The item-based CF methods are introduced as follows:

- **Pearson-based CF (PCF) [27].** This is the standard item-based CF, which estimates item relevance based on Pearson coefficient.
- **SLIM [25].** SLIM learns item relevance by minimizing the loss between the original user-item interaction matrix and the reconstructed one from the item-based CF model.

The ranking method is introduced as follows:

- **DIN [35].** Deep interest network takes target attention to extract the relationship between user's interacted sequence and target item.

6.1.3 Results and Discussion. Table 2 shows the experimental results on three public datasets in terms of HR@10 and NDCG@10. All experiments are repeated 5 times and the average results are reported. Clearly, PDN outperforms all comparison methods under most datasets. We can make the following observations. (1) For personalized retrieval, DSSM performs worst among two-tower methods, which indicates that capturing user interest based on the user behavior is critical for recommender systems. The performance of BST is better than that of YouTube DNN, due to the fact that the transformation layer extracts users' interests by considering the sequential information in their behaviors. PDN achieves the best performance, mainly because these two-tower methods represent each user by one fixed-dimension vector, *i.e.,* a bottleneck for modeling diverse interests. In contrast, PDN utilizes *TrigNet* to extract multi-interest user representation and each dimension describes the user's interest in an interacted item in a fine-grained way. *SimNet* is also effective in deriving the similarity between an interacted item and the target item (ref. Figure 1). Hence, a more accurate relevance can be estimated by considering the potential interest of the user towards the target item.

(2) When deploying online, we adopt SimNet instead of item-based CF to estimate item similarity for item-to-item retrieval. Therefore, we conducted a set of comparison with an item-to-item strategy (*i.e.,* the columns with "Item to Item" in Table 2). SimNet performs best among all methods. The reason is that SimNet explicitly optimizes the similarity between items by integrating the item profile used by the two-tower methods and co-occurrence information used by item-based CF based on deep neural networks, which utilizes more information to solve sparsity problem encountered by CF. (3) PDN performs better than DIN. We believe that DIN ignores users' attention to interacted items, while PDN considers both user attention and item attention for better personalized recommendation. (4) Based on ablation study, PDN yields the best performance, confirming that each component contributes to the final results.

6.2 Online Experiments

6.2.1 A/B Tests. Beyond offline studies, we conduct online A/B experiments by deploying our method in the recommender system of Taobao for two weeks. In the control setup (*i.e.,* Baseline), it includes all matching strategies in our current production system. In the variation experiment setup, we applies *SimNet* instead of the item-based method for inverted-index building. For fair comparison, the same ranking component and business logic are applied on the top of both matching stages. Table 3 and Figure 6 summarize the experimental results. We can find that the proposed PDN improves the e-commerce recommender system for all core business metrics, including page click-through rate (PCTR), user click-through rate (UCTR), clicks per user (ClkPU), average session duration (avgSD),

[2] http://grouplens.org/datasets/movielens/1m/

Table 2: HR and NDCG of different methods on the three datasets. The best results are highlighted in boldface. The improvements over the comparing methods are statistically significance at 0.05 level. 'Personalise' means that the output of the score considers the user information, e.g., the output of PDN and EBR-based method. It is worth noting that CF can't get personalized scores, which is represented by '-'. 'Item to Item' refers to retrieval based on the relevance between interacted and target items. Specifically, the acquisition strategy of item similarity is the output of item-based CF, the output of *SimNet* or the inner product between item features extracted from EBR methods.

Group	Method	MovieLens				Pinterest				Amazon Books			
		Personalise		Item to Item		Personalise		Item to Item		Personalise		Item to Item	
		HR@10	NDCG@10	HR@10	NDCG@10	HR@10	NDCG@10	HR@10	NDCG@10	HR@10	NDCG@10	HR@10	NDCG@10
Two-tower	DSSM [18]	0.4699	0.2603	0.2243	0.1034	0.2730	0.1394	0.1647	0.0759	0.6433	0.4163	0.1981	0.0894
	Youtube DNN [6]	0.6187	0.3579	0.4362	0.2324	0.7585	0.4351	0.6409	0.3028	0.6539	0.4370	0.2398	0.0955
	BST[4]	0.6316	0.3603	0.0561	0.0277	0.8176	0.4955	0.5371	0.2428	0.6923	0.4403	0.1019	0.0574
Item-based CF	PCF [27]	-	-	0.4033	0.2132	-	-	0.2800	0.1470	-	-	0.0547	0.0203
	SLIM [25]	-	-	0.4400	0.2337	-	-	0.6067	0.3323	-	-	0.1621	0.0679
ranking model	DIN [35]	0.6454	0.3767	0.1662	0.0918	0.8185	0.5051	0.4370	0.2059	0.6973	0.4698	0.1325	0.0612
Incorporated	PDN w/o bias net	0.6323	0.3401	0.5080	0.2795	0.8015	0.4814	0.7654	0.4345	0.6869	0.4520	0.2866	0.1546
	PDN w/o direct net	0.6642	0.3930	0.5124	0.2813	0.8123	0.5286	0.7703	0.4397	0.7012	0.4729	0.2962	0.1613
	PDN	**0.6770**	**0.4071**	**0.5152**	**0.2859**	**0.8283**	**0.5358**	**0.7911**	**0.4613**	**0.7019**	**0.4735**	**0.3505**	**0.2049**

as well as diversity[3], which are considered to be good indicators of recommendation satisfaction. Especially, personalized diversity is normally hard for existing production systems, while PDN increases the diversity of recommended items by a large margin, which indicates PDN can capture the diverse interests of users with *TrigNet* and *SimNet* to improve the overall user experience. Besides, we deploy *SimNet* to verify it can be used independently to build index instead of item-based CF for item-to-item retrieval, which also gets a perform gains. Note that, the metrics online are reported with relative improvement, *i.e.*, $(metric_{\text{PDN}} - metric_{base})/metric_{base}$.

Table 3: Online A/B test improvements in Taobao (Observed over two week).

Method	PCTR	ClkPU	AvgSD	Diversity
Baseline	+0.0%	+0.0%	+0.0%	+0.0%
SimNet	+9.25%	+5.43%	+4.08%	+14.68%
PDN	**+18.04%**	**+15.38%**	**+7.87%**	**+19.60%**

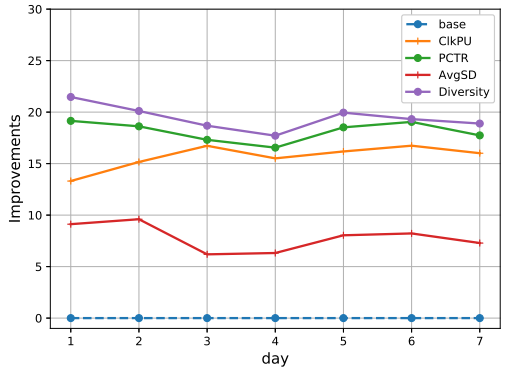


Figure 6: PDN online improvements in one week.

[3]Diversity represents the proportion of category coverage.

6.2.2 Online Efficiency. Here, we report the efficiency of online serving. Specifically, the overall latency from requests to candidate generation can be done within 6.75 milliseconds, *i.e.,* the queries per second (QPS) is in the thousand level, This is comparable to the retrieval with the standard inverted index. Based on the huge performance gain and low latency, we have deployed PDN online to serve the matching stage of recommendation in Taobao.

6.3 Case Study

Table 4: Impact of user behavior sequence length (n) on HR@300 at Taobao offline logs. Percentages in the brackets indicate the relative improvements over BST.

Method	n<=15	15<n<=30	30<n<=45	n>45	all
PCF [27]	0.066 (-46.3%)	0.110 (-59.1%)	0.131 (-31.4%)	0.141 (-41.8%)	0.120
BST [4]	0.123	0.175	0.191	0.200	0.180
PDN	**0.263 (+113.8%)**	**0.326 (+86.3%)**	**0.329 (+72.3%)**	**0.295 (+47.5%)**	0.297

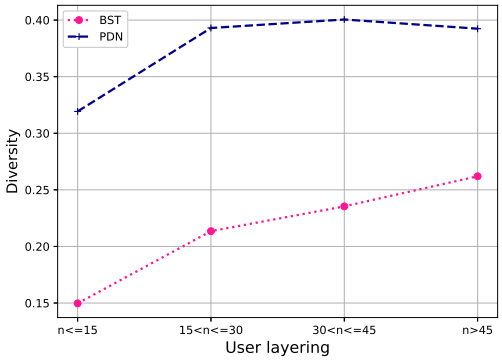
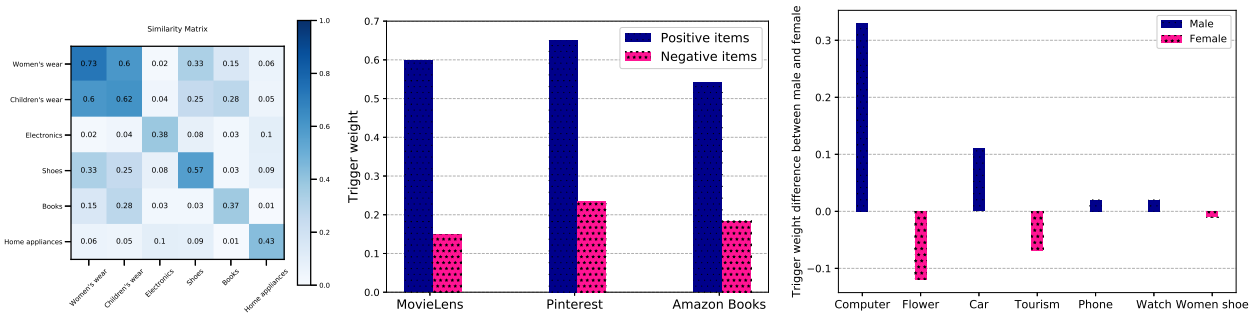


Figure 7: Diversity w.r.t. user behavior sequence length.

6.3.1 Analysis of User Behavior Sequence Length. Based on the model trained with Taobao offline log, we investigate the impact of user behavior sequence length n on performance. As shown in Table 4, we group users based on n, and found that the smaller the n is, the larger the gain is. This result shows that PDN has stronger



(a) (b) (c)

Figure 8: (a) Item-to-item similarity matrix based on *SimNet* at online serving; (b) User-to-item interests based on *TrigNet* on public datasets; (c) User group-to-category interests based on *TrigNet* at online serving.

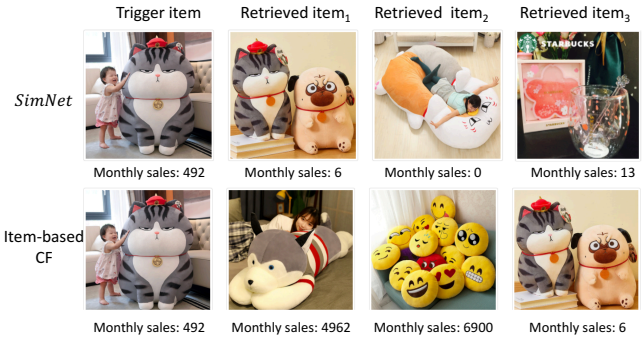


Figure 9: Comparison of item-to-item retrieval results between *SimNet* and item-based CF based on trigger item. The similarity decreases from left to right.

robustness to n, and it can obtain superior performance even in the case of sparse user behavior sequence. As shown in Figure 7, with the increase of n, the diversity of candidates generated by PDN gradually increases, and is better than that of BST by a large margin. These results show that PDN can capture the diverse interests of users in a fine-grained way to improve the user experience.

6.3.2 Diversity and Accuracy of SimNet. To verify the effectiveness of *SimNet*, we utilize the same trigger item to perform item-to-item retrieval using the similarity provided by *SimNet* and item-based CF, respectively. As shown in Figure 9, retrieval through cat-shaped dolls, *SimNet* returns similar cat-shaped dolls and cups printed with cat paw patterns, while item-based CF returns dolls with high monthly sales but not related to cats, *e.g.,* the dog-shaped dolls. This result indicates that *SimNet* can retrieve more relevant and diverse results without being disturbed by the number of interactions, while item-based CF would suffer from "Matthew Effect" [21], which introduces perference bias towards popular items. In other words, item-based CF only calculates similarity based on item co-occurrence information, while our method additionally introduces more information, *e.g.,* item profile, user profile, user behavior, for more accurate item similarity estimation. To further verify the reliability of item similarity based on *SimNet*, we randomly select 1,000 item pairs based on each category pair and calculate their average similarity from the inverted index built by *SimNet*. As shown in Figure 8a, we find that items belonging to the same category have a higher similarity. Besides, the similarity of related categories is higher than that of unrelated categories. For example, both the pair of women's wear and children's wear, and the pair of electronics and home appliances, have higher similarity than other combinations.

6.3.3 Personalization and Effectiveness of TrigNet. On the public datasets, we exploit the user's interest in items based on *TrigNet*. As shown in Figure 8b, the user-to-item trigger weights are averaged on positive and negative items respectively. It is obvious that the weights on positive items are higher than that on negative items. This result indicates the reasonable of *TrigNet*. To verify the personalization ability of *TrigNet*, we randomly selected 1,000 males and 1,000 females, and randomly selected 500 items from several categories, and then apply *TrigNet* to score all user-item pairs and get the average trigger weight difference between male and female. As shown in Figure 8c, we can find that *TrigNet* treats gender differently to reflect the user characteristics precisely. For example, male prefer computer, car, and watch, while female prefer flower, tourism and women shoes.

7 CONCLUSION

In this paper, a novel model named Path-based Deep Network is proposed for candidate item generation. The proposed PDN establishes 2-hop paths from the user to the target item, leading to better personalized yet diverse item retrieval. To the best of our knowledge, this is the first work to build a retrieval architecture based on a 2-hop graph. Besides, we present how to apply this architecture to achieve online retrieval in low latency based on the proposed path retrieval and a greedy strategy. The offline experiments over three real-world datasets are conducted to demonstrate that PDN outperforms existing SOTA alternatives. Moreover, our online experiments further indicates that PDN can improve retrieval quality in terms of five industry metrics. Lastly, PDN has been successfully deployed in our online recommender system in Taobao.

ACKNOWLEDGMENTS

Chenliang Li's work was supported by National Natural Science Foundation of China (No. 61872278).

REFERENCES

[1] Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, Marc Najork, and Thorsten Joachims. 2019. Estimating position bias without intrusive interventions. In *WSDM*.

[2] Immanuel Bayer, Xiangnan He, Bhargav Kanagal, and Steffen Rendle. 2017. A generic coordinate descent framework for learning from implicit feedback. In *WWW*.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[4] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. Behavior sequence transformer for e-commerce recommendation in alibaba. In *DLP-KDD*.

[5] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. 2020. Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance. In *SIGIR*.

[6] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *RecSys*. Association for Computing Machinery.

[7] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *RecSys*.

[8] Mukund Deshpande and George Karypis. 2004. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 143–177.

[9] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *WWW*.

[10] Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social networks. In *ICCV*.

[11] Carlos A Gomez-Uribe and Neil Hunt. 2015. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6, 4 (2015), 1–19.

[12] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*.

[13] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *CIKM*.

[14] X. He, Z. He, J. Song, Z. Liu, Y. Jiang, and T. Chua. 2018. NAIS: Neural Attentive Item Similarity Model for Recommendation. *IEEE Transactions on Knowledge and Data Engineering* 30, 12 (Dec 2018), 2354–2366.

[15] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*.

[16] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *SIGIR*.

[17] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *SIGKDD*.

[18] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. CIKM.

[19] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *SIGKDD*.

[20] Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. https://doi.org/10.1109/MC.2009.263

[21] Adit Krishnan, Ashish Sharma, Aravind Sankar, and Hari Sundaram. 2018. An adversarial approach to improve long-tail performance in neural collaborative filtering. In *CIKM*.

[22] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *CIKM*.

[23] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, 1 (2003), 76–80.

[24] David C Liu, Stephanie Rogers, Raymond Shiau, Dmitry Kislyuk, Kevin C Ma, Zhigang Zhong, Jenny Liu, and Yushi Jing. 2017. Related pins at pinterest: The evolution of a real-world recommender system. In *WWW*.

[25] Xia Ning and George Karypis. 2011. Slim: Sparse linear methods for top-n recommender systems. In *ICDM*.

[26] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*.

[27] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. In *WWW*. Association for Computing Machinery.

[28] Brent Smith and Greg Linden. 2017. Two decades of recommender systems at Amazon. com. *Ieee internet computing* 21, 3 (2017), 12–18.

[29] Xiaoyuan Su and Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009 (2009).

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[31] Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, and Ed H Chi. 2020. Mixed Negative Sampling for Learning Two-tower Neural Networks in Recommendations. In *WWW*.

[32] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *RecSys*.

[33] Han Zhang, Songlin Wang, Kang Zhang, Zhiling Tang, Yunjiang Jiang, Yun Xiao, Weipeng Yan, and Wen-Yun Yang. 2020. Towards Personalized and Semantic Retrieval: An End-to-End Solution for E-commerce Search via Embedding Learning. In *SIGIR*.

[34] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending what video to watch next: a multitask ranking system. In *RecSys*.

[35] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep Interest Network for Click-Through Rate Prediction. In *SIGKDD*. Association for Computing Machinery, New York, NY, USA.