A Feature-Based Coalition Game Framework with Privileged Knowledge Transfer for User-tag Profile Modeling

Xianghui Zhu
954305012@qq.com
Shanghai Jiao Tong University, China

Peng Du
hongyu.dp@alibaba-inc.com
Alibaba Group, China

Shuo Shao
shuoshao@sjtu.edu.cn
Shanghai Jiao Tong University, China

Chenxu Zhu
zhuchenxv@sjtu.edu.cn
Shanghai Jiao Tong University, China

Weinan Zhang
wnzhang@sjtu.edu.cn
Shanghai Jiao Tong University, China

Yang Wang
ywang@sei.ecnu.edu.cn
East China Normal University, China

Yang Cao
yinming.cy@alibaba-inc.com
Alibaba Group, China

ABSTRACT

User-tag profiling is an effective way of mining user attributes in modern recommender systems. However, prior researches fail to extract users' precise preferences for tags in the items due to their incomplete feature-input patterns. To convert user-item interactions to user-tag preferences, we propose a novel feature-based framework named Coalition Tag Multi-View Mapping (CTMVM), which identifies and investigates two special features, ***Coalition Feature*** and ***Privileged Feature***. The former indicates decisive tags in each click where relationships between tags in one item are treated as a coalition game. The latter represents highly informative features that only occur during training. For the coalition feature, we adopt Shapley Value based Empowerment (SVE) to model the tags in items with a game-theoretic paradigm and charge the network to straight master user preferences for essential tags. For the privileged feature, we present Privileged Knowledge Mapping (PKM) to explicitly distill privileged feature knowledge for each tag into one single embedding, which assists the model in predicting user-tag preferences at a more fine-grained level. However, the barren capacity of single embeddings limits the diverse relations between each tag and different privileged features. Therefore, we further propose Adaptive Multi-View Mapping (AMVM) model to enhance effect by handling multiple mapping networks. Excellent offline experiment results on two public and one private datasets show the out-standing performance of CTMVM. After the deployment on Alibaba large-scale recommendation systems, CTMVM achieved improvement by 10.81% and 6.74% in terms of Theme-CTR and Item-CTR respectively, which validates the effectiveness of taking in the two particular features for training.

*Yang Wang is the corresponding author.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599761

CCS CONCEPTS

• **Information systems** → **Personalization**.

KEYWORDS

Recommender System; User-tag Profiling; Personalization; Knowledge Transfer

ACM Reference Format:
Xianghui Zhu, Peng Du, Shuo Shao, Chenxu Zhu, Weinan Zhang, Yang Wang, and Yang Cao. 2023. A Feature-Based Coalition Game Framework with Privileged Knowledge Transfer for User-tag Profile Modeling. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3580305.3599761

1 INTRODUCTION

Personalized services are rapidly gaining popularity for their advantages in reducing information overload and improving user satisfaction [1]. In this field, user-tag profiling plays a crucial role, which outlines user characteristics from item-related information (e.g. fashion tags) [2]. When a user browses an item, the tags bound in the item have a substantial impact on the decision process. To methodically explore the relationships between users and tags, deep learning models are becoming one of the promising techniques.

Due to the indispensable function of user-tag profiling in the recall and ranking stages of recommender systems [3], a large number of industrial applications have sprung up in this field. Suitable recommendation algorithms aim to maximize the commercial value of companies and meet people's interest. In large e-commerce enterprises, advertisers usually make substantial profits through putting advertisements to users who are more anticipated to click the item with target tags [4]. Moreover, the user preference scores for each tag can be introduced as a portrait feature to alleviate the data sparsity issue in cross-domain cold start recommendations. Another business case is the display of recommendation reasons. Prevalent and interpretable recommendation reasons such as "Users who like Lolita Style also like it." are persuasive and generally provide some promotion in the hit ratio.

In recent years, various methods have been proposed to model user-tag profiles. For example, Yan et al. [5] apply a sophisticated multi-head attention mechanism to construct user-tag profiles and

Xu et al. [6] introduce Mixture of Virtual-Kernel Experts to study multi-objective user profile modeling. With the discovery of the data gap between training and testing stages, CWTM [7] is presented to improve the learning procedure in this task. However, two research issues arising with input features still remain:

- $\mathcal{I}1$: For the user-tag profiling task where one item includes several tags, it is rational that not all the tags in the item are equally contributed. However, simply aggregating the tags through a sum-pooling layer [5] assigns equal credits for them, and weightedly masking item tags by an importance network [7] treats each tag as an individual without taking into account the coalition associations between them. Current works lack a theoretic strategy to attribute each tag and instruct the neural network to directly learn users' genuine preferences for each tag by their contributions.
- $\mathcal{I}2$: In the previous methods, only user features and tag features are fed to the model during training since they are the only ones that will be input when evaluation. Nevertheless, ignored item-side features (e.g. store-id, category) may also play a significant role in this task.

In this task, tags in one item work jointly to induce users to click, and their coorperating payment can be referred to the amount of user clicks generated by this tag coalition. Intuitively, interrelationships between tags in one item are comparable to a coalition game in which all tags play the game in an alliance manner [8]. To address $\mathcal{I}1$, we regard item tags as ***Coalition Feature*** and explore the contribution of each tag player by Shapley Value [9]. Subsequently, we adopt **S**hapley **V**alue based **E**mpowerment (**SVE**) to empower tags via their shapley values. During training, each user-item instance are split into several user-tag interaction samples for input and the samples in clicking actions are assigned different gradient descent weights associated with the contributions of tags, which ensures the model straight grasps the precise tag preferences of users.

As for $\mathcal{I}2$, users may click the item for some item-side features except tags. With the help of the knowledge from these features, models may make better predictions. Chen et al. [10] name the features available when training but missing in the evaluation stage as ***Privileged Feature*** for the CTR prediction task and propose the distillation approach to model them. However, without explicitly modeling of the relationships between tags and priviliged features, their work suffers from considerable information loss. To this end, we propose a simple but effective architecture, **P**rivileged **K**nowledge **M**apping (**PKM**), that leverages privileged knowledge of each tag to aid in prediction. This architecture contains two networks and the main network is set for training and inference while the supplementary network is trained simultaneously to get qualified privileged embeddings. Between the two networks, we set the **K**nowledge **M**apping **N**etwork (**PKM**) that receives one tag embedding as input and encodes a **substitute embedding** to replace the privileged features during testing.

Despite the introduction of PKM incorporates privileged knowledge into training, another problem emerges. Constrained by the expression capacity of embedding vectors [11], a single substitute embedding is hard to adequately hold all privileged features that appear in conjunction with the tag, especially in some industrial environments where hundreds of times more privileged features exist than tags. Inspired by multi-interest models [12, 13], we improve PKM into **A**daptive **M**ulti-**V**iew **M**apping (**AMVM**) model that contains multiple mapping networks to adaptively project one tag embedding to multi-view privileged embeddings.

Along this line, we present a universal feature-based framework named **Coalition Tag Multi-View Mapping (CTMVM)**, which identifies two essential features neglected in previous user-tag profiling models and adopts SVE and AMVM modules to cope with them. For each user-item interaction, we forward user context features and every single tag in the current item to the model respectively as the training samples and modify the gradient values with pre-calculated shapley values when users click. In AMVM, we synchronously train a supplementary network with overall features as input and set up several knowledge mapping networks to adaptively encode multiple substitute embeddings restrained by the privileged embeddings via the MSE loss. Then the knowledge aggregating module is introduced to combine these substitute embeddings encoded by several KMNs. Finally, we let our network optimize directly towards the label with a task-oriented auxiliary loss, which mitigates the uncertainty in the substitute embedding.

In this work, we conduct comprehensive experiments in two public real-world datasets and a specific industrial dataset. Besides we have launched CTMVM in Alibaba recommendation systems with careful A/B tests, which achieves observed improvement of Theme-CTR by 10.81% and Item-CTR by 6.74%. The experimental results indicate that the weighted gradients on coalition features and the introduction of privileged features can significantly increase the model performance.

The main contributions of this paper are summarized in three folds:

- We put forward a generic framework CTMVM that for the first time discovers and examines two critical features in user-tag profile modeling: Coalition Feature and Privileged Feature.
- We introduce Shapley Value based Empowerment to determine the credit of each tag in user actions and allow the network to learn about user preferred tags in a targeted way.
- We propose the novel Privileged Knowledge Mapping to effectively utilize the privileged information and furthermore refine it to Adaptive Multi-View Mapping which settles the expressiveness constraints of single substitute embeddings.

2 RELATED WORKS

This section outlines three important areas related to our work: tag recommendations, shapley value for attribution, and knowledge transfer in recommender systems.

2.1 Tag Recommendation System

Integrating tag information into recommender systems has been a long research focus. For instance, Jächke et al. [14] study user-based collaborative filtering and graph-based recommender built on FolkRank. Collaborative filtering algorithm [15] is also cast into tag recommendation issue then. An integrated recommendation method by Sina Weibo [16] that proposes semantic correlations for tag recommendation. In fact, traditional algorithms above usually fail to implement in-depth modeling due to their simple design.

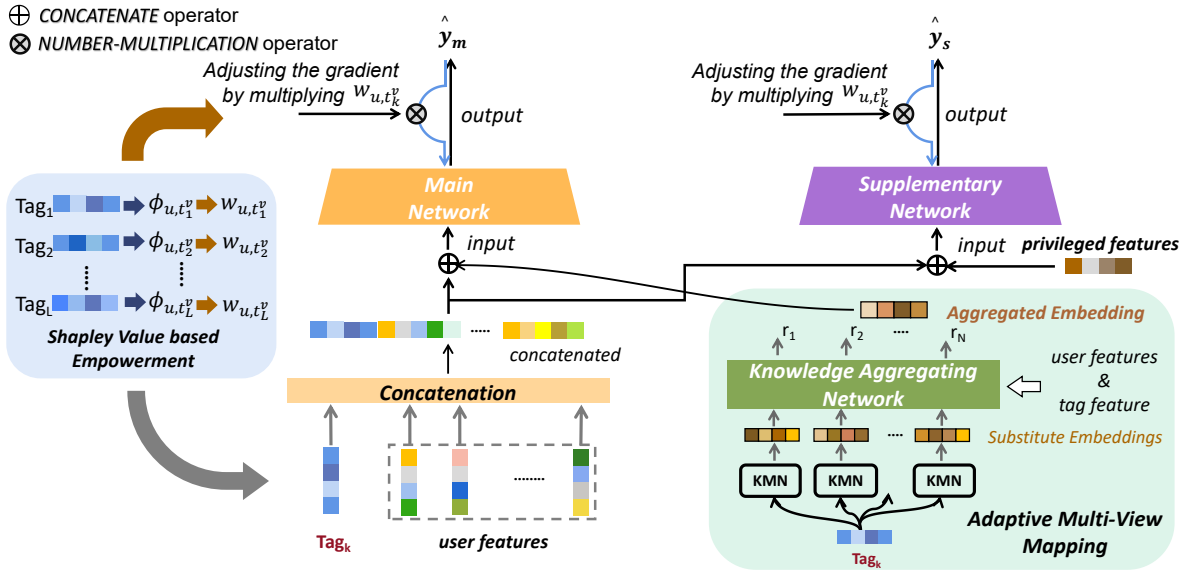


Figure 1: The overview of our proposed model Coalition Tag Multi-View Mapping (CTMVM).

Naturally, deep user profiling models have also come up nowadays. Weixin Group makes improvements to the Youtube model that highlights a multi-head attention mechanism with shared query vectors [5]. Tencent Ads introduces MVKE [6] to learn multiple topic-related user preferences based on different actions unitedly. Alibaba Group[7] presents a complete framework to address the data discrepancy issue in the training and testing stages. However, these models pay little attention to the coalition feature and the privileged feature in this field, which may cause the deviation of the model input and the inference objective.

2.2 Shapley Value for Attribution

Shapley Value [8] is a key attribution technique in coalition game theory, which has wide-ranging applications. Analysis of portfolio risk [17] has been a main purpose of Shapley Value for a long time. With the help of XGBoost, researchers [18] leverage Shapley Value to strengthen the interpretation of BANK A algorithm in credit scoring. Lundberg et al. [19] develops tree solutions for SHAP (SHapley Additive exPlanation) values to obtain unique, consistent, and locally correct imputation values for tree ensemble techniques.

Furthermore, explanation methods with Shapley Value shine brightly in online advertising attribution. Using various conditioning touch points, Shapley et al. [20] employ Shapley Value to quantify the contribution of a specific touch point. Later, Dalessandro et al. [21] demonstrate how the Shapley Value may help simulate the causal effects of various channels. Berman et al. [22] compare Shapley Value with the last-touch approach for ad channels attribution.

Shapley Value is widely applied as the attribution method to interpret feature importance or dig advertising channel contributions. In contrast, our model pioneers to introduce the Shapley Value into the domain of user-tag profiling.

2.3 Knowledge Transfer in Recommendation

Knowlwedge transfer aims to increase learning by transferring information across fields [23, 24]. Domain Adaptation methods are often applied to Cross-Domain Recommendation [25]. Common models for this task adopt embeddings rich in source domain knowledge to help the recommendation of target domains [26–29].

Multi-Task Learning exploits data in multiple tasks to increase the model generalization [30]. Multi-task learning is often divided into multi-behavior and multi-domain learning and recent years have seen lots of studies in this field [31–33].

Graph Neural Network [34–36] have shown outstanding performance by aggregating neighboring nodes to get better representation. However, GNN-based methods usually suffer from large time consumption [37] for online recommendation.

In addition, Knowledge Distillation is an effective form of transfer learning and is often applied to achieve model lightweighting and compression [38, 39]. Recently, Xu et al. [10] propose the privileged feature distillation to enhance the performance of recommendation.

Our proposed PKM and AMVM modules take an easy way to efficiently transfer privileged knowledge in user profiling models.

3 METHODOLOGY

In this section, we first give the formalized definition of the user-tag profiling task. Then we present the technical details of our proposed Coalition Tag Multi-View Mapping framework, the overview of which is shown in Figure 1.

3.1 Problem Formulation

For our task, we have a user set $\mathcal{U} = \{u_1, u_2, ...\}$, an item set $\mathcal{V} = \{v_1, v_2, ...\}$, a tag set $\mathcal{T} = \{t_1, t_2, ...\}$ and a list of privileged feature sets $\mathbf{P}^* = [\mathcal{P}_1, \mathcal{P}_2, ...]$, in which the i-th privileged feature set is denoted as $\mathcal{P}_i = \{p_{i1}, p_{i2}, ...\}$.

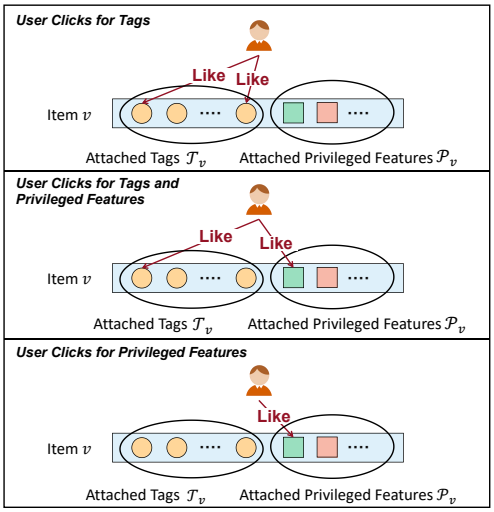


Figure 2: Some examples about the reasons for user clicks.

As shown in Figure 2, there are several tags in the item v and we denote these attached tags as the set $\mathcal{T}_v = \{t_1^v, t_2^v, ...\}$. Except for tags, the item v includes some privileged features $\mathcal{P}_v = \{p_{1i}^v, p_{2j}^v, ...\}$. Interaction labels $\mathcal{Y} = \{y_1, y_2, ...\}$ indicate a list of user behaviors where $y = 1$ for the click action and $y = 0$ for otherwise. Our task aims to extract user preferences for tags from these user-item interactions. All the features will be transformed to vectors through embedding look-up and the input of the base model is the concatenation of the user embedding and the tag embedding in each sample. To validate the model performance, we compute model outputs for each user to retrieve top-K tags that he likes best from the whole tag corpus.

3.2 Shapley Value based Empowerment

As we discussed before, the intrinsic causes of tags leading to user clicks are not yet clear, which obstructs user-tag profile modeling. Inspired by coalition game theory, we utilize a helpful attribution technich Shapley Value to assist the model catching logical user-tag links. This tool takes all subsets of players into account, which models all combinations of tags in a cooperative game. We get the contribution ϕ_{u,t_k^v} of the k-th tag t_k that user u clicks in the item v by:

$$\phi_{u,t_k^v} = \sum_{S \subseteq M \setminus t_k^v} \alpha(|S|)[V(S \cup \{t_k^v\}) - V(S)] \ . \tag{1}$$

Here M denotes the tag coalition $\mathcal{T}_v$ in the item v and S refers to a subset of M, which excludes the tag t_k^v and could be empty. The credit function $V(\cdot)$ is the gain brought by different subsets of tags, which equals the amount of clicks on the corresponding tag set by the interacted user. $S \cup t_k^v$ is the tag subset that includes S plus t_k^v. In Eq.(1), the subtraction formula in square brackets specifies the **marginal benefit** for adding t_k^v to the collection S while the weighting factor $\alpha(|S|)$ means the appearance probability of collection S, which is formulated in Eq.(2). Symbols $|S|$ and $|M|$ represent the number of tags in the set S and M, respectively. In other words, to get the contribution for tag t_k^v in this coalition, we just traverse all the subset collections and average the incremental credit by adding t_k^v.

$$\alpha(|S|) = \frac{|S|!(|M| - |S| - 1)!}{|M|!} \ . \tag{2}$$

By considering the joint effect of the tag coalition in a gaming way, the shapley value implies credible attribution results. With these values, we adjust the back-propagation strength of the samples associating with the item to empower the tags unevenly. To implement Shapley Value based Empowerment (SVE), we first turn each user-item sample into several user-tag samples, which also eliminates the gap between the training and testing stages revealed by [7]. In each click behavior, the gradient value of the sample that represents user u clicking tag t_k^v will be multiplied by the gradient weight w_{u,t_k^v}, which is formulated as followed:

$$w_{u,t_k^v} = \begin{cases} \frac{\exp(\phi_{u,t_k^v}/\tau)}{\sum_{t_j \in \mathbf{T}_v} \exp(\phi_{u,t_j^v}/\tau)} \cdot L \ , & u \text{ clicks } v \text{ with } \mathbf{T_v}. \\ 1 \ , & \text{otherwise.} \end{cases} \tag{3}$$

where τ means **Temperature Coefficient** to control the influence of Shapley Value and L is the number of tags in the item v to normalize the values around one.

Assume the output of the user-tag profiling model is $\hat{y}$, and the back-propagating gradient $\mathcal{G}_{\hat{y}}$ of the output $\hat{y}$ is revised as:

$$\mathcal{G}'_{\hat{y}} = w_{u,t_k^v} \cdot \mathcal{G}_{\hat{y}} \ . \tag{4}$$

where $\mathcal{G}'_{\hat{y}}$ indicates the revised gradient.

The strategy SVE tackles $I1$ by assigning definite credit to each tag in the item. However, in addition to tag features, there are some other features in the item attributing much to user actions. Based on transfer learning, we propose Privileged Knowledge Mapping.

3.3 Privileged Knowledge Mapping

Since users may be interested in store-ids or categories in the items and some tags are likely to frequently occur with particular privileged features, the relationships between them deserve deeply exploiting. Therefore we propose a knowledge mapping mechanism from the tag feature domain to the privileged feature domain. To produce informative privileged feature embeddings, we train a supplementary network $\mathbf{\Phi}_{\mathbf{W}_{\text{supp}}}(\cdot)$ with learnable parameters $\mathbf{W}_{\text{supp}}$, which gets the user embedding $\boldsymbol{X^u}$, the tag embedding $\boldsymbol{X^t}$ and the privileged embedding $\boldsymbol{X^p}$ as input. This network performs forward-propagation through the tower structure and predicts the clicking probability $\hat{y}_s$ as followed:

$$\hat{y}_s = \mathbf{\Phi}_{\mathbf{W}_{\text{supp}}}(\boldsymbol{X^u}, \boldsymbol{X^t}, \boldsymbol{X^p}) \ . \tag{5}$$

Synchronized training of the main and supplementary networks produces similar outputs and makes it more smooth to transmit knowledge across models, so we optimize the two networks $\mathbf{\Phi}_{\mathbf{W}_{\text{main}}}(\cdot)$ and $\mathbf{\Phi}_{\mathbf{W}_{\text{supp}}}(\cdot)$ at same pace. Specifically, the main network $\mathbf{\Phi}_{\mathbf{W}_{\text{main}}}(\cdot)$ will be fed with only user features and tag features while the privileged features are not involved in training. To transfer privileged knowledge to the main network, we then put forward a knowledge mapping mechanism for mapping tag embeddings to privileged embeddings, which is shown in Figure 3.

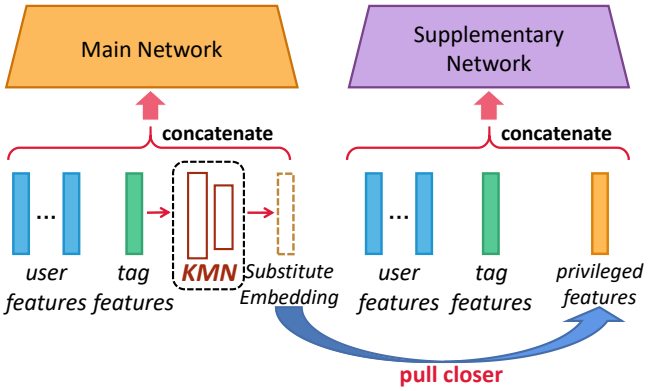


Figure 3: Prileged Knowledge Mapping that encodes the substitute privileged embedding and transfers privileged knowledge from the supplementary network to the main network.

Receiving instructions from the supplementary network, the Knowledge Mapping Network $f(\cdot)$ substitutes the privileged embedding $f(\boldsymbol{X^t})$ by encoding tag embedding $\boldsymbol{X^t}$. KMN contains two-layer neurons and a **Tanh** activation layer. Furthermore, to keep the consistency between training and testing stages, we only feed user features and tag features into the framework and make predictions as:

$$\hat{y}_{\mathrm{m}} = \Phi_{\mathbf{W}_{\mathrm{main}}}(\boldsymbol{X^u}, \boldsymbol{X^t}, f(\boldsymbol{X^t})) \,. \tag{6}$$

The main network and the supplementary network are trained synchronously and minimize the following cross-entropy loss:

$$\mathcal{L}(y, \hat{y}_{m/s}) = -y \log(\hat{y}_{m/s}) - (1-y)\log(1-\hat{y}_{m/s}) \,, \tag{7}$$

where $y \in \{0, 1\}$ and the subscript $\boldsymbol{m/s}$ indicates the logits $\hat{y}$ from the main network or the supplmentary network, repectively.

To guide the substitute embedding by privileged information, we adopt a mapping loss function $\mathcal{L}_{\mathrm{mapping}}$ to supervise $f(\boldsymbol{X^t})$ closer to privileged embeddings $\boldsymbol{X^p}$, which we choose as Mean Square Error (MSE). The mapping loss is formulated as:

$$\mathcal{L}_{\mathrm{mapping}} = ||\oslash(\boldsymbol{X^p}) - f(\boldsymbol{X^t})||^2 \,, \tag{8}$$

where $\oslash(\cdot)$ represents the **stop-gradient** operator to eliminate the adverse impact towards supplementary tower during training.

De-biasing Policy for SVE. To keep the training of the two networks in sync, we apply the SVE strategy to the both two networks for de-biasing. The policy is illustrated as:

$$\mathcal{G}'_{\hat{y}_{m/s}} = w_{u,t^v_k} \cdot \mathcal{G}_{\hat{y}_{m/s}} \tag{9}$$

However, we discover that many tags appear in more than one store or category, and even hundreds of times as many privileged features associate with one tag. In that case, a single substitute embedding in PKM can hardly accommodate all the privileged information. To solve this problem, we bring AMVM model that can fully capitalize on the privileged features.

3.4 Adaptive Multi-View Mapping

Instead of utilizing a single substitute embedding to encode the missing privileged features, AMVM takes several knowledge mapping networks $\{f_1(\cdot), f_2(\cdot), .., f_k(\cdot), ..\}$ to encode various embeddings for each tag, which is elucidated in Figure 4. To facilitate end-to-end learning, we employ a competitive training strategy to adaptively train diverse substitute embeddings.

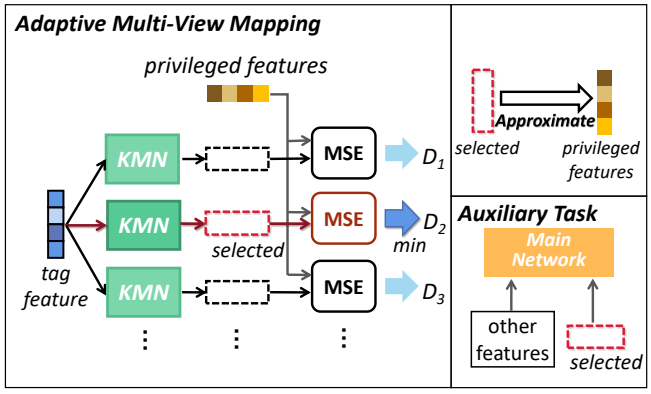


Figure 4: Illustration of AMVM details.

A simple idea to train numerous substitute embeddings is to approximate one specific embedding each time. Therefore, we select the closest substitute embedding $f_{\mathrm{selected}}(\boldsymbol{X^t})$ to the privileged embedding where the MSE is least. After sufficient training iterations, the difference between substitute embeddings accumulates, allowing it to be seperated into $\mathbf{N}$ clusters. Thus, $\mathcal{L}'_{\mathrm{mapping}}$ can be modified as adaptive multi-view mapping loss $\mathcal{L}_{\mathrm{AMVM}}$:

$$\mathcal{L}_{\mathrm{AMVM}} = ||\oslash(\boldsymbol{X^p}) - f_{\mathrm{selected}}(\boldsymbol{X^t})||^2 \,. \tag{10}$$

Knowledge Aggregating. Here we use a knowledge aggregating network to determine the adaptation ratings r_k of the k-th substitute embedding corresponding to the tag and aggregate them, defined as:

$$\begin{aligned} r'_k &= g(\boldsymbol{X^u}, \boldsymbol{X^t}, f_k(\boldsymbol{X^t})) \,, \\ r_k &= \frac{\exp(r'_k)}{\sum_{1\le l\le N}\exp(r'_l)} \,, \end{aligned} \tag{11}$$

where $f_k(\cdot)$ denotes the k-th encoder network, and $g(\cdot)$ denotes the knowledge aggregagting network to capture the importance between the encoded embeddings and the other features.

Then each encoded embedding will be merged to form the aggregation embedding $\boldsymbol{E}_{\mathrm{AGG}}$ defined as:

$$\boldsymbol{E}_{\mathrm{AGG}} = \sum_{k=1}^{N} r_k \cdot f_k(\boldsymbol{X^t}) \,. \tag{12}$$

Subsequently, the click probability $\hat{y}'_m$ and the prediction loss $\mathcal{L}_{\mathrm{predict}}$ of the main network should be revised as:

$$\hat{y}'_m = \Phi_{\mathbf{W}_{\mathrm{main}}}(\boldsymbol{X^u}, \boldsymbol{X^t}, \boldsymbol{E}_{\mathrm{AGG}}) \,, \tag{13}$$

$$\mathcal{L}_{\mathrm{predict}} = -y \log(\hat{y}'_m) - (1-y)\log(1-\hat{y}'_m) \,. \tag{14}$$

Task-Oriented Auxiliary Loss. Obviously, $f_{\mathrm{selected}}(\boldsymbol{X^t})$ should perform a comparable role to the privileged feature $\boldsymbol{X^p}$ in the current instance. Consequently, we introduce an auxiliary loss to train mapping networks towards labels and enhance the substitute embedding quality, which is shown in Eq.(15)~(16). To calculate

the auxiliary loss, the encoded embedding that is the closest to the privileged feature will be fed into the main network alone.

$$\hat{y}_{\text{aux}} = \Phi_{\mathbf{W}_{\text{main}}}(\boldsymbol{X}^{\boldsymbol{u}}, \boldsymbol{X}^{\boldsymbol{t}}, f_{\text{selected}}(\boldsymbol{X}^{\boldsymbol{t}})) , \tag{15}$$

$$\mathcal{L}_{\text{aux}} = -y \log(\hat{y}_{\text{aux}}) - (1-y) \log(1-\hat{y}_{\text{aux}}) . \tag{16}$$

The final combined loss function $\mathcal{L}_{\text{final}}$ is formulated in an end-to-end pattern as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{predict}} + \mathcal{L}_{\text{aux}} + \rho \cdot \mathcal{L}_{\text{AMVM}} , \tag{17}$$

where ρ is the factor that controls the learning intensity of privileged knowledge.

4 EXPERIMENTS

In this section, we demonstrate the experimental setup and detailed experiment results, which helps us better answer the following research questions:

- **RQ1:** Compared to state-of-the-art models, does our proposed CTMVM achieve the best performance?
- **RQ2:** What is the role of each component in the final model?
- **RQ3:** How sensitive is the model CTMVM to hyper-parameters τ, ρ and N?
- **RQ4:** Can CTMVM get a significant boost in online recommendation scenarios?

4.1 Experimental Setup

4.1.1 Dataset Descriptions. We perform comprehensive experiments in an industrial dataset and two open real-world datasets with privileged features in their items to assess the performance of our proposed CTMVM.

- **Taobao:** Taobao dataset records randomly sampled 10 days' logs of user historic behaviors on fashion apparel products, which is collected from real recommendation scenarios of the e-commerce company Alibaba. The samples from the previous nine days are utilized for training, while the tenth day's samples are used for model validation. This dataset consists of 1.49 million users, 4.95 million items and 0.15 billion samples, in which the user-side features comprise some user attributes (including ages, genders and cities) and the item-side features contain 0.43 million store-ids, 560 categories and 52 tags covering styles, patterns and material areas. We select store-ids and categories as the privileged features, which consumes two sets of AMVM modules to encode substitute embeddings with unshared mapping networks.
- **MovieLens**[1]**:** MovieLens dataset is gathered from the platform *grouplens.org*. Each user in the dataset holds a collection of tag assignments. Movie categories are privileged features in this dataset.
- **Delicious**[2]**:** Delicious dataset is a social recommendation dataset in which each user freely tags a list of bookmarks. Bookmark urls are regarded as privileged features.

In our experiments, we take the same preprocessing approach for the public datasets as in previous works [5, 40]. We randomly select 80% of all users for training and the remaining 20% for testing. In MovieLens dataset, ratings more than or equal to 4 are labeled as positive and the others are labeled as negative. For the interacted items of each user (i.e. movies, bookmarks), we sort them by timestamp . We select tags and privileged attributes in the top 80% as historical behaviors, and set the bottom 20% as interaction samples for the model. We eliminate tags that appear less than 5 times in MovieLens and 15 times in Delicious to prevent sparsity and set all unique Delicious url-ids that occur only once to *others*. The statistics of these datasets are summarized in Table 1.

Table 1: Datasets statistics. ***Privileged*** **indicates privileged features in the dataset.**

Dataset	# Users	# Items	# Tags	# Privileged	# Assignments
Taobao	1,490,386	4,953,070	52	433,688 & 560	158,886,300
MovieLens	138,495	12,404	1,046	20	1,194,429
Delicious	1,848	65,877	3,508	6,490	431,666

4.1.2 Evaluation Criterion. Top-K recommendation metrics are commonly used to assess the model quality of user-tag profiling. In this work, we evaluate the model performance by 4 metrics [41–43]: Precision, Recall, Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) [44].

4.1.3 Compared methods. To answer **RQ1**, the following benchmarks are chosen as comparisons to provide more convincing experimental results:

- **YouTube [45]:** YouTube model is a deep collaborative filtering model able to successfully deal with a large amount of data and learn the feature interaction in high-dimension space.
- **UTPM [5]:** UTPM model contains shared query vectors, a cross feature layer, and a joint loss. The attention method captures field attributes and reasonably weighs them.
- **CWTM [7]:** CWTM model addresses the gap between the training and inference procedure with the tag selection modules RMM and WMM. In addition, contrastive learning method is adopted to remain the lost aggregation information.

As is known that the single-tower form benefits model performance in contrast to two-tower form, our model CTMVM is constituted in the form of single-tower. For fair comparison, the compared methods above are all reproduced in single-tower form.

4.1.4 Model Details. During training, we first perform embedding look-up for user attributes, tags and privileged features. Then we adopt SVE to input the single tag and empower the main network and the supplementary network. Privileged features are also input for the training of AMVM. During testing, the model is only input user features and the target tag feature, while the privileged features will be replaced with the substitute embedding encoded by well-trained multi-view mapping networks.

4.1.5 Parameter Settings. Hidden layers in our model have [128, 64, 1] neurons and the activation functions are [$tanh$, $tanh$, $sigmoid$]. The dimensionality of embeddings is set to 16 for public datasets and 64 for Taobao dataset. For PKM and AMVM, we simultaneously train the supplemental network and the main network using Adam optimizer with the same learning rate 0.001. We adopt grid search to find the ideal model parameters. Parameter τ in SVE is set to

[1] http://www.grouplens.org
[2] http://www.delicious.com

0.1 for MovieLens and 0.5 for Delicious. For the coefficient of loss $\mathcal{L}_{\text{AMVM}}$, $\rho = 0.05$ for MovieLens and $\rho = 0.1$ for Delicious. For the number of AMVM, $\mathbf{N} = 4$ in MovieLens and $\mathbf{N} = 6$ in Delicious.

Table 2: Model performance in 3 datasets. Best results are in boldface. *Impr.* denotes the averaged relative improvement over each benchmark.

Dataset	Metric	YouTube	UTPM	CWTM	**CTMVM**
Taobao	Precision@3	0.0654	0.1566	0.2142	**0.2362**
	Recall@3	0.0219	0.0606	0.0796	**0.0910**
	HR@3	0.1779	0.4013	0.4367	**0.4920**
	NDCG@3	0.1121	0.2944	0.3183	**0.3419**
	Impr.	239.65%	34.93%	11.18%	—
	Precision@5	0.0976	0.1582	0.2386	**0.2666**
	Recall@5	0.0556	0.1004	0.1502	**0.1712**
	HR@5	0.3660	0.5593	0.6087	**0.6648**
	NDCG@5	0.1917	0.3571	0.3931	**0.4222**
	Impr.	145.68%	44.02%	10.59%	—
MovieLens	Precision@10	0.0219	0.0266	0.0741	**0.0891**
	Recall@10	0.0165	0.0211	0.0561	**0.0660**
	HR@10	0.1882	0.2069	0.4621	**0.5316**
	NDCG@10	0.0931	0.1015	0.2207	**0.2704**
	Impr.	244.86%	192.97%	18.86%	—
	Precision@20	0.0208	0.0239	0.0702	**0.0789**
	Recall@20	0.0318	0.0352	0.0949	**0.1170**
	HR@20	0.3210	0.3268	0.6059	**0.6820**
	NDCG@20	0.1255	0.1284	0.2621	**0.3036**
	Impr.	200.54%	176.83%	16.04%	—
Delicious	Precision@10	0.0209	0.0341	0.0568	**0.0632**
	Recall@10	0.0098	0.0177	0.0252	**0.0282**
	HR@10	0.1881	0.2613	0.3831	**0.4111**
	NDCG@10	0.0837	0.1292	0.1965	**0.2246**
	Impr.	169.40%	68.89%	11.20%	—
	Precision@20	0.0196	0.0348	0.0482	**0.0514**
	Recall@20	0.0180	0.0348	0.0413	**0.0451**
	HR@20	0.3108	0.4487	0.5389	**0.5514**
	NDCG@20	0.1140	0.1755	0.2238	**0.2609**
	Impr.	129.73%	37.23%	8.65%	—

4.2 Performance Comparison (RQ1)

In this section, we evaluate the performance of our proposed CTMVM model and some compared models for the user-tag profile modeling task. Considering the noteble magnitude difference of tag numbers in each dataset , we set $K \in \{3, 5\}$ in Taobao dataset and $K \in \{10, 20\}$ in MovieLens and Delicious datasets, respectively. The experiment results are summarized in Table 2, which reports the averaged results under multiple runs with 5 random seeds. Compared with the YouTube baseline, CTMVM has made consistently improvements in three datasets. It is obvious that our model has greater capacity to predict user preferred tags by charging the model with tag contributions and incorporating privileged features into training. In Taobao dataset, there is a relative boost of 239.65% for $K = 3$, while in public datasets MovieLens and Delicious, there are relative boosts of 244.86% and 169.40% for $K = 10$. Note that the boosts brought by each component in the model are different. We will show and discuss the specific details in Section 4.3.

Results from the experiments reveal that CTMVM achieves significant promotion over all baselines. Particularly, we can observe that the newest model CWTM, brought by Alibaba Group, exceeds other benchmarks by a wide margin. Despite the huge gain, our model still reaches 11.18% improvement of Precision@3 in the Taobao dataset, 18.86% and 11.20% improvements of Precision@10 in MovieLens and Delicious datasets compared to CWTM, which suggests that the feature-based framework CTMVM is more amenable than methods such as CWTM and UPTM to settling user-tag profile modeling.

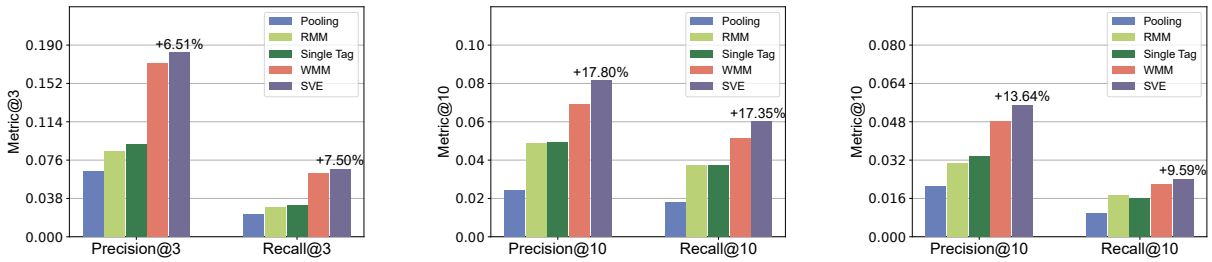
4.3 Further Analysis(RQ2)

4.3.1 The effectiveness of Shapley Value based Empowerment. To analyze the effect of SVE, we conduct comparison experiments on different modules that process tag features. The results are shown in Figure 5. In particular, we list five methods to input tags: (1) Pooling. Tag features will be input after one pooling layer. (2) RMM [7]. RMM module applies random masking on tags to get rid of the bias between training and testing. (3) Single Tag. Splitting all tags in one item into several user-tag instances for input can also reduce the training-tesing gap. (4) WMM [7]. This method selects the tag with high attention scores to input in each instance. (5) SVE. This is our model to empower the network by the shapley value on the basis of Single Tag. From the figure, we obtain the following observations:

- **Assigning right credits for tags during training can much promote the model.** Although RMM and Single Tag modules address the input bias to elevate the baseline performance, WMM and SVE attain ultra-high gains over them, by giving higher learning weights to the important tags in the item.
- **The module SVE achieves the best performance among all tag-input modules.** We can see that the module SVE performs the best in contrast to other modules. Moreover, compared to the latest technique WMM, SVE raises Precision@3 by 6.51% in Taobao dataset and Precision@10 by 17.80% and 13.64% in MovieLens and Delicious datasets, respectively.
- **Shapley value shows a greater advantage over the importance network in tag attribution.** According to the outcome in three datasets, we can find that Single Tag achieves similar results as RMM. However, single-tag based SVE consistently outperforms RMM-based WMM in all datasets. The result implies that the coalition game is better at demonstrating the pattern of tag cooperation in one item.

Table 3: Impact of privileged different knowledge transfer modules. *P@K* denotes *Precision@K* and *R@K* denotes *Recall@K*. Bolded values are best results and nderlined values are second best results.

Dataset	Taobao		MovieLens		Delicious	
Metric	P@3	R@3	P@10	R@10	P@10	R@10
YouTube	0.0654	0.0219	0.0219	0.0165	0.0209	0.0098
YouTube+KL	0.0898	0.0336	0.0248	0.0190	0.0247	0.0109
YouTube+MMD	0.0912	0.0391	0.0223	0.0168	0.0230	0.0103
LightGCN	0.0894	0.0343	0.0252	0.0189	0.0261	0.0130
PFD	0.1045	0.0374	0.0250	0.0191	<u>0.0268</u>	<u>0.0133</u>
PKM	<u>0.1088</u>	<u>0.0394</u>	<u>0.0253</u>	<u>0.0199</u>	0.0266	0.0129
AMVM	**0.1580**	**0.0568**	**0.0302**	**0.0232**	**0.0369**	**0.0167**



(a) Taobao (b) MovieLens (c) Delicious

Figure 5: The ablation study of different methods to handle Coalition Features in three datasets.

4.4 The effectiveness of privileged knowledge transfer

We investigate three model variants and one knowledge transfer technique to assess the knowledge transfer modules PKM and AMVM: (1) ***YouTube***: The baseline of our model. (2) ***YouTube+KL*** [46, 47]: Reducing the distance between the intermediary hidden layers of the supplementary (*teacher*) and main (*student*) networks is another transfer learning strategy [48]. In this model variant, we pull closer the first layer of both networks and measure the distance by the KL Divergence metric. (3) ***YouTube+MMD*** [49]: Similar to (2), we adopt Maximum Mean Discrepancy(MMD) as the distance metric. (4) ***LightGCN*** [35] To ensure the acceptable time delay for online inference [37], we choose the simple graph neural network framework GCN as the privileged knowledge transfer baseline. (5) ***PFD*** [10]:This is the state-of-the-art approach for inheriting privileged recommender features, which helps *student* network distill privileged knowledge from *teacher* network. We also list the results of PKM and AMVM, tag features of which are input after pooling layers. From the results in Table 3, we claim the following observations:

- **The importance of privileged features cannot be overlooked.** The enhancement of PKM over baseline YouTube is superior with one single embedding containing privileged knowledge. Other models involved privileged information can also improve over the baseline.
- **Excellent results exemplify the benefits of using embeddings to take privileged knowledge.** According to the table, PKM outperforms most knowledge transfer models. It is challenging to accurately integrate privileged knowledge in ***YouTube+KL*** and ***YouTube+MMD*** since the main network inputs lack embeddings to hold abundant privileged information, which greatly limits the expressiveness of the model.
- **Distillation of model inputs can achieve similar performance as the distillation of model outputs.** Evidently, PFD and PKM models behave similar across all datasets. While the former realizes distillation by approximating the inputs of the *teacher* and *student* networks and the latter performs distillation by pulling the outputs closer.
- **Multi-perspective privileged features are more expressive than single substitute embeddings.** AMVM with multiple mapping networks outperforms the model PKM and all other knowledge transfer models. Besides, we can see that privileged features outnumber tags a lot in Delicious and Taobao datasets. Hence the improvement of AMVM over PKM in them is bigger than in MovieLens dataset. These results suggest the diversified relationships of privileged features compared to tags.

4.4.1 Module Compatibility. To verify the compatibility of our modules, we further perform ablation study by adding modules PKM and AMVM to the SVE. The experimental results are displayed in Table 4. We can conclude from the results that:

- With the addition of PKM and AMVM, the performance gets significant boost in all datasets, showing good compatibility of these modules. The results prove that both the coalition feature and the privileged feature make sense in this task.
- The performance of SVE+AMVM reaches greater improvement compared to SVE+PKM in three datasets, which is consistent with the previous experimental results in 4.4 . Meanwhile, in Taobao and Delicious datasets, the relative improvement of SVE+AMVM over SVE+PKM outperforms that in MovieLens dataset for the same reasons explained in section 4.4.

Table 4: Ablation study for CTMVM, where *P@K* and *R@K* denote *Precision@K* and *Recall@K*. The best results are bolded.

Dataset	Taobao		MovieLens		Delicious	
Sub-model	P@3	R@3	P@10	R@10	P@10	R@10
SVE	0.1832	0.0674	0.0814	0.0602	0.0550	0.0240
SVE+KMN	0.1921	0.0701	0.0848	0.0627	0.0574	0.0251
SVE+AMVM	**0.2362**	**0.0910**	**0.0891**	**0.0660**	**0.0632**	**0.0282**

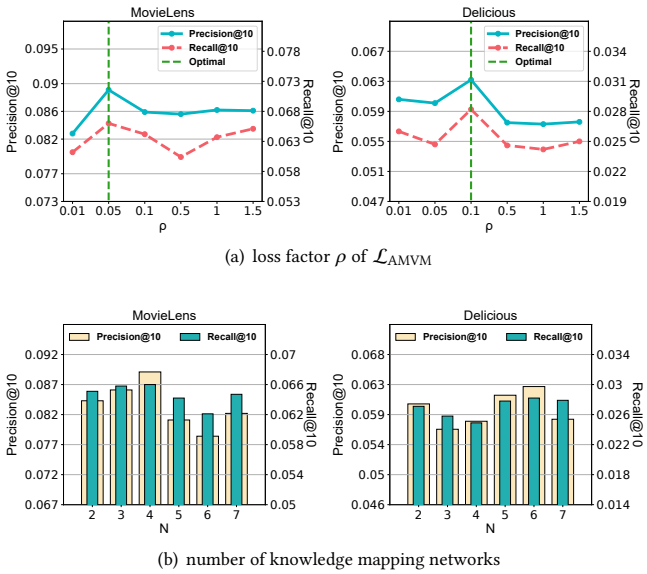
4.4.2 Qualitative Analysis. In our findings, the introduction of coalition features and privileged features could have remarkably rational influences on the recommendation results, and we will clarify this in the Appendix.

4.5 Hyper-Parameters Analysis (RQ3)

4.5.1 The temperature coefficient. This section examines sensitivity of temperature τ in SVE. Table 5 displays the performance fluctuation with the change of parameter τ. The model results first improve as the temperature coefficient increases, then declines. Experimental results peak at $\tau = 0.1$ in MovieLens and $\tau = 0.5$ in Delicious. It is not hard to explain that if the temperature coefficient is too big, the shapley values are not strong enough to empower the model, and if it is too small, the tag with a small shapley value can not get fully learned, which hurts the model performance.

Table 5: Analysis of temperature coefficient τ, where *P@K* denotes *Precision@K*, *R@K* denotes *Recall@K* and *N@K* denotes *NDCG*. Bolded values are the optimal results.

	MovieLens			Delicious		
Parameter	P@10	R@10	N@10	P@10	R@10	N@10
$\tau = 0.01$	0.0873	0.0651	0.2602	0.0592	0.0271	0.2110
$\tau = 0.05$	0.0889	0.0652	0.2657	0.0614	**0.0286**	0.2169
$\tau = 0.1$	**0.0891**	**0.0660**	**0.2704**	0.0613	0.0283	0.2137
$\tau = 0.5$	0.0663	0.0491	0.1954	**0.0632**	0.0282	**0.2246**
$\tau = 1.0$	0.0546	0.0404	0.1604	0.0567	0.0260	0.2088



(a) loss factor ρ of $\mathcal{L}_{\text{AMVM}}$

(b) number of knowledge mapping networks

Figure 6: Hyper-parameter study.

4.5.2 The loss factor of $\mathcal{L}_{AMVM}$. We study the influence of the loss factor ρ of $\mathcal{L}_{AMVM}$. As seen in Figure 6(a), privileged knowledge inheritance influences model performance. Best performance is at $\rho = 0.05$ in MovieLens and $\rho = 0.1$ in Delicious and model performance diminishes when ρ changes.

4.5.3 The number of knowledge mapping networks. Figure 6(b) illustrates model performance of public datasets by adjusting the number **N** of knowledge mapping networks. As is shown, too many or too few knowledge mapping networks can damage performance. The best results are achieved when $\mathbf{N} = 4$ for MovieLens and $\mathbf{N} = 6$ for Delicious. We can find that the optimal N for Delicious is greater than that for MovieLens, which confirms the intuitive premise that the more privileged features there are, the more knowledge mapping networks should be applied to maximize performance.

4.6 Industrial Results (RQ4)

Except for these offline experiments, we also conduct online A/B test on Alibaba's large-scale recommendation systems. Our recommendation algorithm shows consumers personalized card themes like "Vintage style theme". In the first stage of our business scenario, we use Theme-CTR(Click-Through-Rate) to measure the user's preference for themes. Once users click the theme, it jumps to a new page with lots of items which have the same tag with the theme. We focus on Item-CTR to measure the user's preference for items in the second stage. Since our themes are related to tags and our items are organized by tags, higher Theme-CTR or Item-CTR both represent a better user-tag preference.

For the control group, we adopt the current solution working online to recommend themes to users, which is mainly based on theme items and some basic theme features such as theme IDs and historical statistic features. For the experimental group, we employ our user-tag profiling model to select proper themes based on theme tags in the first stage. For fair comparison, the two groups share the same matching and ranking strategies. With careful online A/B test, our model contributes gain by 10.81% to the Theme-CTR and by 6.74% to the Item-CTR, which proves the superior efficacy of our proposed model. It is worth noting that all the experiments only introduce the user-tag preferences in the first stage, while the recommended items and recommendation strategies in the second stage remain unchanged. However, the Item-CTR in the second stage was also greatly improved, which further illustrates the importance of user-tag profiling in the recommendation systems.

5 CONCLUSION

In this paper, we propose the CTMVM framework for user-tag profile modeling, which is crucial in personalization services. We discover two prominant features particular in this task, Coalition Feature and Privileged Feature. To model these features, we first adopt Shapley Value based Empowerment to assign credit for each tag, and then propose Knowledge Mapping Network to transfer the privileged information in evaluation. Eventually we further put forward Adaptive Multi-View Mapiing model to alleviate poor distillation effect caused by insufficient expression of single embeddings. Excellent offline and online results prove the validity of our model. In the future, our team will conduct research on the interests mining of user long-term tag behaviors.

ACKNOWLEDGMENTS

The work is partially supported by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Artificial Intelligence Innovation and Development Fund (2020-RGZN-02026) and Alibaba Group through Alibaba Innovative Research (AIR) Program.

REFERENCES

[1] Dwayne Ball, Pedro S Coelho, and Manuel J Vilares. Service personalization and loyalty. *Journal of services marketing*, 2006.

[2] Wanvimol Nadee. *Modelling user profiles for recommender systems*. PhD thesis, Queensland University of Technology, 2016.

[3] Tarmo Robal and Ahto Kalja. Applying user profile ontology for mining web site adaptation recommendations. *ADBIS Research Communications*, 325, 2007.

[4] Jong-Hyuk Roh and Seunghun Jin. Personalized advertisement recommendation system based on user profile in the smart phone. In *2012 14th International Conference on Advanced Communication Technology (ICACT)*, pages 1300–1303. IEEE, 2012.

[5] Su Yan, Xin Chen, Ran Huo, Xu Zhang, and Leyu Lin. Learning to build user-tag profile in recommendation system. In *CIKM*, pages 2877–2884, 2020.

[6] Zhenhui Xu, Meng Zhao, Liqun Liu, Xiaopeng Zhang, and Bifeng Zhang. Mixture of virtual-kernel experts for multi-objective user profile modeling. *arXiv preprint arXiv:2106.07356*, 2021.

[7] Chenxu Zhu, Peng Du, Xianghui Zhu, Weinan Zhang, Yong Yu, and Yang Cao. User-tag profile modeling in recommendation system via contrast weighted tag masking. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4630–4638, 2022.

[8] L. S. Shapley. A value for n-person games. 1952.

[9] Eyal Winter. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054, 2002.

[10] Chen Xu, Quan Li, Junfeng Ge, Jinyang Gao, Xiaoyong Yang, Changhua Pei, Fei Sun, Jian Wu, Hanxiao Sun, and Wenwu Ou. Privileged features distillation at taobao recommendations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2590–2598, 2020.

[11] Yanping Zhang, Juan Jiang, Yongliang Zha, Heng Zhang, and Shu Zhao. Research on embedding capacity and efficiency of information hiding based on digital images. 2013.

[12] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. Multi-interest network with dynamic routing for recommendation at tmall. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2615–2623, 2019.

[13] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2942–2951, 2020.

[14] Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in folksonomies. In *European conference on principles of data mining and knowledge discovery*, pages 506–514. Springer, 2007.

[15] Leandro Balby Marinho and Lars Schmidt-Thieme. Collaborative tag recommendations. In *Data Analysis, Machine Learning and Applications*, pages 533–540. Springer, 2008.

[16] Deqing Yang, Yanghua Xiao, Hanghang Tong, Junjun Zhang, and Wei Wang. An integrated tag recommendation algorithm towards weibo user profiling. In *International conference on database systems for advanced applications*, pages 353–373. Springer, 2015.

[17] Stephane Mussard and Virginie Terraza. The shapley decomposition for portfolio risk. *Applied Economics Letters*, 15(9):713–715, 2008.

[18] RIDA Abdollah. Machine and deep learning for credit scoring: A compliant approach. 2019.

[19] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[20] Lloyd S Shapley. A value for n-person games. *Classics in game theory*, 69, 1997.

[21] Brian Dalessandro, Claudia Perlich, Ori Stitelman, and Foster Provost. Causally motivated attribution for online advertising. In *Proceedings of the sixth international workshop on data mining for online advertising and internet economy*, pages 1–9, 2012.

[22] Ron Berman. Beyond the last touch: Attribution in online advertising. *Marketing Science*, 37(5):771–792, 2018.

[23] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[24] Weike Pan. A survey of transfer learning for collaborative recommendation with auxiliary data. *Neurocomputing*, 177:447–453, 2016.

[25] Lili Zhao, Sinno Jialin Pan, Evan Wei Xiang, Erheng Zhong, Zhongqi Lu, and Qiang Yang. Active transfer learning for cross-system recommendation. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[26] Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. Cross-domain recommendation: An embedding and mapping approach. In *IJCAI*, volume 17, pages 2464–2470, 2017.

[27] Yongchun Zhu, Zhenwei Tang, Yudan Liu, Fuzhen Zhuang, Ruobing Xie, Xu Zhang, Leyu Lin, and Qing He. Personalized transfer of user preferences for cross-domain recommendation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1507–1515, 2022.

[28] Yongchun Zhu, Kaikai Ge, Fuzhen Zhuang, Ruobing Xie, Dongbo Xi, Xu Zhang, Leyu Lin, and Qing He. Transfer-meta framework for cross-domain recommendation to cold-start users. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1813–1817, 2021.

[29] Ye Bi, Liqiang Song, Mengqiu Yao, Zhenyu Wu, Jianming Wang, and Jing Xiao. A heterogeneous information network based cross domain insurance recommendation system for cold start users. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2211–2220, 2020.

[30] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[31] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939, 2018.

[32] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1137–1140, 2018.

[33] Dongbo Xi, Zhen Chen, Peng Yan, Yinger Zhang, Yongchun Zhu, Fuzhen Zhuang, and Yu Chen. Modeling the sequential dependence among audience multi-step conversions with multi-task learning in targeted display advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3745–3755, 2021.

[34] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[35] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020.

[36] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[37] Zhihao Jia, Sina Lin, Mingyu Gao, Matei Zaharia, and Alex Aiken. Improving the accuracy, scalability, and performance of graph neural networks with roc. *Proceedings of Machine Learning and Systems*, 2:187–198, 2020.

[38] Guorui Zhou, Ying Fan, Runpeng Cui, Weijie Bian, Xiaoqiang Zhu, and Kun Gai. Rocket launching: A universal and efficient framework for training well-performing light net. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[39] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[40] Bo Chen, Wei Guo, Ruiming Tang, Xin Xin, Yue Ding, Xiuqiang He, and Dong Wang. Tgcn: Tag graph convolutional network for tag-aware recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 155–164, 2020.

[41] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958, 2019.

[42] Zhenghua Xu, Cheng Chen, Thomas Lukasiewicz, Yishu Miao, and Xiangwu Meng. Tag-aware personalized recommendation using a deep-semantic similarity model with negative sampling. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1921–1924, 2016.

[43] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174, 2019.

[44] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[45] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[46] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[47] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[48] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[49] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

A QUALITATIVE ANALYSIS

We find via our research that different modules cause the model to choose various items for each user. This motivates us to draw additional conclusions about how our modules affect the recommendation results. To further demonstrate the user-tag profiles generated by the CTMVM model, we introduce two metrics to evaluate the effectiveness of its components.

A.1 The Influence of Coalition Feature

The module SVE pre-calculates the shapley values of each tag as the tag contribution weights and empowers the model according to these values in positive samples. Therefore, the model tends to recommend tags with high shapley values to the users. To verify this point, we propose the metric $AvgShapley@K$, which is defined as follows:

$$AvgShapley@K = \frac{1}{N} \cdot \sum_{i=1}^{N} AvgShapley\_User_i \tag{18}$$

$$AvgShapley\_User_i = \frac{1}{K} \cdot \sum_{j=1}^{K} Shap(Tag_{ij}) \tag{19}$$

where Tag_{ij} denotes the j-th tag that is recommended to the i-th user, and $Shap(Tag_{ij})$ is obtained by dividing the sum of the Tag_{ij}'s shapley values in the whole training samples by its occurrence number.

Table 6: AvgShapley@K between different models during inference in MovieLens dataset

Model	AvgShapley@10	AvgShapley@20
YouTube	0.0143	0.0156
YouTube+SVE	0.0237	0.0240

In Table 6, we summarizes the performance of models YouTube and YouTube+SVE. From this table, we can conclude that the module SVE enhances the model performance by inducing the model to recommend significant tags to users, which corroborates the validity of shapley values.

A.2 The Influence of Privileged Feature

The module PKM and AMVM can help the model distinguish item tags in a more fine-grained way and achieves more personalized tag recommendations. To figure out the effectiveness of introducing privileged feature knowledge into training, we supplement another experiment showned in Table 7.

Table 7: Tag_Privileged_HR@K between different models during inference in Delicious dataset

Sub-Model	Tag_Privileged_HR@10	Tag_Privileged_HR@20
SVE	0.4892	0.5432
SVE+PKM	0.5243	0.5784
SVE+AMVM	0.5649	0.6011

In this table, we display the metric $Tag\_Privileged\_HR@K$ of SVE, SVE+PKM and SVE+AMVM, respectively. The metric is defined as follows:

$$Tag\_Privileged\_HR@K = \frac{1}{N} \cdot \sum_{i=1}^{N} I(|P_i@K \cap C_i|) \tag{20}$$

$$P_i@K = P_{i1} \cup P_{i2} \cup ... \cup P_{iK} \tag{21}$$

$$I(x) = \begin{cases} 1 & x > 0 \\ 0 & otherwise \end{cases} \tag{22}$$

where P_{ij} is the set of privileged features that have been in the same items with the k-th tag, which is recommended to the i-th user, and C_i is the set of privileged features that the i-th user clicked.

From the results, we can conclude that the introduction of privileged features motivates the model to recommend tags that have higher relevance with the privileged faetures that users like, which brings in privileged information to the model.