# Beyond Co-occurrence: Multi-modal Session-based Recommendation

Xiaokun Zhang, Bo Xu, Fenglong Ma, Chenliang Li, Liang Yang and Hongfei Lin

**Abstract**—Session-based recommendation is devoted to characterizing preferences of anonymous users based on short sessions. Existing methods mostly focus on mining limited item co-occurrence patterns exposed by item ID within sessions, while ignoring what attracts users to engage with certain items is rich multi-modal information displayed on pages. Generally, the multi-modal information can be classified into two categories: descriptive information (*e.g.*, item images and description text) and numerical information (*e.g.*, price). In this paper, we aim to improve session-based recommendation by modeling the above multi-modal information holistically. There are mainly three issues to reveal user intent from multi-modal information: (1) How to extract relevant semantics from heterogeneous descriptive information with different noise? (2) How to fuse these heterogeneous descriptive information to comprehensively infer user interests? (3) How to handle probabilistic influence of numerical information on user behaviors? To solve above issues, we propose a novel multi-modal session-based recommendation (MMSBR) that models both descriptive and numerical information under a unified framework. Specifically, a pseudo-modality contrastive learning is devised to enhance the representation learning of descriptive information. Afterwards, a hierarchical pivot transformer is presented to fuse heterogeneous descriptive information. Moreover, we represent numerical information with Gaussian distribution and design a Wasserstein self-attention to handle the probabilistic influence mode. Extensive experiments on three real-world datasets demonstrate the effectiveness of the proposed MMSBR. Further analysis also proves that our MMSBR can alleviate the cold-start problem in SBR effectively.

**Index Terms**—Session-based recommendation, Multi-modal learning, Pseudo-modality contrastive learning, Hierarchical pivot transformer, Probabilistic modeling.

✦

## 1 INTRODUCTION

As an important tool to combat information overload, recommender system (RS) plays a vital role in present information era. Especially in context of e-commerce, RS facilitates online consumption by offering personalized services to individuals. Assuming the user identity information is accessible, conventional RS [1], [2] relies on user profiles and long-term behaviors to predict their preferences. However, in most real-world scenarios, the user identification is not available due to privacy policy or unlogged-in cases, where what RS could use is the short behavior sequences of anonymous users (*i.e.*, sessions). Apparently, conventional RS methods are no longer applicable or satisfactory in this case. To handle this situation, session-based recommendation (SBR) is proposed to predict next items interested by anonymous users within short sessions [3]. Nowadays, SBR has drawn significant attention from both academia and industry due to its highly practical value [4], [5].

With capacity in capturing transition patterns among items in a session, various neural networks are employed to improve SBR, such as recurrent neural networks (RNN) [4], [6], convolutional neural networks (CNN) [8], attention

- *Xiaokun Zhang, Bo Xu, Liang Yang and Hongfei Lin are with the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. E-mail: dawnkun1993@gmail.com, {xubo, liang, hflin}@dlut.edu.cn*
- *Fenglong Ma is with the College of Information Sciences and Technology, Pennsylvania State University, Pennsylvania, USA. E-mail: fenglong@psu.edu*
- *Chenliang Li is with the School of Cyber Science and Engineering, Wuhan University, Wuhan, China. E-mail: cllee@whu.edu.cn*
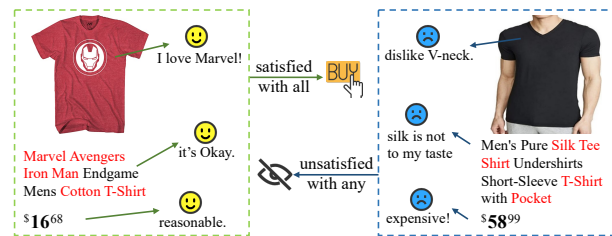
Fig. 1: A user makes the decision after evaluating all multi-modal information displayed on pages including item images, description text and price.

mechanisms [7], [9], and graph neural networks (GNN) [10], [11]. Despite having made impressive progress, most existing methods still rely on mining *co-occurrence patterns* exposed by item ID within short sessions. This significantly limits their performance since that a session usually contains a few items under the scenario of SBR (as shown in Table 2, the average session length is no more than 3). In other words, there are not enough co-occurrence patterns for them to exploit for user intent modeling in SBR. Fortunately, the available *multi-modal information* of items provides a promising antidote to improve SBR.

Intuitively, it is the multi-modal information displayed on pages that drives users to engage with certain items. As shown in Fig. 1, a user makes a decision usually after looking through item *images*, reading description *text*, and checking *price*. Since relying on different vehicles for conveying particular item features, the multi-modal information can be categorized into two groups: *descriptive* information and *nu-*

*merical* information. The descriptive information portrays an item with image and text that can intuitively describe some item features like style, color and material. For numerical information, *i.e.,* price, it delivers abstract value of an item through real numbers. In most cases, as illustrated in Fig. 1, a user would not click an item unless she is satisfied with its all aspects. Obviously, the above multi-modal information jointly determines a user's choice.

In fact, different from item ID that merely contains item co-occurrence patterns, multi-modal information presents extensive characteristics of items and encodes user fine-grained preferences. For example, a Marvel fan may have a high probability to purchase a T-shirt with the logo of iron man. Unfortunately, most existing models take neither images nor text into consideration, leading to their failure for accurate intent understanding. Moreover, co-occurrence based methods usually suffer from cold-start problem where there is no sufficient data to signify relations among new items [12]. This issue will be smoothly solved, if we can understand user preferences from multi-modal features instead of dull item ID. Although some recent models try to incorporate side information to facilitate user preferences learning, such as item category [13], description text [12] and price [14], they are still unable to reveal user intent holistically with such fragmentary information. Thus, to fully understand user fine-grained preferences, we should consider the entire multi-modal information displayed on pages. However, it's nontrivial to utilize multi-modal information in SBR due to following obstacles:

(1) *Descriptive information representation*. Under SBR scenario, images and text possess distinct noise. Normally, an item image not only contains the item for sale such as a cloth but also extra contents like accessories of the cloth. Similarly, an item description text usually includes redundant words like exaggerated statements to attract user attention. The existence of such noise in images and text increases the difficulty of extracting item semantics, hindering precise user preferences learning. Therefore, the first challenge is how to obtain relevant semantics from heterogeneous descriptive information with different noise.

(2) *Descriptive information fusion*. For an item, both image and text are utilized to describe its characteristics. Obviously, there exists shared information between them. At the same time, they also hold different purposes and focus on presenting distinct properties of items. To be specific, images are more intuitive than text to describe item colors and styles. Text can clearly express the material, *e.g.,* silk or cotton, whereas we can hardly understand it from images. Thus, the image and text complement each other and present an item in a united way. Accordingly, to comprehensively infer user interest, another challenge is how to fuse these heterogeneous descriptive information.

(3) *Numerical information modeling*. In general, a user's taste is *deterministic* on descriptive information. For instance, a user who prefers crewneck T-shirts may not click suggested ones with V-neck. In contrast, numerical price affects user behaviors in a *probabilistic* way. More precisely, as long as the item price falls in a user's acceptable range, it does not matter if the price is slightly lower or higher. Thus, the last challenge is how to handle the probabilistic influence of numerical information on user behaviors.

In order to tackle above challenges, we propose a novel Multi-Modal Session-Based Recommendation (MMSBR) that customizes both *deterministic and probabilistic modelings* to handle descriptive and numerical information respectively. In the deterministic modeling, we devise a *pseudo-modality contrastive learning* to refine descriptive information representations. In particular, contrastive learning is used to enhance representation learning by pushing semantically similar (positive) pairs close, while pulling dissimilar (negative) pairs apart [15]. Since different modalities of an item refer to similar contents, it is intuitive to view them as positive pairs to tackle the noise issue. However, there are semantic gaps between distinct modalities, making it inappropriate to directly contrast them. To address this issue, we propose to utilize one modality to generate pseudo-information (namely pseudo-modality) in another modality via data generation techniques. The actual and pseudo modalities which are aligned in the same semantic space are then used as positive pairs in contrastive learning to mitigate noise existing in images and text.

Moreover, to fuse descriptive information, we present a *hierarchical pivot transformer* in deterministic modeling. With the ability in modeling complex relations in sequences, Transformer structure has shown to be effective for merging multi-modal signals [17], [19]. Inspired by this, we further create a pivot, which serves as a mixer of valuable information, in each transformer layer to govern the fusion of heterogeneous information. The pivot hierarchically extracts and integrates useful information from images and text under Transformer operations. We then view the pivot as the comprehensive embedding of descriptive information.

In probabilistic modeling, we first represent item price as a *Gaussian distribution embedding*, which enables MMSBR to perceive range property of item price. The *Wasserstein self-attention* is then developed to handle price distribution embeddings for obtaining user acceptable price range. With the capacity in distinguishing differences between Gaussian distributions, the Wasserstein distance [20], [21] is used in the Wasserstein self-attention to determine the relevance among price distribution embeddings. Finally, the proposed MMSBR provides personalized services for users via evaluating the entire multi-modal information displayed on pages. In summary, the main contributions of our work are as follows:

- We propose a novel MMSBR to characterize user preferences based on multi-modal information, which is more in line with user decision-making process than conventional co-occurrence based methods. To our best knowledge, this is the first work to reveal user intent from multi-modal information in SBR.
- We classify multi-modal information into descriptive and numerical types. Accordingly, we customize deterministic and probabilistic modeling that consist of several innovative techniques for comprehensively mining user intent.
- Extensive experiments over three public benchmarks demonstrate the superiority of MMSBR over state-of-the-art methods. Further analysis also justifies the effectiveness of MMSBR under cold-start scenario.

## 2 RELATED WORK

Considering that this work aims to improve session-based recommendation by incorporating multi-modal information, we briefly review the related work from following two aspects: (1) session-based recommendation including co-occurrence based methods and side information enhanced methods; (2) multi-modal recommendation.

### 2.1 Session-based Recommendation

**Co-occurrence based methods**. Recent years, with the tremendous achievements of neural networks in various applications, we have witnessed the transition from traditional methods to neural models in SBR [3]. With the intrinsic ability to handle sequential data, RNN as well as its variants are the first neural networks applied in SBR. For example, GRU4Rec [6] utilizes gated recurrent unit (GRU) to capture sequential patterns within sessions. NARM [4] enhances GRU4Rec with attention mechanism to explore user main purpose. Afterwards, many neural architectures are employed to model user sequential behaviors such as CNN [8], attention mechanism [7], [9], [23], GNN [10], [11], and reinforcement learning [50]. Some approaches further enhance the learning for co-occurrence patterns via exploring extra sessions [22], [24], [25], multi behaviors [26], multi user intents [27] and multi relations among items [5], [28]. Contrastive learning is an emerging technique whose target is to improve embeddings by enlarging the distance between positive and negative pairs. Many recent models utilize the technique to enable robust representation learning for accurate user intent modeling [15], [29]. Also, other methods design new distance functions with metric learning to optimize user preferences learning [21]. Although greatly promoting the development of SBR, all of these methods, essentially, focus on mining co-occurrence patterns reflected by item ID. They fail to perceive user fine-grained preferences concealing in multi-modal information, which becomes a bottleneck limiting their performance.

**Side information enhanced methods**. Considering that side information can help to unveil user unique taste, some methods try to incorporate various kinds of information to improve recommendation performance such as time (*aka* positions) [30], [32], categories [13], [31], price [14], text [12], [33] and images [34]. There are also a few works [16], [35], [36] taking both text and images into account to handle long sequential behaviors of users. These methods have proved the effectiveness of side information in understanding user interest. However, most of them conduct information fusion with simple concatenation or addition, leading to their failure in effectively merging various information. Moreover, they can neither alleviate noise in various modalities nor distinguish influence modes of distinct modalities on user behaviors. In addition, to our best knowledge, none of existing methods collectively considers entire multi-modal information displayed on the websites, *i.e.*, images, text and price, to simulate user behaviors. Thus, we propose a novel MMSBR to holistically reveal user intent from these multi-modal information, which is consistent with genuine decision-making process.

TABLE 1: Important Notations.

| Notation | Description |
| --- | --- |
| $\mathcal{I}, n/|\mathcal{I}|$ | item set, the total number of items |
| $x_i$ | an item |
| $\mathcal{S} = [x_1, x_2, ..., x_m]$ | a session with $m$ items |
| $x_i^{img}, x_i^{txt}, x_i^{pri}$ | item image, description text and price |
| $v_i^{pri}$ | price encoding |
| $\mathbf{e}_i^{img}$ | actual image embedding |
| $\mathbf{e}_i^{txt}$ | actual text embedding |
| $\mathbf{e}_i^{pseimg}$ | pseudo image embedding |
| $\mathbf{e}_i^{psetxt}$ | pseudo text embedding |
| $\mathbf{e}_i$ | descriptive information embedding |
| $\mathbf{e}_i^{pri}$ | numerical information embedding |
| $\mathbf{s}_d$ | representing user deterministic taste |
| $\mathbf{s}_p$ | representing user acceptable price range |

### 2.2 Multi-modal Recommendation

Multi-modal recommendation has received increasing attention recently, since that we humans perceive the world by concurrently processing and fusing multi-modal information [19], [37]. To name a few, some methods [38], [39] employ Graph Neural Networks (GNN) and incorporate item images and text into the user-item interaction graph to facilitate user preferences and item characteristics learning. Beside, another line of research [40], [41] utilizes the pre-training technique to inject rich knowledge from item visual and textual modalities into recommender systems. More recently, BM3 [42] improves user and item representations by optimizing three multi-modal objectives including replicating user-item interaction graph and aligning modality features in inter- and intra-modality. Unfortunately, these methods fail to handle the situation of SBR because that they require users' identification and long-term behaviors to guide the model learning. Furthermore, there is no efforts bridging multi-modal information and SBR, hence we are the first to fill this research gap.

## 3 PRELIMINARIES

### 3.1 Problem Statement

Session-based recommendation (SBR) is proposed to provide personalized services for anonymous users based on their short behavior sequences. Let $\mathcal{I}$ signify the set of all unique items, where $|\mathcal{I}| = n$ is the total number of items. Normally, as depicted in Fig. 1, an item $x_i \in \mathcal{I}$ ($1 \leqslant i \leqslant n$) is presented to users in the form of multi-modal information including item image ($x_i^{img}$), description text ($x_i^{txt}$) and price ($x_i^{pri}$), *i.e.*, $x_i = \{x_i^{img}, x_i^{txt}, x_i^{pri}\}$. In SBR, an anonymous user has chronologically interacted with $m$ items in a certain interval, producing a session $\mathcal{S} = [x_1, x_2, ..., x_m]$, where $x_m \in \mathcal{I}$. Our goal is to predict next item the user will prefer based on $\mathcal{S}$. Note that, we rely on rich multi-modal information users can access instead of dull item ID to reveal user intent, which enables our MMSBR to capture user fine-grained preferences and support cold-start scenario easily. The important notations used in this work are detailed in Table 1.

## 3.2 Multi-modal Information Encoding

Considering that distinct modalities are presented to users with completely different forms, *i.e.*, images in RGB, text in symbolic words and price in real numbers, we need to handle these information via special methods so that they can serve as inputs to neural models. In the next, we will detail how we encode these modalities, *i.e.*, image ($x_i^i$), text ($x_i^t$) and price ($x_i^p$).

**Image embedding.** The first thing that a user may notice while browsing e-commerce websites is the item image. Due to the strong ability of GoogLeNet [44] in extracting semantics from images [35], we apply it to obtain image embedding $\mathbf{e}_i^{img} \in \mathbb{R}^d$ from original image $x_i^{img}$ via,

$$\mathbf{e}_i^{img} = \text{imgEmb}(x_i^{img}), \tag{1}$$

where the $\text{imgEmb}(\cdot)$ denotes the GoogLeNet model pre-trained on a large number of images.

**Text embedding.** After watching the image, the user further approaches the item by reading its description text. BERT [43] has been proved to be good at extracting text semantics by many studies [41], [45]. Therefore, we employ it to learn text embedding $\mathbf{e}_i^{txt} \in \mathbb{R}^d$ from original description text $x_i^{txt}$ via,

$$\mathbf{e}_i^{txt} = \text{textEmb}(x_i^{txt}), \tag{2}$$

where the $\text{textEmb}(\cdot)$ denotes the BERT model pre-trained on large text corpus.

**Price encoding.** After evaluating descriptive image and text of an item, the user would check the item price to determine whether to purchase it. The absolute price cannot accurately indicate weather an item is expensive or not because that the price varies greatly across different categories (*e.g.,* tens of dollars for clothes and hundreds of dollars for electronics). Thus, for an item with price $x_i^{pri}$ in a certain category, we encode its price level via,

$$v_i^{pri} = \lfloor \frac{x_i^{pri} - \min}{\max - \min} \times \rho \rfloor, \tag{3}$$

where $[\min, \max]$ is the price range of its category, and $\rho$ is the total number of price levels. Notably, such a operation enables item price to be compared across different categories [14].

## 4 METHODOLOGY

In this section, we will elaborate on the proposed MMSBR which is illustrated in Fig. 2. MMSBR is mainly composed of following interdependent components: (1) *Deterministic modeling* is devised to handle descriptive information, *i.e.,* item image and description text, to capture user deterministic taste; (2) *Probabilistic modeling* is developed to copy with numerical information, *i.e.,* item price, for modeling user acceptable price range; (3) *Prediction* provides personalized services for individuals based on entire multi-modal information displayed on pages.

## 4.1 Deterministic Modeling

Deterministic modeling employs: (1) *pseudo-modality contrastive learning* to refine descriptive information representations; (2) *hierarchical pivot transformer* to fuse heterogeneous descriptive information; (3) *vanilla attention* to capture user deterministic taste.

### 4.1.1 *Pseudo-modality Contrastive Learning*

As stated before, there exists noise in item images and text, leading to inaccurate item semantics extraction. Contrastive learning can tackle this issue by maximizing the agreements between semantically similar pairs. However, image and text embeddings coming from an item locate in distinct semantic spaces. Thus, it will corrupt the original semantics if we directly view them as positive pairs. To obtain effective contrastive signals, we resort to data generation techniques to generate pseudo modality which is aligned in the same space as the corresponding actual modality. Afterwards, with the generated contrastive signals, the contrastive learning is utilized to refine image and text embeddings.

**Pseudo-modality generation.** DALL·E [46] is an emerging technique to produce vivid images according to short text. For a piece of text $x_i^{txt}$, therefore, we feed it into the DALL·E to generate pseudo image $x_i^{pseimg}$. We then use the $\text{imgEmb}(\cdot)$ to get the pseudo image embedding $\mathbf{e}_i^{pseimg} \in \mathbb{R}^d$ via,

$$\mathbf{e}_i^{pseimg} = \text{imgEmb}(x_i^{pseimg}). \tag{4}$$

As to the image $x_i^{img}$, we obtain its pseudo text by image classification. Concretely, we input $x_i^{img}$ into GoogLeNet to perform image classification with 1,000 categories, where each category label signifies a short text. The predicted top-$l$ categories, *i.e.,* a set of short texts, are then concatenated as pseudo text $x_i^{psetxt}$. Afterwards, we get the pseudo text embedding $\mathbf{e}_i^{psetxt} \in \mathbb{R}^d$ via,

$$\mathbf{e}_i^{psetxt} = \text{textEmb}(x_i^{psetxt}). \tag{5}$$

**Contrastive learning.** The embeddings of actual and corresponding pseudo modalities, *i.e.,* $\mathbf{e}_i^{img}$ to $\mathbf{e}_i^{pseimg}$ (and $\mathbf{e}_i^{txt}$ to $\mathbf{e}_i^{psetxt}$), describe the same item and locate in the same semantic space. Naturally, we view them as positive pairs in contrastive learning to enhance image and text embeddings via,

$$\mathcal{L}_{con} = -\frac{\exp(\text{sim}(\mathbf{e}_i^{img}, \mathbf{e}_i^{pseimg}))}{\sum_{k=1}^{n} \exp(\text{sim}(\mathbf{e}_i^{img}, \mathbf{e}_k^{pseimg}))} \\ -\frac{\exp(\text{sim}(\mathbf{e}_i^{txt}, \mathbf{e}_i^{psetxt}))}{\sum_{k=1}^{n} \exp(\text{sim}(\mathbf{e}_i^{txt}, \mathbf{e}_k^{psetxt}))}, \tag{6}$$

where the $\text{sim}(\cdot)$ is cosine similarity. In the first term, for an item image ($\mathbf{e}_i^{img}$), we view its pseudo image embedding ($\mathbf{e}_i^{pseimg}$) referring similar semantics as positives, while regarding other items' pseudo image embeddings ($\mathbf{e}_k^{pseimg}$) containing different contents as negatives. With pushing the positives close while pulling negatives apart, the MMSBR can enhance image embeddings. The second term does the same for refining text embeddings. With rich knowledge about corresponding modalities, the used data generation models not only align the positive pairs in the
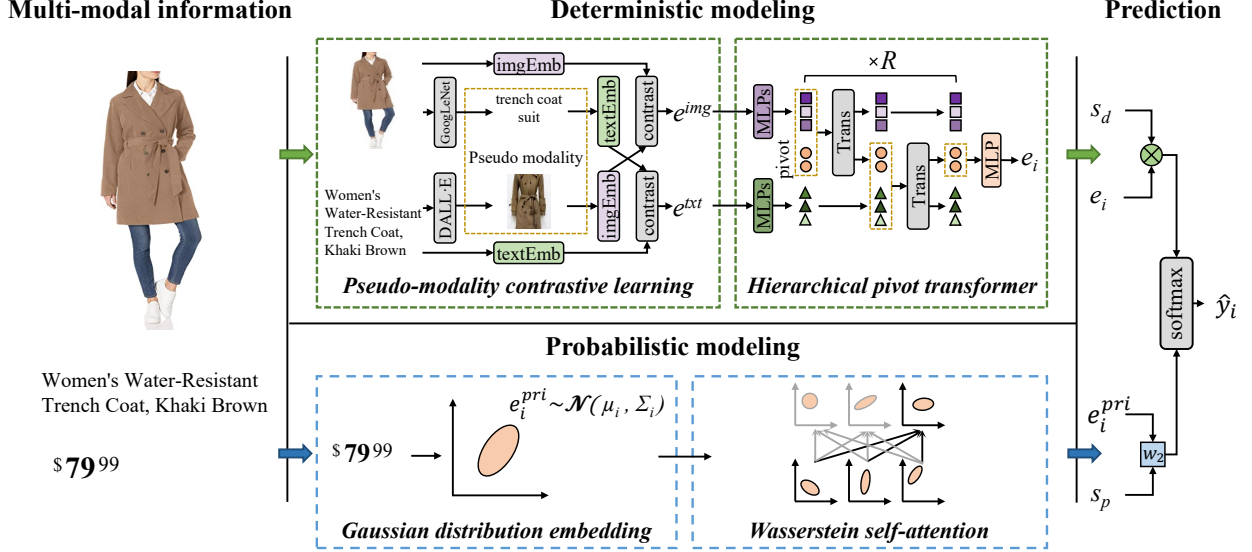
Fig. 2: The proposed MMSBR customizes deterministic and probabilistic modeling to handle descriptive and numerical information respectively. In deterministic modeling, a pseudo-modality contrastive learning is designed to enhance descriptive information representations, a hierarchical pivot transformer is presented to fuse heterogeneous descriptive information, and a vanilla attention is used to capture user deterministic taste. The probabilistic modeling represents item price with Gaussian distribution embedding and devises Wasserstein self-attention to model user acceptable price range. Finally, we predict user behaviors based on the multi-modal information.

same space but also make pseudo modality contain core semantics of the actual modality. As shown in Fig. 2, the pseudo image retains core contents cloth and filters out redundant pants and shoes. Obviously, this is of benefit to the pseudo-modality contrastive learning for alleviating noisy information existing in distinct modalities.

### 4.1.2 Hierarchical Pivot Transformer

As demonstrated early, we need to fuse image and text features for comprehensive user interest understanding. Transformer structure has shown great potential in merging multi-modal signals, given that it can effectively mine complex relations among tokens in a sequence [17], [19]. Inspired by this, we first apply several distinct MLPs to convert image/text embedding into different item feature embeddings and formulate feature sequence for image/text accordingly. Based on the feature sequences, a hierarchical pivot transformer is further proposed for effective descriptive information fusion.

**Image/Text features generation.** We apply MLP to obtain feature embeddings because many studies have demonstrated the effectiveness of MLP in capturing semantics of input data [18], [45]. Formally, an item image/text feature sequence ($\mathbf{Z}_{img}/\mathbf{Z}_{txt}$) is formulated via,

$$\mathbf{Z}_{img} = \{\text{MLP}_1^{img}(\mathbf{e}_i^{img}), ..., \text{MLP}_C^{img}(\mathbf{e}_i^{img})\}, \quad (7)$$
$$\mathbf{Z}_{txt} = \{\text{MLP}_1^{txt}(\mathbf{e}_i^{txt}), ..., \text{MLP}_C^{txt}(\mathbf{e}_i^{txt})\}, \quad (8)$$

where $\text{MLP}_k^{img}$ and $\text{MLP}_k^{txt}$ denote feed-forward neural networks with two hidden layers, and $C$ is the number of MLPs used for image/text features extracting. Note that, The $\text{MLP}_k^{img}(\mathbf{e}_i^{img})$ and $\text{MLP}_k^{txt}(\mathbf{e}_i^{txt}) \in \mathbb{R}^d$ are certain feature embeddings of image and text respectively.

**Hierarchical pivot transformer.** A vanilla transformer layer mainly contains three modules: Multi-head Self-Attention (MSA), Layer Normalisation (LN) and Fully Connected Layer (FCL). We can define a transformer layer with the input sequence $\mathbf{F}^l = [f_1^{in}, f_2^{in}, ..., f_k^{in}]$ and output sequence $\mathbf{F}^{l+1} = [f_1^{out}, f_2^{out}, ..., f_k^{out}]$ as $\mathbf{F}^{l+1} = \text{Trans}(\mathbf{F}^l)$ via,

$$\mathbf{F}_*^l = \text{MSA}(\text{LN}(\mathbf{F}^l)) + \mathbf{F}^l, \quad (9)$$
$$\mathbf{F}^{l+1} = \text{FCL}(\text{LN}(\mathbf{F}_*^l)) + \mathbf{F}_*^l. \quad (10)$$

Based on this, we further create a pivot $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_T]$ in each transformer layer to govern the fusion of multi-modal information, where $\mathbf{p}_i \in \mathbb{R}^d$ is a trainable token embedding used to assist information transmission. The hierarchical pivot transformer integrates the information of image ($\mathbf{Z}_{img}$) and text ($\mathbf{Z}_{txt}$) via:

$$[\mathbf{Z}_{img}^{l+1}, \mathbf{P}_{img}^l] = \text{Trans}([\mathbf{Z}_{img}^l, \mathbf{P}^l]), \quad (11)$$
$$\mathbf{p}_*^l = (\mathbf{P}_{img}^l + \mathbf{P}^l)/2, \quad (12)$$
$$[\mathbf{Z}_{txt}^{l+1}, \mathbf{P}_{txt}^l] = \text{Trans}([\mathbf{Z}_{txt}^l, \mathbf{P}_*^l]), \quad (13)$$
$$\mathbf{p}^{l+1} = (\mathbf{P}_{txt}^l + \mathbf{P}_*^l)/2, \quad (14)$$

where $\mathbf{P}^0 = \mathbf{P}$ (random initialization), $\mathbf{Z}_{img}^0 = \mathbf{Z}_{img}$ and $\mathbf{Z}_{txt}^0 = \mathbf{Z}_{txt}$. In each transformer layer, the pivot extracts and fuses important information from different modalities. Taking Eq. (13) as an example, the pivot absorbs text information and transmits image information to the text modality. To fully fuse descriptive information, we stack the hierarchical pivot transformer defined by Eqs. (11)-(14) $R$ times. Finally, the last layer pivot passed by a MLP is used to represent the descriptive information of an item $x_i$ as $\mathbf{e}_i \in \mathbb{R}^d$ via,

$$\mathbf{e}_i = \text{MLP}(\mathbf{P}^R) = \text{MLP}([\mathbf{p}_1^R; \mathbf{p}_2^R; ...; \mathbf{p}_T^R]), \quad (15)$$

where $[;]$ denotes the concatenation operation, and MLP is a feed-forward neural network with two hidden layers.

### 4.1.3 Vanilla Attention

For an item $x_i$, we have obtained its embedding $\mathbf{e}_i$ for descriptive information involving image and text. Apparently, a user deterministic taste hidden in items she has interacted with. Thus, based on item sequence with descriptive information $\mathbf{E}_d = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_m]$, we can apply the vanilla attention as used in [10], [14] to obtain user deterministic taste $\mathbf{s}_d \in \mathbb{R}^d$ via,

$$\mathbf{s}_d = \sum_{k=1}^{m} \alpha_k \mathbf{e}_k, \tag{16}$$

$$\alpha_k = \mathbf{u}\sigma(\mathbf{A}_1\mathbf{e}_k + \mathbf{A}_2\bar{\mathbf{e}} + \mathbf{b}), \tag{17}$$

where $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{d \times d}$ and $\mathbf{b}$ are learnable parameters, $\mathbf{u}^T \in \mathbb{R}^d$ is a trainable vector used to determine items' importance in the session, and $\bar{\mathbf{e}} = \frac{1}{m}\sum_{k=1}^{m}\mathbf{e}_k$.

## 4.2 Probabilistic Modeling

Probabilistic Modeling employs: (1) *Gaussian distribution embedding* to represent item price; (2) *Wasserstein self-attention* to model user acceptable price range.

### 4.2.1 Gaussian Distribution Embedding

As discussed before, user preferences on price present range instead of point-wise property. Therefore, we represent the price level $v_i^{pri}$ of an item $x_i$ with the Gaussian distribution via,

$$\hat{\mathbf{e}}_i^{pri} = \text{Gaussian}(v_i^{pri}) \sim \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i), \tag{18}$$

where $\hat{\mu}_i$ and $\hat{\Sigma}_i \in \mathbb{R}^d$ are mean and covariance vectors respectively. MMSBR learns them with two distinct lookup embedding tables based on item price level. Note that, the mean and covariance vectors collectively signify the price range where the item falls in. As indicated in [14], the user price preferences are related with item category, so we further incorporate category information to formulate price embedding $\mathbf{e}_i^{pri}$ for the item $x_i$ via,

$$\mathbf{e}_i^{pri} \sim \mathcal{N}(\mu_i, \Sigma_i) = \mathcal{N}(\hat{\mu}_i + \mathbf{e}_i^c, \hat{\Sigma}_i + \mathbf{e}_i^c), \tag{19}$$

where $\mathbf{e}_i^c \in \mathbb{R}^d$ is the category embedding of the item. It is noted that an item price is represented by Gaussian distribution instead of widely used point-wise vector embedding, which endows MMSBR with the ability to perceive range property of item price.

### 4.2.2 Wasserstein Self-attention

Self-attention is employed by various approaches [14], [17] to model behavior sequences due to its capacity in capturing item-item transition patterns. However, the conventional self-attention calculates similarity between point-wise vector embeddings with dot product, which is unsuitable for our settings where the price is represented by Gaussian distribution. Therefore, we devise a Wasserstein self-attention which applies Wasserstein distance [20], [21] to obtain attention scores between price distribution embeddings. Formally, the Wasserstein distance between two Gaussian distribution embeddings $\mathcal{G}_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{G}_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ is defined as,

$$\mathcal{W}_2(\mathcal{G}_1, \mathcal{G}_2) = \sqrt{\|\mu_1 - \mu_2\|_2^2 + \left\|(\Sigma_1)^{\frac{1}{2}} - (\Sigma_2)^{\frac{1}{2}}\right\|_2^2}. \tag{20}$$

Referring to conventional self-attention, Wasserstein self-attention (WSA) handles price sequence $\mathbf{E}_p = [\mathbf{e}_1^{pri}, \mathbf{e}_2^{pri}, ..., \mathbf{e}_m^{pri}]$ via,

$$\mathbf{H} = \text{WSA}(A^Q\mathbf{E}_p, A^K\mathbf{E}_p, A^V\mathbf{E}_p), \tag{21}$$

where $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_m\}$ is output and $A^* = (A_\mu^*, A_\Sigma^*)$ ($* \in \{Q, K, V\}$) is used to map each distribution in $\mathbf{E}_p$ into query, key and value spaces respectively. $A_\mu^*$ or $A_\Sigma^* \in \mathbb{R}^{d \times d}$ converts mean or covariance embeddings to corresponding spaces. Afterwards, the Wasserstein distance is employed to calculate the attention scores between query $A^Q\mathbf{e}_i^{pri}$ and key $A^K\mathbf{e}_j^{pri}$ via,

$$\begin{aligned} a_{ij} &= \mathcal{W}_2(A^Q\mathbf{e}_i^{pri}, A^K\mathbf{e}_j^{pri}) \\ &= \mathcal{W}_2(\mathcal{N}(A_\mu^Q\mu_i, A_\Sigma^Q\Sigma_i), \mathcal{N}(A_\mu^K\mu_j, A_\Sigma^K\Sigma_j)). \end{aligned} \tag{22}$$

We then linearly sum up each value $A^V\mathbf{e}_j^{pri} \sim \mathcal{N}(A_\mu^V\mu_j, A_\Sigma^V\Sigma_j)$ according to its attention scores to $i$-th item price $a_{ij}$ to obtain the $i$-th output $\mathbf{h}_i \sim \mathcal{N}(\mathbf{h}_i^\mu, \mathbf{h}_i^\Sigma)$ via,

$$\mathbf{h}_i^\mu = \sum_{j=1}^{m} a_{ij}A_\mu^V\mu_j, \text{ and } \mathbf{h}_i^\Sigma = \sum_{j=1}^{m} a_{ij}^2 A_\Sigma^V\Sigma_j. \tag{23}$$

Finally, the hidden state $\mathbf{h}_m$ is used to represent acceptable price range $\mathbf{s}_p$ for the user via,

$$\mathbf{s}_p = \mathbf{h}_m \sim \mathcal{N}(\mathbf{h}_m^\mu, \mathbf{h}_m^\Sigma). \tag{24}$$

## 4.3 Prediction

So far, for an item $x_i$, we have obtained its comprehensive representation $(\mathbf{e}_i, \mathbf{e}_i^{pri})$ based on its multi-modal information, where $\mathbf{e}_i$ is derived from descriptive information (image and text) and $\mathbf{e}_i^{pri} \sim \mathcal{N}(\mu_i, \Sigma_i)$ comes from numerical information (price). As to an anonymous user, $\mathbf{s}_d$ represents her deterministic taste on descriptive information, and $\mathbf{s}_p$ indicates her acceptable price range. Based on the entire multi-modal information displayed on pages, thus, we can infer the probability of the user clicking item $x_i$ via,

$$\hat{y}_i = softmax(\mathbf{e}_i\mathbf{s}_d + \mathcal{W}_2(\mathbf{e}_i^{pri}, \mathbf{s}_p)), \tag{25}$$

where we evaluate user deterministic behaviors with dot-product and user probabilistic behaviors with Wasserstein distance. As in [4], [10], [14], we employ cross-entropy to improve recommendation performance via:

$$\mathcal{L}_{rec} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i), \tag{26}$$

where $y_i$ is ground-truth label and $\hat{y}_i$ is predicted probability of item $x_i$ to be clicked. Finally, we train our MMSBR under the joint supervision of recommendation and contrastive learning via,

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda\mathcal{L}_{con}, \tag{27}$$

where the $\lambda$ is a constant controlling the strength of contrastive learning task.

TABLE 2: Statistics of all datasets.

| Datasets | Cellphones | Grocery | Sports |
|---|---|---|---|
| #item | 8,614 | 11,638 | 18,796 |
| #category | 48 | 665 | 1,259 |
| #interaction | 196,376 | 364,728 | 566,504 |
| #session | 78,026 | 127,548 | 211,959 |
| avg.length | 2.52 | 2.86 | 2.67 |

## 5 EXPERIMENTAL SETUP

### 5.1 Research Questions

We conduct extensive experiments to validate the effectiveness of MMSBR by answering following research questions:

- **RQ1** Does the proposed MMSBR achieve state-of-the-art performance? (ref. Section 6.1)
- **RQ2** What is the effect of various novel techniques proposed in MMSBR? (ref. Section 6.2-6.4)
- **RQ3** What is the performance of MMSBR under cold-start scenario? (ref. Section 6.5)
- **RQ4** How does session length influence the performance of SBR? (ref. Section 6.6)
- **RQ5** What is the influence of different modalities on the performance of SBR? (ref. Section 6.7)
- **RQ6** What is the influence of key hyperparameters on MMSBR? (ref. Section 6.8)

### 5.2 Datasets and Preprocessing

We evaluate our MMSBR and all baselines on three datasets covering different characteristics and domains from Amazon[1], *i.e.*, Cell Phones and Accessories (**Cellphones**), Grocery and Gourmet Food (**Grocery**), as well as Sports and Outdoors (**Sports**). Following [14], we organize user behaviors within one day to imitate SBR scenario. The last item in a session is taken as predicted target, and remaining items are used to model user intent. As in [4], [10], we filter out sessions whose length is 1 and items appearing less than 5 times. Also, we delete items with missing or invalid images/text. We chronologically split each dataset into three parts with the ratio of 7:2:1 for training, validation and testing respectively. Relying on item ID, existing models [4], [9], [13], [14], [47] can not handle cold-start items which do not appear in training sets, so they simply delete these items from test sets. Following their settings, we also remove the cold-start items, where datasets' statistics is shown in Table 2. Besides, we retain cold-start items to investigate the performance of MMSBR under cold-start scenario in Section 6.5, where the cold-start situation is reported in Table 5. Note that, although our setting is ubiquitous in real scenes, available datasets containing images, text and price are very scarce instead. Thus, we sincerely hope that our work can foster the development of multi-modal datasets for SBR.

### 5.3 Evaluation Metrics

As in [4], [10], [14], we evaluate the performance of MMSBR and baselines with following two widely used metrics: **Prec@k** (Precision) calculates the proportion of cases where the target item is within recommendation list; **MRR@k**

(Mean Reciprocal Rank) is the average of reciprocal ranks of target item among recommendation list. Similar as [13], [14], [47], [48], the k is set as 10 and 20 in this work.

### 5.4 Baselines

The following two groups of competitive methods are selected as baselines for performance comparison:

**Co-occurrence based methods** focus on mining item co-occurrence patterns to provide recommendation:

- **S-POP** recommends the most frequent items in the current session;
- **SKNN** predicts next items based on items' co-occurrence frequency in all sessions;
- **NARM** [4] utilizes GRU with attention mechanism to capture user main intent;
- **SASRec** [49] applies Transformer architecture to model transitions among items;
- **BERT4Rec** [9] employs bidirectional self-attention to model user behaviors;
- **SR-GNN** [10] captures complex relations among items via GNN;
- **COTREC** [47] enhances item embeddings by contrastive learning.
- **MSGIFSR** [48] studies fine-grained co-occurrence relations by dividing a session into multiple snippets.

**Side information enhanced methods** utilize extra information to facilitate user preferences learning:

- **MGS** [13] exploits item categories for more accurate preferences estimation;
- **UniSRec** [12] incorporates description text of items to obtain universal sequence representations;
- **CoHHN** [14] emphasizes the significance of price in determining user choices.

We have not included MML [16], which focuses on text and image-based long sequence learning, in our baselines. This decision was made because MML randomly deletes some items within a sequence during model training, which is unsuitable for SBR where a session typically consists of only a few items (as shown in Table 2).

### 5.5 Implementation Details

To ensure fair comparison, we fix embedding size of all methods at 64. The other hyperparameters of MMSBR and all baselines are determined via grid search according to their performance on Prec@20 in validation set. For main hyperparameters of MMSBR, we investigate the number of stacked layers for hierarchical pivot transformer $R$ in $\{1, 2, 3, 4, 5\}$, the number of generated features $C$ in $\{2, 4, 6, 8, 10\}$ and the number of tokens in pivot $T$ in $\{1, 2, 3, 4, 5, 6\}$. Besides, we fix balance coefficient $\lambda = 0.01$, retain top-2 ($l$=2) categories from image classification as pseudo text, and set the number of price levels as 100 ($\rho = 100$) for all datasets. The mini-batch size and initial learning rate is 100 and 0.001, respectively. Given that the output dimension of GoogLeNet and BERT are 1024 and 768 respectively, we utilize PCA algorithm to reduce them to 64. We have released the source code of our work[2].

TABLE 3: Performance comparison of MMSBR with baselines over three datasets. The results (%) produced by the best baseline and the best performer in each column are underlined and boldfaced respectively. Statistical significance of pairwise differences for MMSBR against the best baseline (*) is determined by the t-test ($p < 0.01$).

| Method | Cellphones | | | | Grocery | | | | Sports | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Prec@10 | MRR@10 | Prec@20 | MRR@20 | Prec@10 | MRR@10 | Prec@20 | MRR@20 | Prec@10 | MRR@10 | Prec@20 | MRR@20 |
| S-POP | 5.32 | 2.71 | 7.24 | 2.85 | 20.65 | 17.00 | 23.64 | 17.25 | 15.61 | 14.56 | 17.59 | 14.69 |
| SKNN | 21.07 | 9.95 | 24.71 | 10.21 | 39.83 | 25.15 | 41.88 | 25.29 | 31.79 | 21.31 | 33.86 | 21.46 |
| NARM | 20.59 | 15.32 | 24.12 | 15.56 | 40.39 | 34.53 | 42.41 | 34.62 | 31.64 | 26.94 | 34.17 | 27.12 |
| SASRec | 23.37 | 15.47 | 27.58 | 15.76 | 40.97 | 34.76 | 43.02 | 34.92 | 31.54 | 26.68 | 34.11 | 26.87 |
| BERT4Rec | 22.28 | 14.39 | 27.09 | 14.73 | 40.59 | 34.09 | 42.93 | 34.31 | 31.57 | 26.85 | 34.32 | 27.07 |
| SR-GNN | 21.80 | 15.60 | 25.08 | 15.77 | 40.81 | 34.89 | 42.74 | 35.01 | 31.96 | 27.43 | 34.29 | 27.51 |
| COTREC | 23.78 | 10.82 | 28.33 | 11.13 | 41.28 | 30.60 | 43.24 | 30.75 | 32.16 | 23.28 | 35.13 | 23.46 |
| MSGIFSR | 20.92 | 14.53 | 24.51 | 14.77 | 41.34 | 35.25 | 43.40 | 35.47 | 32.28 | 27.56 | 34.95 | 27.72 |
| MGS | 21.74 | 14.29 | 25.21 | 14.54 | 40.92 | 35.06 | 42.79 | 35.20 | 31.63 | 26.75 | 33.76 | 26.89 |
| UniSRec | 22.73 | 15.36 | 26.65 | 15.63 | 41.40 | 35.12 | 43.44 | 35.24 | 31.90 | 26.91 | 34.41 | 27.04 |
| CoHHN | 23.60 | 15.77 | 27.71 | 15.96 | 41.58 | 35.33 | 43.59 | 35.58 | 32.12 | 27.13 | 35.02 | 27.31 |
| **MMSBR** | **24.37*** | **16.47*** | **29.22*** | **16.81*** | **42.10*** | **35.91*** | **44.27*** | **36.06*** | **32.89*** | **28.10*** | **35.64*** | **28.28*** |

## 6 RESULTS AND ANALYSIS

### 6.1 Overall Performance (RQ1)

We report the performance of MMSBR and all baselines in Table 3, where the following observations are noted:

(1) Among co-occurrence based methods, COTREC and MSGIFSR achieve competitive performance. We speculate that COTREC's good performance comes from its utilization of contrastive learning to improve session embeddings. As to MSGIFSR, it divides a session into many snippets containing consecutive items, enabling it to capture fine-grained co-occurrence relations among items.

(2) For methods with side information enhancement, Co-HHN (price) and UniSRec (text) have obvious advantages over MGS (category). As opposed to category, price and text are what users can immediately observe on item pages. This observation supports our claim that modeling what displays on websites is of benefit to capturing user intent.

(3) Compared with co-occurrence based methods, the side information enhanced methods generally perform better. This signifies the validity of extra information in modeling user behaviors. It makes sense since that side information enables models to mine various user preferences, leading to effective intent understanding.

(4) Different baselines have varying performance on various datasets. Taking CoHHN as an example, it achieves the best performance on Grocery among all baselines, while its results on Sports left some to be desired. These methods just focus on a part of information that users may access, either item ID, category, text or price. In fact, however, a user evaluates all available information before making decisions. Therefore, they are incapable of capturing user preferences holistically, which results in their inferiority in discerning user intent across various context (*i.e.*, datasets).

(5) The proposed MMSBR consistently outperforms all baselines in terms of all evaluation metrics on all datasets, which demonstrates its effectiveness for SBR. In particular, MMSBR surpasses the best baselines in Prec@20 and MRR@20 by 3.14% and 5.33% on Cellphones, 1.56% and 1.35% on Grocery and 1.45% and 2.02% on Sports. Given
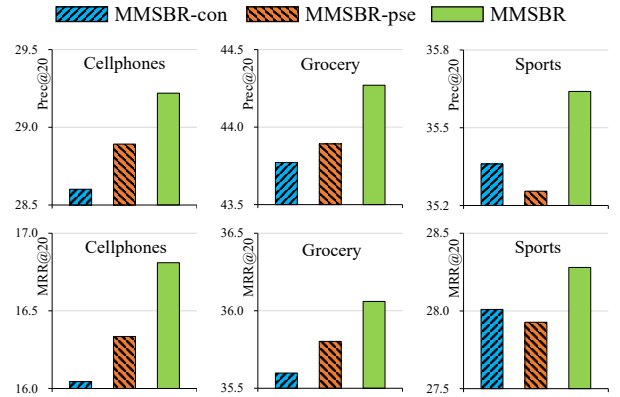


Fig. 3: The effect of pseudo-modality contrastive learning.

that a user makes decisions by evaluating item images, text and price, the modeling for entire multi-modal information in MMSBR is in line with decision process, contributing to revealing her intent more effectively. Besides, with reference to Table 2 and Table 3, we find that MMSBR obtains largest improvements in Cellphones that contains least items among all datasets. We argue that the introduction of multi-modal information enriches data and enables MMSBR to understand user demands from multiple perspectives. Therefore, the proposed MMSBR achieves impressive performance under the condition of sparsity data. It also reminds researchers that the multi-modal information is an antidote to copy with sparsity issue.

### 6.2 Effect of Pseudo-modality Contrastive Learning (RQ2)

To obtain relevant semantics from descriptive information under distinct noise, we propose a pseudo-modality contrastive learning. It refines image and text embeddings via contrastive learning, where generated pseudo modalities are used as contrastive signals. MMSBR-con removes contrastive learning from MMSBR, *i.e.*, it directly fuses outputs of different modality encoders without handling

TABLE 4: The effect of hierarchical pivot transformer.

| Method | Cellphones | | Grocery | | Sports | |
|---|---|---|---|---|---|---|
| | Prec@20 | MRR@20 | Prec@20 | MRR@20 | Prec@20 | MRR@20 |
| COTREC | 28.33 | 11.13 | 43.24 | 30.75 | 35.13 | 23.46 |
| MSGIFSR | 24.51 | 14.77 | 43.40 | 35.47 | 34.95 | 27.72 |
| $MMSBR_{mlp}$ | 26.74 | 15.95 | 42.93 | 35.28 | 34.67 | 27.86 |
| **MMSBR** | **29.22*** | **16.81*** | **44.27*** | **36.06*** | **35.64*** | **28.28*** |

distinct noise. MMSBR-pse projects embeddings of different modalities into a space via MLP and conducts contrastive learning accordingly like in [15], [18], [29], while ignoring the semantic gaps existing in distinct modalities.

As shown in Fig. 3, in Cellphones and Grocery, both MMSBR and MMSBR-pse outperform MMSBR-con, demonstrating that contrastive learning can enhance modality representation. Besides, MMSBR-pse is defeated by MMSBR-con on Sports. It proves our hypothesis that semantic gaps between distinct modalities may impede representation learning. Thus, directly contrasting different modalities of an item in turn leads to performance degradation in this case. Moreover, MMSBR performs much better than MMSBR-pse in all datasets, which indicates that generated pseudo modalities can fill such semantic gaps. Additionally, MMSBR achieves the best performance across all variants, which indicates the superiority of pseudo-modality contrastive learning on mitigating distinct noise existing in different modalities.

## 6.3 Effect of Hierarchical Pivot Transformer (RQ2)

A user usually makes the decision after evaluating shared and distinct information from descriptive information. Therefore, we propose a novel hierarchical pivot transformer for heterogeneous information fusion. Following conventional operations [18], [45], $MMSBR_{mlp}$ maps image and text into the same space by MLP and concatenates their embeddings to fuse item descriptive information.

As shown in Table 4, $MMSBR_{mlp}$ is defeated by MMSBR with a large margin, which indicates the effectiveness of hierarchical pivot transformer in capturing complementary information from images and text. We believe that the pivot in each transformer layer is able to extract and integrate meaningful information from distinct modalities, thus facilitating effective information fusion. Furthermore, $MMSBR_{mlp}$ achieves competitive performance (especially in MRR@20) compared with the best baselines MSGIFSR and COTREC. It serves as more evidence that modeling multimodal information rather than only mining co-occurrence patterns of item ID can assist to better user intent learning.

## 6.4 Effect of Probabilistic Modeling (RQ2)

As discussed previously, different from descriptive information where a user's taste is deterministic, the numerical information, *i.e.,* item price, affects user behaviors in a probabilistic way. Therefore, we propose a probabilistic modeling to handle this situation, where the Gaussian distribution and Wasserstein Self-attention are devised to represent item price and model user acceptable price range respectively. Following [14], the variant $MMSBR_{de}$ represents item price
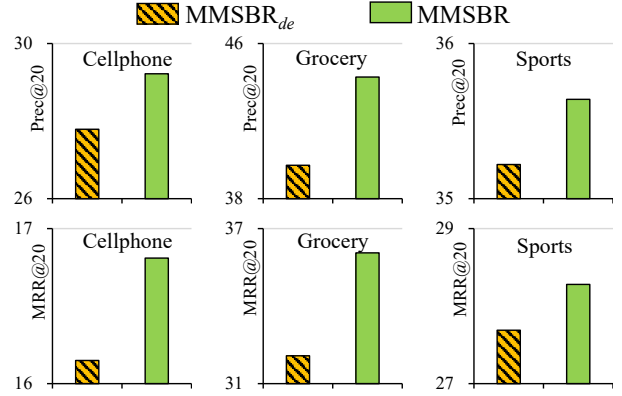


Fig. 4: The effect of probabilistic modeling.

TABLE 5: Statistics of datasets with cold-start items.

| Datasets | Cellphones+ | Grocery+ | Sports+ |
|---|---|---|---|
| #item | 10,245(+1631) | 13,493(+1855) | 22,049(+3253) |
| #category | 48(-) | 678(+13) | 1,312(+53) |
| #interaction | 199,065(+2689) | 367,674(+2946) | 571,789(+5285) |
| #session | 78,987(+961) | 128,510(+962) | 213,787(+1828) |
| avg.length | 2.52(-) | 2.86(-) | 2.67(-) |

with point-wise vector embeddings instead of distribution ones. Specifically, $MMSBR_{de}$ first discretizes continuous item price into discrete price-level, and then obtains point-wise embedding for the price via look-up embedding table. In other words, it does not discriminate distinct influence modes of various information on user choices.

As presented in Fig. 4, MMSBR significantly outperforms $MMSBR_{de}$ in all cases, confirming the validity of the proposed probabilistic modeling in tackling numerical information. Moreover, it demonstrates that users exhibit different behavioral patterns on different information, *i.e.,* deterministic/probabilistic mode on the descriptive/numerical information. By utilizing Gaussian distribution embeddings and Wasserstein self-attention, MMSBR is able to learn user acceptable price range, leading to its good performance on user behaviors modeling. In addition, distinguishing influence modes of different type information in a fine-grained manner is advantageous to user behaviors modeling, which is a valuable reference to future research.

## 6.5 Performance in Cold-start Scenario (RQ3)

Recommendation systems have long struggled with the cold-start problem, where they are required to show users new items that never appear in the system before. To evaluate the performance of MMSBR under cold-start scenario, we retain fresh items which do not appear in training sets in tests, where the statistics is presented in Table 5. We can get following insights from Fig. 5, :

(1) When encountering with new items, all models show a deteriorated performance, indicating that the cold-start is truly a tricky issue in SBR. Fortunately, the incorporation of extra information can aid in portrayal for new items, leading to impressive performance in cold-start situation. For instance, in Sports, COTREC/MSGIFSR based on co-occurrence patterns defeats CoHHN incorporating price
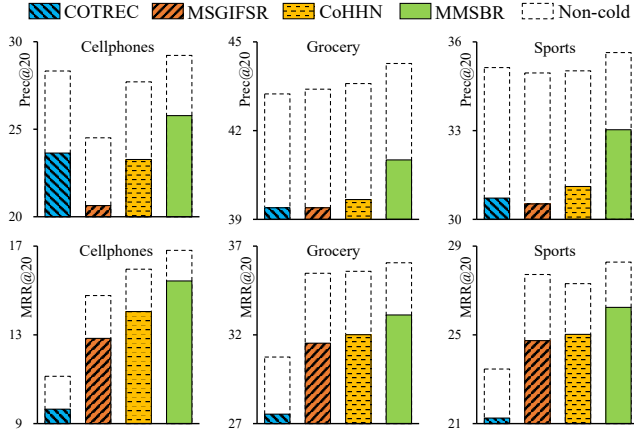
Fig. 5: Performance in cold-start scenario.



Fig. 6: Impact of various session lengths.

TABLE 6: The influence of different modalities.

| Method | Cellphones | | Grocery | | Sports | |
|---|---|---|---|---|---|---|
| | Prec@20 | MRR@20 | Prec@20 | MRR@20 | Prec@20 | MRR@20 |
| (a) w/o image | 27.45 | 14.85 | 41.23 | 35.20 | 32.14 | 27.50 |
| (b) w/o text | 27.19 | 14.69 | 41.11 | 35.08 | 32.22 | 27.42 |
| (c) w/o price | 25.10 | 13.35 | 42.98 | 35.57 | 34.78 | 27.68 |
| **MMSBR** | **29.22*** | **16.81*** | **44.27*** | **36.06*** | **35.64*** | **28.28*** |

and category on Prec@20/MRR@20 for non-cold items, whereas CoHHN outperforms them both in cold-start scenario. (2) The co-occurrence based methods are prone to fail in cold-start scenario, which is intuitive as there are not co-occurrence patterns for them to learn. Solely relying on co-occurrence patterns exposed by item ID, these methods could do nothing but blindly guess user interest in new items with random embeddings, resulting in their inferior performance. (3) The proposed MMSBR outperforms all methods with a large margin under cold-start scenario, indicating that MMSBR can effectively alleviate cold-start problem. Furthermore, our MMSBR has the least performance degradation compared with other methods in the cold start scenario. We believe that holistically modeling multi-modal information that a user can access enables MMSBR to mine her fine-grained preferences to the maximum, thus achieving impressive results. It also reminds researchers that utilizing multi-modal information is a promising way to copy with cold-start issue.

### 6.6 Impact of Various Session Lengths (RQ4)

The session length can significantly affect recommendation performance since it signifies how much information we can obtain to model user intent. Therefore, we investigate the performance of MMSBR under different session lengths. As shown in Fig. 6, following observations are noted:

(1) The proposed MMSBR achieves larger improvement over baselines on short sessions ($\leqslant 3$) than long sessions ($> 3$). Obviously, it is hard for co-occurrence based methods to accurately predict user behaviors within short sessions, since there is limited information for them to capture user intent. In contrast, our MMSBR can identify fine-grained preferences of users from rich multi-modal information, which alleviates data sparsity existing in short sessions. (2) Models perform better in short sessions than in long ones on Cellphones and Sports. Instead, they perform well in long sessions but poorly in short ones on Grocery. According to Table 2, sessions in Grocery are much longer than that in Cellphones and Sports. We speculate that much more instances concentrated in long sessions make models achieve better performance in long sessions on Grocery. (3) MMSBR achieves the best results in all cases, which demonstrates its effectiveness on modeling user behaviors in SBR again.
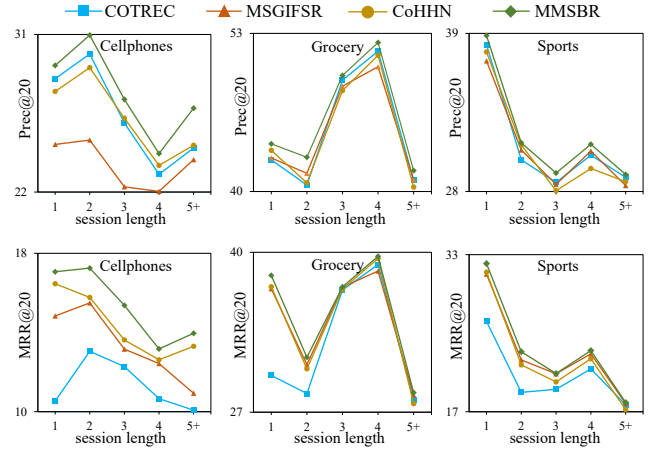
### 6.7 Ablation Study (RQ5)

In this part, we further zoom into each modality to see its specific influence on MMSBR. We successively remove each modality from MMSBR to conduct ablation study. Notably, in (a)/(b) of Table 6, the item image/text is only used to refine text/image in pseudo-modality contrastive learning while we do not include it to model user interest.

As shown in Table 6, different modalities show various influence on MMSBR's performance in distinct context. For instance, without price, (c) is overwhelmed by (a) and (b) in Cellphones, while its performance is better than (a) and (b) in other datasets. We speculate that, for electronics, users are concerned with its price because there may be a huge price gaps between cellphones with different brands. As to Grocery, users tend to care its practicality instead of price. Moreover, MMSBR achieves much better performance than all variants. It supports our motivation that a user behaviors are determined by the entire multi-modal information displayed on pages. Thus, it is rationale and imperative to model user preferences by considering these multi-modal information holistically.

### 6.8 Hyperparameter Study (RQ6)

In this section, we investigate the influence of three main hyperparamers on MMSBR.

**The number of stacked layers for hierarchical pivot transformer $R$.** From the first row in Fig. 7, we can find that the optimal $R$ for Cellphones/Grocery and Sports is 3 and 4 respectively. As shown in Table 2, Sports contains much more items than other datasets. We speculate that MMSBR needs to repeat hierarchical pivot transformer more times to fully integrate heterogeneous information in larger dataset.
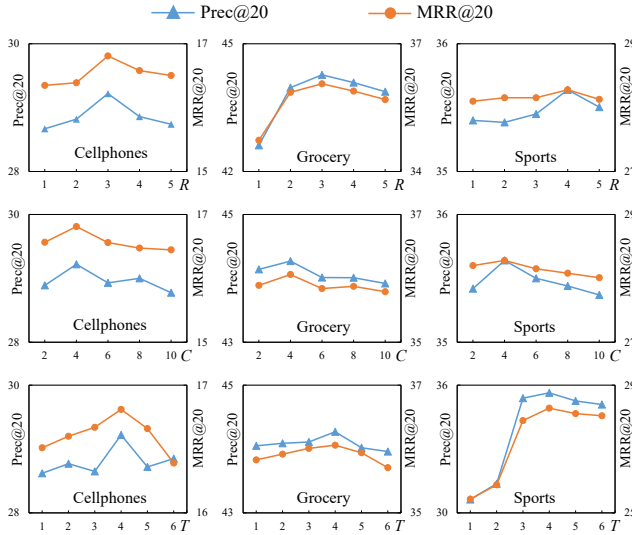
Fig. 7: Impact of hyperparameters.

**The number of generated features** $C$**.** As shown in middle row in Fig. 7, the small number of $C$, *i.e.*, 4, can make MMSBR achieve satisfactory results. It is consistent with cognitive anthropology that humans can only pay attention to a few aspects (features) of a matter (item) simultaneously.

**The number of tokens in pivot** $T$**.** Refer to the last row in Fig. 7, if $T$ is set too small, the pivot can not effectively extract information from different modalities. In contrast, if it is set too large, the information is sparsely distributed in each token, which is also adverse to information fusion. Accordingly, we empirically fix $T$ at 4 in all datasets.

## 7 CONCLUSION AND FUTURE WORK

Existing methods for session-based recommendation mostly concentrate on mining limited item co-occurrence patterns exposed by item ID, while ignoring that it is rich multi-modal information displayed on pages that attracts users to engage with certain items. Based on this motivation, we propose a novel MMSBR to characterize user preferences by modeling multi-modal information including descriptive information (images and text) and numerical information (price). Specifically, we devise a pseudo-modality contrastive learning to obtain relevant semantics of item images and text. Afterwards, a hierarchical pivot transformer is presented to effectively fuse heterogeneous descriptive information. For numerical information, we first represent item price with Gaussian distribution and devise a Wasserstein self-attention to model user acceptable price range. Comprehensive experiments conducted on three public datasets demonstrate the superiority of MMSBR over state-of-the-art baselines. Additional research also validates the effectiveness of MMSBR under cold-start scenario.

In the future, we plan to explore user reviews on items for further mining user fine-grained preferences for SBR. Besides, despite tailored for SBR, the proposed pseudo-modality contrastive learning and hierarchical pivot transformer can be easily extended to other multi-modal tasks for effective multi-modal learning.

## REFERENCES

[1] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *WWW*. ACM, 2001, pp. 285–295.

[2] Y. Koren, R. M. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, pp. 30–37, 2009.

[3] S. Wang, L. Cao, Y. Wang, Q. Z. Sheng, M. A. Orgun, and D. Lian, "A survey on session-based recommender systems," *ACM Comput. Surv.*, pp. 154:1–154:38, 2022.

[4] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma, "Neural attentive session-based recommendation," in *CIKM*, 2017, pp. 1419–1428.

[5] Q. Han, C. Zhang, R. Chen, R. Lai, H. Song, and L. Li, "Multi-faceted global item relation learning for session-based recommendation," in *SIGIR*, 2022, pp. 1705–1715.

[6] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *ICLR*, 2016.

[7] Q. Liu, Y. Zeng, R. Mokhosi, and H. Zhang, "STAMP: short-term attention/memory priority model for session-based recommendation," in *KDD*, 2018, pp. 1831–1839.

[8] F. Yuan, A. Karatzoglou, I. Arapakis, J. M. Jose, and X. He, "A simple convolutional generative network for next item recommendation," in *WSDM*, 2019, pp. 582–590.

[9] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *CIKM*, 2019, pp. 1441–1450.

[10] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, "Session-based recommendation with graph neural networks," in *AAAI*, 2019, pp. 346–353.

[11] T. Chen and R. C. Wong, "Handling information loss of graph neural networks for session-based recommendation," in *KDD*, 2020, pp. 1172–1180.

[12] Y. Hou, S. Mu, W. X. Zhao, Y. Li, B. Ding, and J. Wen, "Towards universal sequence representation learning for recommender systems," in *KDD*, 2022, pp. 585–593.

[13] S. Lai, E. Meng, F. Zhang, C. Li, B. Wang, and A. Sun, "An attribute-driven mirror graph network for session-based recommendation," in *SIGIR*, 2022, pp. 1674–1683.

[14] X. Zhang, B. Xu, L. Yang, C. Li, F. Ma, H. Liu, and H. Lin, "Price DOES matter!: Modeling price and interest preferences in session-based recommendation," in *SIGIR*, 2022, pp. 1684–1693.

[15] Y. Wang, H. Zhang, Z. Liu, L. Yang, and P. S. Yu, "Contrastvae: Contrastive variational autoencoder for sequential recommendation," in *CIKM*, 2022, pp. 2056–2066.

[16] X. Pan, Y. Chen, C. Tian, Z. Lin, J. Wang, H. Hu, and W. X. Zhao, "Multimodal meta-learning for cold-start sequential recommendation," in *CIKM*, 2022, pp. 3421–3430.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[18] Z. Liu, Y. Ma, M. Schubert, Y. Ouyang, and Z. Xiong, "Multi-modal contrastive pre-training for recommendation," in *ICMR*, 2022, pp. 99–108.

[19] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *NeurIPS*, 2021, pp. 14 200–14 213.

[20] C. Ma, L. Ma, Y. Zhang, R. Tang, X. Liu, and M. Coates, "Probabilistic metric learning with adaptive margin for top-k recommendation," in *KDD*, 2020, pp. 1036–1044.

[21] Z. Fan, Z. Liu, Y. Wang, A. Wang, Z. Nazari, L. Zheng, H. Peng, and P. S. Yu, "Sequential recommendation via stochastic self-attention," in *WWW*, 2022, pp. 2036–2047.

[22] X. Zhang, H. Lin, B. Xu, C. Li, Y. Lin, H. Liu, and F. Ma, "Dynamic intent-aware iterative denoising network for session-based recommendation," *Inf. Process. Manag.*, vol. 59, p. 102936, 2022.

[23] X. Zhang, H. Lin, L. Yang, B. Xu, Y. Diao, and L. Ren, "Dual part-pooling attentive networks for session-based recommendation," *Neurocomputing*, vol. 440, pp. 89–100, 2021.

[24] M. Wang, P. Ren, L. Mei, Z. Chen, J. Ma, and M. de Rijke, "A collaborative session-based recommendation approach with parallel memory modules," in *SIGIR*, 2019, pp. 345–354.

[25] L. Guo, H. Yin, Q. Wang, T. Chen, A. Zhou, and N. Q. V. Hung, "Streaming session-based recommendation," in *KDD*, 2019, pp. 1569–1577.

[26] J. Yuan, W. Ji, D. Zhang, J. Pan, and X. Wang, "Micro-behavior encoding for session-based recommendation," in *ICDE*, 2022, pp. 2886–2899.

[27] P. Zhang, J. Guo, C. Li, Y. Xie, J. Kim, Y. Zhang, X. Xie, H. Wang, and S. Kim, "Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network," in *WSDM*, 2023, pp. 168–176.

[28] W. Wang, W. Zhang, S. Liu, Q. Liu, B. Zhang, L. Lin, and H. Zha, "Incorporating link prediction into multi-relational item graph modeling for session-based recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2683–2696, 2023.

[29] X. Li, A. Sun, M. Zhao, J. Yu, K. Zhu, D. Jin, M. Yu, and R. Yu, "Multi-intention oriented contrastive learning for sequential recommendation," in *WSDM*, 2023, pp. 411–419.

[30] W. Ye, S. Wang, X. Chen, X. Wang, Z. Qin, and D. Yin, "Time matters: Sequential recommendation with complex temporal information," in *SIGIR*, 2020, pp. 3940–3946.

[31] Y. Xie, P. Zhou, and S. Kim, "Decoupled side information fusion for sequential recommendation," in *SIGIR*, 2022, pp. 1611–1621.

[32] J. Wu, R. Cai, and H. Wang, "Déjà vu: A contextualized temporal attention mechanism for sequential recommendation," in *WWW*, 2020, pp. 2199–2209.

[33] Y. Liu, Z. Ren, W. Zhang, W. Che, T. Liu, and D. Yin, "Keywords generation improves e-commerce session-based recommendation," in *WWW*, 2020, pp. 1604–1614.

[34] A. Rashed, S. Elsayed, and L. Schmidt-Thieme, "Context and attribute-aware sequential recommendation via cross-attention," in *RecSys*, 2022, pp. 71–80.

[35] Q. Cui, S. Wu, Q. Liu, W. Zhong, and L. Wang, "MV-RNN: A multi-view recurrent neural network for sequential recommendation," *IEEE Trans. Knowl. Data Eng.*, pp. 317–331, 2020.

[36] G. de Souza Pereira Moreira, S. Rabhi, R. Ak, M. Y. Kabir, and E. Oldridge, "Transformers with multi-modal features and post-fusion context for e-commerce session-based recommendation," *CoRR*, vol. abs/2107.05124, 2021.

[37] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 423–443, 2019.

[38] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T. Chua, "MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video," in *MM*, 2019, pp. 1437–1445.

[39] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie, "Dualgnn: Dual graph neural network for multimedia recommendation," *IEEE Trans. Multim.*, vol. 25, pp. 1074–1084, 2023.

[40] T. Han, P. Wang, S. Niu, and C. Li, "Modality matches modality: Pretraining modality-disentangled item representations for recommendation," in *WWW*, 2022, pp. 2058–2066.

[41] Y. Liu, S. Yang, C. Lei, G. Wang, H. Tang, J. Zhang, A. Sun, and C. Miao, "Pre-training graph transformer with multimodal side information for recommendation," in *MM*, 2021, pp. 2853–2861.

[42] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, and F. Jiang, "Bootstrap latent representations for multi-modal recommendation," in *WWW*, 2023, pp. 845–854.

[43] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[45] X. Song, L. Jing, D. Lin, Z. Zhao, H. Chen, and L. Nie, "V2P: vision-to-prompt based multi-modal product summary generation," in *SIGIR*, 2022, pp. 992–1001.

[46] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, 2021, pp. 8821–8831.

[47] X. Xia, H. Yin, J. Yu, Y. Shao, and L. Cui, "Self-supervised graph co-training for session-based recommendation," in *CIKM*, 2021, pp. 2180–2190.

[48] J. Guo, Y. Yang, X. Song, Y. Zhang, Y. Wang, J. Bai, and Y. Zhang, "Learning multi-granularity consecutive user intent unit for session-based recommendation," in *WSDM*, 2022, pp. 343–352.

[49] W. Kang and J. J. McAuley, "Self-attentive sequential recommendation," in *ICDM*, 2018, pp. 197–206.

[50] L. Guo, J. Zhang, T. Chen, X. Wang, and H. Yin, "Reinforcement learning-enhanced shared-account cross-domain sequential recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 7397–7411, 2023.
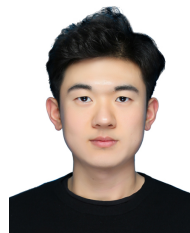
**Xiaokun Zhang** is currently pursuing the PhD degree with the School of Computer Science and Technology, Dalian University of Technology, China. His research interests include data mining and information retrieval, mainly focusing on intelligent recommender systems.



**Bo Xu** received the BSc and PhD degrees from the Dalian University of Technology, China, in 2011 and 2018. He is currently an associate professor in School of Computer Science and Technology of Dalian University of Technology. His current research interests include information retrieval and natural language processing.



**Fenglong Ma** is an assistant professor in the College of Information Sciences and Technology at the Pennsylvania State University. He received his Ph.D. from the Department of Computer Science and Engineering, University at Buffalo (UB) in 2019. His research interests lie in data mining and machine learning, with an emphasis on mining health-related data. His research interests also include natural language processing, social network mining and security.



**Chenliang Li** is a full Professor with School of Cyber Science and Engineering, Wuhan University China. His research areas include Information Retrieval, Recommendation System, Natural Language Processing and Social Computing. He has published over 80 papers in leading conferences and journals, and was the recipient of SIGIR 2016 Best Student Paper Award Honorable Mention. Currently, He serves as an Associate Editor for ACM TOIS, ACM TALLIP, and an editorial board member of JASIST and IPM. He has been a PC member of many leading conference, such as SIGIR, WWW, ACL, WSDM, AAAI, CIKM.



**Liang Yang** received the BSc and PhD degrees from the Dalian University of Technology, China, in 2009 and 2017, respectively. He is currently a lecturer in School of computer science and technology at the Dalian University of Technology. His current research interests include sentiment analysis and text mining.



**Hongfei Lin** received the BSc degree from the Northeastern Normal University in 1983, the MSc degree from the Dalian University of Technology in 1992, and the PhD degree from the Northeastern University in 2000. He is currently a professor in School of Computer Science and Technology at the Dalian University of Technology. He has published more than 500 research papers in various journals, conferences, and books. His research interests include information retrieval, text mining for biomedical literatures, biomedical hypothesis generation, information extraction from huge biomedical resources, learning-to-rank. He is the director of Information Retrieval Lab. at Dalian University of Technology.