

Intent-aware Ranking Ensemble for Personalized Recommendation

Jiayu Li
jy-li20@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing, China

Peijie Sun
sun.hfut@gmail.com
DCST, Tsinghua University
Beijing, China

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

Weizhi Ma
mawz@tsinghua.edu.cn
AIR, Tsinghua University
Beijing Academy of Artificial
Intelligence
Beijing, China

Yangkun Li
liyangkun17@gmail.com
DCST, Tsinghua University
Beijing, China

Min Zhang*
z-m@tsinghua.edu.cn
DCST, Tsinghua University
BNRist
Beijing, China

Zhoutian Feng
fengzhoutian@hotmail.com
Meituan Inc.
Beijing, China

Daiyue Xue
xuedaiyue@meituan.com
Meituan Inc.
Beijing, China

ABSTRACT

Ranking ensemble is a critical component in real recommender systems. When a user visits a platform, the system will prepare several item lists, each of which is generally from a single behavior objective recommendation model. As multiple behavior intents, e.g., both clicking and buying some specific item category, are commonly concurrent in a user visit, it is necessary to integrate multiple single-objective ranking lists into one. However, previous work on rank aggregation mainly focused on fusing homogeneous item lists with the same objective while ignoring ensemble of heterogeneous lists ranked with different objectives with various user intents.

In this paper, we treat a user's possible behaviors and the potential interacting item categories as the user's intent. And we aim to study how to fuse candidate item lists generated from different objectives aware of user intents. To address such a task, we propose an Intent-aware ranking Ensemble Learning (IntEL) model to fuse multiple single-objective item lists with various user intents, in which item-level personalized weights are learned. Furthermore, we theoretically prove the effectiveness of IntEL with point-wise, pair-wise, and list-wise loss functions via error-ambiguity decomposition. Experiments on two large-scale real-world datasets also show significant improvements of IntEL on multiple behavior objectives simultaneously compared to previous ranking ensemble models.

*Min Zhang is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).
SIGIR '23, July 23–27, 2023, Taipei, Taiwan.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9408-6/23/07...\$15.00
<https://doi.org/10.1145/XXXXXX.XXXXXX>

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Personalization*.

KEYWORDS

Ranking ensemble, User intents, Personalized recommendation

ACM Reference Format:

Jiayu Li, Peijie Sun, Zhefan Wang, Weizhi Ma, Yangkun Li, Min Zhang, Zhoutian Feng, and Daiyue Xue. 2023. Intent-aware Ranking Ensemble for Personalized Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, July 23–27, 2023, Taipei, Taiwan. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/XXXXXX.XXXXXX>

1 INTRODUCTION

Users typically have various intents when using recommender systems. For instance, when shopping online, users may intend to buy snacks or browse clothes. Generally, we call the users' behaviors the behavior intents and their interacted item categories the item category intents. Multiple behavior intents may be concurrent in a visit, and users need distinct items with different item category intents. Therefore, user intents are essential to recommender systems for recommendation list generation. In this paper, we follow the definition of user intents by Chen et al. [10] as a combination of user behavior and item category, such as booking an item with a hotel category or clicking an item in a phone category.

From the systems' viewpoint, since users usually have diverse intents, multiple item lists will be generated when a user visits the platform. These lists generally come from recommendation models optimized with different behavior objectives, such as clicking, consuming, or browsing duration. Existing research has made promising achievements with a single objective, such as predicting Click Through Rate (CTR) [4, 20, 48] and Conversion Rate (CVR) [14, 32].

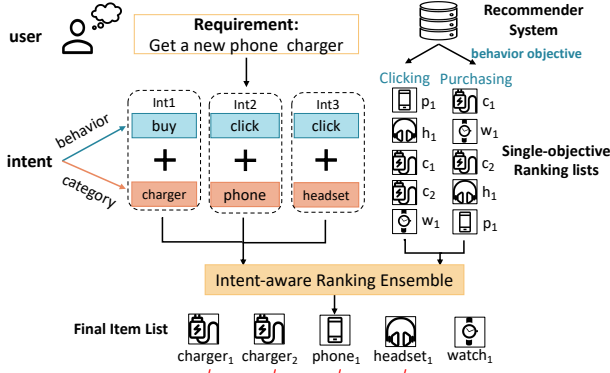


Figure 1: An example of fusing two single-objective item lists into a final list aware of multiple concurrent user intents.

However, as multiple intents of a user may appear in a visit, it is crucial to aggregate multiple heterogeneous single-objective ranking lists aware of the user’s current intents.

An example of the intent-aware ranking ensemble on an online shopping platform is shown in Figure 1. A user needs to buy a phone charger, and she also wants to browse new products about phones and headsets. The system has two single-objective ranking lists ready when she visits the platform. These lists are produced by two recommendation models optimized with users’ consumption and clicking histories, respectively. To satisfy the user’s diverse intents at once, an intent-aware ranking ensemble model is adopted to aggregate two ranking lists for a final display, where items are re-ordered according to both basic ranking lists and the user’s intents. Thus, charger_1 , charger_2 , phone_1 , and headset_1 are placed at the front of the final list, satisfying users’ preference better than both single-objective ranking lists. Therefore, intent-aware ranking ensemble is important for promoting recommendation performance.

However, there have been few attempts to combine heterogeneous single-objective ranking lists (Hereinafter referred to as basic lists) considering user intents. In industry, a common strategy is simply summing basic lists with pre-defined list-level weights, which ignores users’ personalized preference. While in academia, existing studies are not adequate to handle ranking ensemble for personalized recommendation. Widely-explored unsupervised rank aggregation methods [3, 21, 23] are mostly studied in information retrieval tasks rather than recommendation scenario. Recently, supervised methods [1, 2, 30] have been proposed to combine different item lists in recommendation. Nevertheless, these studies focused on combining homogeneous item lists optimized for the same behavior, not the heterogeneous rank lists for different objectives. Users’ intents are also overlooked in the ranking ensemble stage.

To aggregate basic lists aware of user intents, we aim to learn different weights for different basic lists and item categories to sum basic lists’ scores. However, it is challenging since numerous weights should be assigned for all items in all basic lists, which may be hard to learn. Therefore, we first prove its effectiveness theoretically. Unlike previous studies, we aim to assign ensemble weights at item level rather than list level. We prove the effectiveness of this form of ranking ensemble and verify that the loss of the ensemble list can be smaller than the loss of any basic models with

point-wise, pair-wise, and list-wise loss functions. An ambiguity term is derived from the proof and used for optimization loss.

With theoretical guarantees, another challenge in practice is to infer users’ intents and integrate the intents into ranking ensemble of heterogeneous basic lists. To address this challenge, we propose an Intent-aware ranking Ensemble Learning (IntEL) method for personalized ensemble of multiple single-objective ranking lists adaptively. A sequential model is adopted to predict users’ intents. And a ranking ensemble module is designed to integrate basic list scores, item categories, and user intents. Thus, the learnable ranking ensemble model can adaptively adjust the integration of multiple heterogeneous lists with user intents.

We conducted experiments on a public-available online shopping recommendation dataset and a local life service dataset. Our method, IntEL, is compared with various ensemble learning baselines and shows significant improvements. The main contributions of this work are as follows:

- To our knowledge, it is the first work that aims to generalize ranking ensemble learning with item-level weights on multiple heterogeneous item lists. We theoretically prove the effectiveness of ranking ensemble in this new setting.
- A novel intent-aware ranking ensemble learning model, IntEL, is proposed to fuse multiple single-objective recommendation lists aware of user intents adaptively. In the model, ambiguity loss, ranking loss, and intent loss have been proposed and integrated.
- Experiments on two large-scale real-world recommendation datasets indicate IntEL is significantly superior to previous ranking ensemble models on multiple objectives.

2 RELATED WORK

2.1 Ranking Ensemble

Ranking ensemble, i.e., fusing multiple ranking lists for a consensus list, has been long discussed in IR scenario [17, 19] and proved to be NP-hard even with small collections of basic lists [16].

In general, rank aggregation includes unsupervised and supervised methods. Unsupervised methods only rely on the rankings. For instance, Borda Count [5] computed the sum of all ranks. MRA [18] adopted the median of rankings. Comparisons among basic ranks were also used, such as pair-wise similarity in Out-rank [19] and distance from null ranks in RRA [21]. Recently, researchers have paid attention to supervised rank aggregation methods. For example, the Evolutionary Rank Aggregation (ERA) [30] was optimized with genetic programming. Differential Evolution algorithm [1, 2] and reinforcement learning [47] were also adopted for rank aggregation optimization. However, these rank aggregation methods only utilized the rank or scores of items in basic lists without considering item contents and users in recommendation.

Another view on the fusion problem comes from ensemble learning. It is a traditional topic in machine learning [35], which has been successfully applied to various tasks [29, 39, 42]. A basic theory in ensemble learning is error-ambiguity (EA) decomposition analysis [22], which proves better performance can be achieved with aggregated results with good and diverse basic models. It was proved in classification and regression with diverse loss functions [6, 45]. Liu et al. [24] generalized EA decomposition to model-level weights

in ranking ensemble with list-wise loss, where different items in a list shared the same weights.

The differences between the previous studies and our method are mainly twofold: First, rather than calculate a general weight for each basic model, we extend to assign item-level weights considering item category and user behavior intents. We theoretically prove the effectiveness of this extension. Second, we aim to combine heterogeneous lists generated for different behavior objectives and simultaneously improve performance on multiple objectives.

2.2 Multi-Intent Recommendation

Since we aggregate ranking lists aware of users' multiple intents, we briefly introduce recent methods on multi-intents and multi-interests in recommender systems. Existing studies focused on capturing dynamic intents in the sequential recommendation [10, 25, 26, 40, 43]. For instance, AIR [10] predicted intents and their migration in users' historical interactions. Wang et al. [40] modeled users' dynamic implicit intention at the item level to capture item relationships. MIND [34] and ComiRec [9] adopted dynamic routing from historical interactions to capture users' multi-intents and diversity. TimiRec [41] distilled target user interest from predicted distribution on multi-interest of the users. With the development of contrastive learning, implicit intent representations were also applied as constraints on contrastive loss [11, 15].

Previous studies usually mixed "intent" and "interest" and paid attention to intent on item contents in single-behavior scenarios. However, we follow [10] to consider both behavior intents and item category intents. Moreover, instead of learning user preference for each intent, we utilize intents as guidance for fusing user preference with different behavior objectives.

2.3 Multi-Objective Recommendation

Another brunch of related but different work is the multi-objective recommendation. It mainly contains two groups of studies. One group provides multiple recommendation lists for different objectives with shared information among objectives, such as MMOE [28] and PLE [38], where different lists are evaluated on corresponding objectives separately. The other group tried to promote the model performance on a target behavior objective with the help of other objectives, such as MB-STR [46] predicting users' click preferences. However, instead of generating multiple lists or specifying a target behavior, we fuse a uniform list on which multiple objectives are evaluated simultaneously.

Some studies that tried to jointly optimize ranking accuracy and other goals are also called multi-objective recommendation, such as fairness [44], diversity [8], etc. They sought to promote other metrics while maintaining utility on some behavior. But we aim to concurrently promote performance on multiple objectives by aggregating various recommendation lists.

3 PRELIMINARIES

3.1 Ranking Ensemble Learning Definition

Let $\mathcal{F} = \{f^1, f^2, \dots, f^K\}$ be K basic models that are trained for K different objectives (such as click, buy, and favorite, etc.), $\mathcal{I}(u, c) = \{i_1, i_2, \dots, i_N\}$ be the union set of K recommended basic item lists for user u in session environment context c (e.g., time and location),

Table 1: Notations. u and c denote user and context, respectively.

Notation	Description
$\mathcal{F} = \{f^1, \dots, f^K\}$	Set of K basic models.
$\mathcal{I}(u, c)$	The union set of items generated with \mathcal{F} .
$S_n^k(u, c)$	Predicted score from basic model k on item n .
$z_{mn}^k(u, c)$	The difference between scores $S_n^k(u, c) - S_m^k(u, c)$.
$w_n^k(u, c)$	Ensemble weight of item n in basic model k .
$S_n^{ens}(u, c)$	The final ensemble score of item n .
$\pi_n(u, c)$	Ground truth ranking of item n .
$\pi^{u,c}$	List of ground truth, $\pi^{u,c} = \{\pi_1(u, c), \dots, \pi_N(u, c)\}$.
Int	Distribution of user intent.
l_m, l_b, l_{p-l}	Point-wise, pair-wise and list-wise loss function.
A	The ambiguity term in ensemble learning loss.

and $S_n^k(u, c) = f^k(i_n, u, c)$ be the predicted score given by basic model k on item n . The goal of ranking ensemble learning is to learn a weighted ensemble score $S_n^{ens}(u, c)$ for each item i_n in \mathcal{I} ,

$$S_n^{ens}(u, c) = \sum_{k=1}^K w_n^k(u, c) \cdot S_n^k(u, c) \quad (1)$$

Where $w_n^k(u, c) \in \mathbb{R}$ denotes the weight of the k -th basic model for item i_n . The weights are learnable with the help of side information, e.g., user intents and item categories. The items in \mathcal{I} are sorted according to S_n^{ens} , and are compared with a ground truth order of ranking $\pi^{u,c} = \{\pi_1, \pi_2, \dots, \pi_N\}$, which is sorted to users' interactions with a pre-defined priority of user feedback, e.g., Buy>Click>Examine. The priority can be defined by business realities and will not influence the model learning strategy. The definition of ranking ensemble learning is similar to previous work [1, 24, 31], except that we conduct ensemble on heterogeneous basic models with different objectives, which makes the problem more difficult. The main notations are shown in Table 1.

3.2 User Intent Definition

When aggregating basic models optimized with different objectives, users' intent about behaviors and item categories are both essential. Therefore, we define a user's intent in a visit as a probability distribution of item categories and behaviors,

$$Int \sim P_{int}(I, B), \sum_{I \times B} P_{int}(I, B) = 1 \quad (2)$$

Where I and B indicate the item category intents and behavior intents, respectively. The types of categories and behaviors vary with recommendation scenarios. For instance, in online shopping, I can be product class, and B may include clicking and buying. In music recommender systems, I may be music genre, while B can contain listening and purchasing albums. In experiments, user intents Int are predicted from users' historical interactions and environment context.

3.3 Ranking Losses

Three representative losses are generally leveraged in the recommendation scenario, namely point-wise, pair-wise, and list-wise loss. We will theoretically and empirically illustrate the effectiveness of ranking ensemble with three losses in the following sections.

For a given user u under session context c with multi-level ground truth ranking $\pi^{u,c} = \{\pi_1, \pi_2, \dots, \pi_N\}$ on an item set $\mathcal{I} =$

$\{i_1, i_2, \dots, i_N\}$, the loss of score list $S(u, c) = \{S_1, S_2, \dots, S_N\}$ is defined as (u and c are omitted):

- **Point-wise Loss** As π is a multi-level feedback based on a group of user feedback, the Mean Squared Error (MSE) loss is utilized as a representative point-wise loss,

$$l_m(\pi, S) = \frac{1}{N} \sum_{n=1}^N l_m(\pi_n, S_n) := \frac{1}{N} \sum_{n=1}^N (S_n - \pi_n)^2 \quad (3)$$

- **Pair-wise Loss** We leverage the Bayesian Personalized Ranking (BPR) loss [33]. Following the negative sampling strategy for multi-level recommendation [27], a random item from one level lower is paired with a positive item at each level,

$$l_b(\pi, S) := \frac{1}{N^+} \sum_{l=1}^L \sum_{n, m \in I_l^+, I_l^-} l_b(S_n, S_m) \quad (4)$$

$$l_b(S_n, S_m) = -\log \sigma(S_n - S_m)$$

Where L is the number of interaction levels (e.g., buy, click, and exposure), N^+ is the number of positive items of all levels, I_l^+ and I_l^- are positive and one-level-lower negative item set for level l , and σ is the sigmoid function.

- **List-wise Loss** Following [24], we adopt the Plackett-Luce (P-L) model as the likelihood function of ranking predictions,

$$P_{p-l}(\pi|S) = \frac{1}{N} \prod_{n=1}^N \frac{\exp(S_{\pi_n})}{\sum_{m=n}^N \exp(S_{\pi_m})} \quad (5)$$

Where π_n indicates the n -th item sorted by ground truth π . The corresponding list-wise loss function is

$$l_{p-l}(\pi, S) := -\log[P_{p-l}(\pi|S)] \quad (6)$$

4 THEORETICAL EFFECTIVENESS OF RANKING ENSEMBLE LEARNING

To prove the effectiveness of our proposed item-level ranking ensemble learning in Eq.1, we aim to prove that the loss of ensemble learning scores $S^{ens} = \{S_n^{ens}\}$ can be smaller than any of the loss of basic-model scores $S^k = \{S_n^k\}$ for point-wise, pair-wise, and list-wise loss, i.e. $l(\pi, S^{ens}) \leq \sum_{k=1}^K w^k l(\pi, S^k)$, $\forall w^k, l \in \{l_m, l_b, l_{p-l}\}$. In this way, we can claim that there exist some combinations of weights w^k to achieve results better than all basic models.

Inspired by previous studies in ensemble learning, error ambiguity (EA) decomposition [22] provides an upper bound for ensemble loss $l(\pi, S^{ens})$, which helps conduct the above proof. For basic lists with loss $\{l(\pi, S^k)\}$, EA decomposition tries to split ensemble loss $l(\pi, S^{ens})$ into a weighted sum of basic-model loss ($\sum_k w_k l(\pi, S^k)$, $\forall w_k$) minus a positive ambiguity term A^1 of basic models, so that the upper bound of $l(\pi, S^{ens})$ is controlled by both basic-model losses and ambiguity. It was recently proved in ranking tasks with the same weights for a basic list (i.e., $w_n^k = w_m^k, \forall n = m$) [24]. However, different weights should be assigned for different items in our setting. Therefore, we need to verify whether EA decomposition is still available. To summarize, we try to prove that loss functions can be rewritten as $l(\pi, S^{ens}) \leq \sum_k \sum_n w_n^k l_n^k(\pi, S_n^k) - A$, $\forall w_n^k$ for point-wise, pair-wise, list-wise loss in the following.

¹The ambiguity A is sometimes called the *diversity* in EA decomposition. We use *ambiguity* to denote it to avoid confusion with the term *item diversity* in recommendation.

4.1 Point-wise Loss

THEOREM 1 (GENERALIZED EA DECOMPOSITION THEORY FOR POINT-WISE LOSS). *Given a set of score lists $\{S_n^k | k \in \{1, 2, \dots, K\}, n \in \{1, \dots, N\}\}$ from K basic models on N items, and a weighted ensemble model $S_n^{ens} = \sum_{n=1}^N w_n^k S_n^k$ with $w_n^k \geq 0$ and $\sum_{k=1}^K w_n^k = 1$, the MSE loss of the n -th ensemble score S_n^{ens} can be decomposed into two parts,*

$$l_m(\pi_n, S_n^{ens}) = \sum_{k=1}^K w_n^k l_m(\pi_n, S_n^k) - \sum_{k=1}^K w_n^k A_n^k \quad (7)$$

where A_n^k indicates the ambiguity term,

$$A_n^k = (S_n^k - S_n^{ens})^2 \quad (8)$$

PROOF. For each basic-model score S_n^k , we expand the MSE loss $l_m(\pi_n, S_n^k)$ in Eq. 3 around point S_n^{ens} by Taylor expansion with Lagrange type reminder (π_n is removed when the meaning is clear),

$$l_m(S_n^k) = l_m(S_n^{ens}) + \frac{\partial l_m(\tilde{S}_n^{ens})}{\partial \tilde{S}_n^{ens}} (S_n^k - S_n^{ens}) + \frac{1}{2!} \frac{\partial^2 l_m(\tilde{S}_n^{ens})}{(\partial \tilde{S}_n^{ens})^2} (S_n^k - S_n^{ens})^2 \quad (9)$$

Where \tilde{S}_n^{ens} is an interpolation point between S_n^{ens} and S_n^k . Define A_n^k as Eq.8, we weighted sum losses of all basic models as follows,

$$\begin{aligned} \sum_{k=1}^K w_n^k l_m(S_n^k) &= \sum_{k=1}^K [w_n^k l_m(S_n^{ens}) + \frac{\partial l_m(\tilde{S}_n^{ens})}{\partial \tilde{S}_n^{ens}} w_n^k (S_n^k - S_n^{ens}) + w_n^k A_n^k] \\ &= l_m(S_n^{ens}) + \frac{\partial l_m(\tilde{S}_n^{ens})}{\partial \tilde{S}_n^{ens}} \left(\sum_{k=1}^K w_n^k S_n^k - S_n^{ens} \right) + \sum_{k=1}^K w_n^k A_n^k \\ &= l_m(S_n^{ens}) + \sum_{k=1}^K w_n^k A_n^k \end{aligned} \quad (10)$$

The first equation is due to $\sum_{k=1}^K w_n^k = 1$ and $\partial^2 l_m / (\partial S)^2 = 2$, and the second equation is due to Eq. 1. Therefore,

$$l_m(S_n^{ens}) = \sum_{k=1}^K w_n^k l_m(S_n^k) - \sum_{k=1}^K w_n^k A_n^k \quad (11)$$

Proof done. \square

Since the ambiguity A_n^k in Eq. 8 is positive and $w_n^k \geq 0$, Eq. 7 follows the form of EA decomposition. Ranking ensemble with item-level weights for point-wise loss is effective theoretically, since the ensemble loss is smaller than weighted sum of basic model losses with any w_n^k as long as $w_n^k \geq 0$ and $\sum_{k=1}^K w_n^k = 1$.

For brevity, we will omit statements of score lists and ensemble formulas in the following theorems.

4.2 Pair-wise Loss

THEOREM 2 (GENERALIZED EA DECOMPOSITION THEORY FOR PAIR-WISE LOSS). *When $w_n^k \geq 0$, $\sum_{k=1}^K w_n^k = 1$, and $|w_m^k - w_n^k| \leq \delta$, $\forall m, n$, the BPR loss of a pair of ensemble scores S_n^{ens} and S_m^{ens} can be decomposed into*

$$l_b(S_n^{ens}, S_m^{ens}) < \sum_{k=1}^K w_n^k l_b(S_n^k, S_m^k) + \delta \sum_{k=1}^K S_m^k - \sum_{k=1}^K w_n^k A_{nm}^k \quad (12)$$

Where A_{nm}^k is the ambiguity of scores generated from basic models,

$$A_{nm}^k = \sigma(\tilde{z}^{ens})(1 - \sigma(\tilde{z}^{ens})) \sum_{k=1}^K w_n^k (z_{nm}^k - z_{nm}^{ens})^2 \quad (13)$$

$z_{nm}^* = S_n^* - S_m^*$ denotes the differences between scores.

Due to space limitation, we only show key steps in the proof:

PROOF. Let $z_{nm}^k = S_n^k - S_m^k$ and $l_b(z_{nm}^*) = l_b(S_n^*, S_m^*)$ in Eq.4, we expand $l_b(z_{nm}^k)$ around z_{nm}^{ens} by Taylor expansion,

$$\begin{aligned} l_b(z_{nm}^k) &= l_b(z_{nm}^{ens}) + \frac{\partial l_b(\tilde{z}^{ens})}{\partial \tilde{z}^{ens}} (z_{nm}^k - z_{nm}^{ens}) + \frac{1}{2!} \frac{\partial^2 l_b(\tilde{z}^{ens})}{(\partial \tilde{z}^{ens})^2} \\ &:= l_b(z_{nm}^{ens}) - B_{nm}^k + A_{nm}^k \end{aligned} \quad (14)$$

Where \tilde{z}^{ens} is an interpolation point between z_{nm}^{ens} and z_{nm}^k . With the limitation that $\sum_{k=1}^K w_n^k = 1$ and $|w_n^k - w_m^k| \leq \delta$, the weighted sum of B_{nm}^k is limited by

$$\begin{aligned} \sum_{k=1}^K w_n^k B_{nm}^k &= [1 - \sigma(\tilde{z}^{ens})] \sum_{k=1}^K w_n^k (z_{nm}^k - z_{nm}^{ens}) \\ &\leq \sum_{k=1}^K |w_n^k - w_m^k| |S_m^k| \leq \sigma \sum_{k=1}^K S_m^k \end{aligned} \quad (15)$$

Sum both sides of Eq.14 with weights, we get

$$l_b(z_{nm}^{ens}) < \sum_{k=1}^K w_n^k l_b(z_{nm}^k) + \sigma \sum_{k=1}^K S_m^k - \sum_{k=1}^K w_n^k A_{nm}^k \quad (16)$$

Proof done. \square

The range limitation of w_n^k leads to $\delta \leq 1$. And in pair-wise loss, the order rather than the values of scores matters. So the second term in Eq. 12 ($\delta \sum_{k=1}^K S_m^k < \sum_{k=1}^K S_m^k$) can be arbitrarily small. Meanwhile, the ambiguity A_{nm}^k is semi-positive. Therefore, Eq.12 follows the form of EA decomposition, and our ranking ensemble method with pair-wise loss is effective theoretically.

4.3 List-wise Loss

THEOREM 3 (GENERALIZED EA DECOMPOSITION THEORY FOR LIST-WISE LOSS). When $w_n^k \geq 0$, $\sum_{k=1}^K w_n^k = 1$, and $|w_n^k - w_m^k| \leq \delta$ for any m and n , the list-wise loss of ensemble scores $S^{ens} = \{s_1^{ens}, s_2^{ens}, \dots, s_n^{ens}\}$ (sorted with π) can be decomposed as

$$l_{p-l}(\pi, S^{ens}) < \sum_{k=1}^K w_{\max}^k l_{p-l}(\pi, S^k) + \delta N S_{sum}^{\max} - \sum_{k=1}^K \sum_{n=1}^N w_n^k A_n^k \quad (17)$$

Where w_{\max}^k denotes the maximum of all weights in list k , A_n^k is the ambiguity at position n ,

$$A_n^k = \frac{\left[\sum_{m=n+1}^N \exp(-\tilde{z}_{nm}^{ens})(z_{nm}^k - z_{nm}^{ens}) \right]^2}{\left(1 + \sum_{m=n+1}^N \exp(\tilde{z}_{nm}^{ens}) \right)^2} \quad (18)$$

S_{sum}^{\max} is defined as

$$S_{sum}^{\max} = \max_{m=1}^N \sum_{k=1}^K S_m^k \quad (19)$$

$z_{nm}^* = S_n^* - S_m^*$ denotes the differences between scores.

Due to space limitation, we only show key steps in the proof:

PROOF. We define the score difference $z_{n:N} = [z_{n+1}, \dots, z_N] = [S_n - S_{n+1}, \dots, S_n - S_N]$ and the logarithm pseudo-sigmoid function,

$$g_n(z_{n:N}) = \log \left(1 + \sum_{m=n+1}^N \exp(-z_{nm}) \right) \quad (20)$$

For each basic model of a list of items $S^k = \{S_n^k | n \in \{1, 2, \dots, N\}\}$, the PL loss is $l_{p-l}(S^k) = \sum_{n=1}^N g_n(z_{n:N}^k)$. We expand $g_n(z_{n:N}^k)$ around point $z_{n:N}^{ens}$ by Taylor expansion with Lagrange type reminder,

$$\begin{aligned} g_n(z_{n:N}^k) &= g_n(z_{n:N}^{ens}) + [\nabla g_n(z_{n:N}^{ens})]^T [z_{n:N}^k - z_{n:N}^{ens}] \\ &\quad + \frac{1}{2!} [z_{n:N}^k - z_{n:N}^{ens}]^T H_n(z_{n:N}^{ens}) [z_{n:N}^k - z_{n:N}^{ens}] \\ &:= g_n(z_{n:N}^{ens}) - B_n^k + A_n^k \end{aligned} \quad (21)$$

With the limitation that $\sum_{k=1}^K w_n^k = 1$, $w_n^k \geq 0$, and $|w_n^k - w_m^k| < \delta$, $\forall n, m$, the weighted sum of B_n^k on K basic models will be

$$\begin{aligned} \sum_{k=1}^K w_n^k B_n^k &= \frac{\sum_{m=n+1}^N \exp(-\tilde{z}_{nm}^{ens}) \sum_{k=1}^K w_n^k [z_{nm}^k - z_{nm}^{ens}]}{1 + \sum_{m=n+1}^N \exp(-\tilde{z}_{nm}^{ens})} \\ &< \delta \cdot S_{sum}^{\max} \end{aligned} \quad (22)$$

Therefore, sum from $n = 1$ to $n = N$, we get

$$l_{p-l}(S^{ens}) < \sum_{k=1}^K w_{\max}^k l_{p-l}(S^k) + \delta N S_{sum}^{\max} - \sum_{n=1}^N \sum_{k=1}^K w_n^k A_n^k \quad (23)$$

Proof done. \square

Because in the list-wise optimization, the order rather than the values of scores matters, S_n^{ens} can be arbitrarily small. Meanwhile, the ambiguity term A_n^k is semi-positive. Therefore, Eq.17 conforms to the EA decomposition theory, and our ranking ensemble method with list-wise loss is effective.

4.4 Ensemble Loss for Model Training

The above theorems guarantee our proposed ranking ensemble learning method in theory for three representative loss functions. With EA decomposition theory, we prove that the loss of ensemble list is smaller than any weighted sum combination of losses of basic lists: $l_{ens}(\pi, S^{ens}) \leq \sum_k w_k l_k(\pi, S^k) - A + \Delta$, $\forall w_k \leq 0$, $\sum_{k=1}^K w_k = 1$, where A is a positive ambiguity term, and Δ is arbitrarily small. Therefore, the ensemble loss $l_{ens}(\pi, S^{ens})$ (i.e., differences between ensemble list and ground truth) is possible to be smaller than any basic list loss with suitable weights $\{w_n^k\}$, and larger ambiguity A will lead to a smaller bound of ensemble loss. Thus, it can be effective for our ranking ensemble task.

In practice, since basic lists are fixed (so $l_k(\pi, S^k)$ are constants), we aim to minimize the ensemble ranking loss $l_{ens}(\pi, S^{ens})$ and maximize the ambiguity A . Therefore, the loss function for ranking ensemble learning, l_{el} , is defined as follows,

$$l_{el} = l_{ens}(\pi, S^{ens}) - \alpha A \quad (24)$$

where $l_{ens}(\pi, S^{ens})$ can be any of the $l_m(\pi, S^{ens})$, $l_b(\pi, S^{ens})$, and $l_{p-l}(\pi, S^{ens})$, and A indicates the ambiguity term. For BPR and P-L loss, there exists an interpolation $\tilde{S}_n^{ens} = S_n^{ens} + \theta(S_n^k - S_n^{ens})$ in A . To simplify the calculation, we let $\theta \rightarrow 0$ without loss of generality.

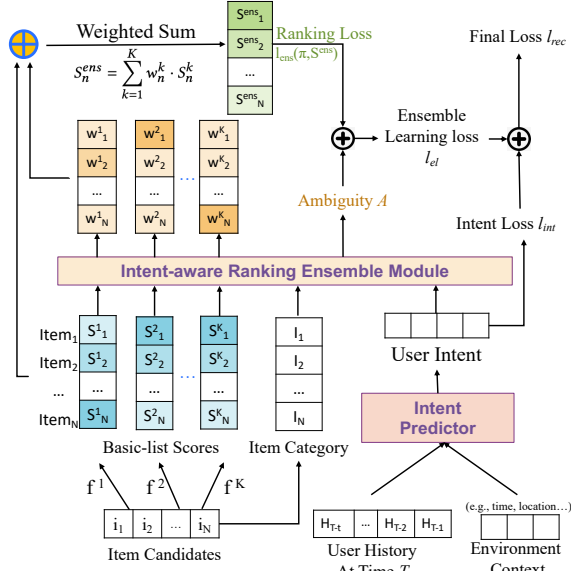


Figure 2: Overall framework of the Intent-aware Ranking Ensemble Learning (IntEL) model, where l_{ens} and l_{rec} are generated by Eq.24 and Eq.29, respectively.

5 INTENT-AWARE RANKING ENSEMBLE METHOD

5.1 Overall Framework

After we proved the effectiveness of item-level weights $\{w_n^k\}$ for ranking ensemble with three different loss functions, we need to design a neural network for learning the weights w_n^k . As shown in Section 3.2, users' intents about behaviors and item categories help aggregate the basic lists, while intents are not available in advance. Therefore, an intent predictor and an intent-aware ranking ensemble module are designed for our method.

The main framework of our Intent-aware Ensemble Learning (IntEL) method is shown in Figure 2. For a user u at time T , user intents Int are predicted with an intent predictor from her historical interactions and current environment context. Then, with N candidate items generated from K basic models, an intent-aware ranking ensemble module is adopted to integrate basic list scores, item categories, and the predicted user intents. The output of the ensemble module is item-level weights $\{w_n^k\}$ for each item n and basic model k . Eventually, weighted sum of all basic list scores constructs the ensemble scores $\{S_n^k\}$ for a final list. Since we focus on the ranking ensemble learning problem, a straight-forward sequential model is used for intent prediction in Section 5.2, and we pay more attention to the design of ensemble module in Section 5.3. IntEL is optimized with a combination of ranking loss $l_{ens}(\pi, S)$, ambiguity loss A , and intent prediction loss l_{int} . Details about the model learning strategy will be discussed in Section 5.4.

5.2 User Intent Predictor

As defined in Section 3.2, user intent describes a multi-dimensional probability distribution Int over different item categories and behaviors at each user visit. The goal of the intent predictor is to generate an intent probability distribution for each user visit. We predict

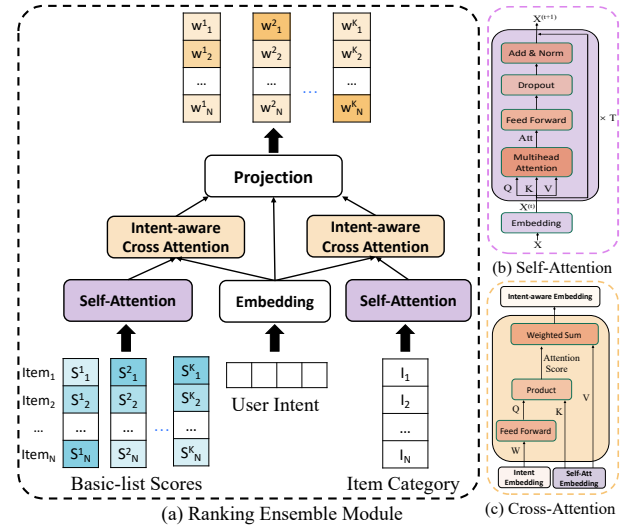


Figure 3: Structure of the intent-aware ranking ensemble module.

intents with users' historical interactions and environment context, as both historical habits and current situations will influence users' intents.

For a user u at time T , her historical interactions from $T - t$ to $T - 1$ and environment context (such as timestamp and location) at T are adopted to predict her intent at T , where t is a pre-defined time window. Environment context is encoded into embedding $c(u, T)$ with a linear context encoder. Two sequential encoders are utilized to model historical interactions at user visit (i.e., session) level and item level. Session-level history helps learn users' habits about past intents, while item-level interactions express preferences about item categories in detail. At the session level, the intents and context of each historical session are embedded with two linear encoders, respectively. Then two embeddings are concatenated and encoded with a sequential encoder to form an embedding $h_s(u, T)$. At the item level, "intent" of each positive historical interaction can also be represented by its behavior type and item category. Then item-level "intent"s are embedded with the same intent encoder as session-level, and fed into a sequential encoder to form item-level history $h_i(u, T)$. The sequential encoder can be any sequential model, such as GRU [12], transformer [37], etc. Finally, context $c(u, T)$, session-level $h_s(u, T)$, and item-level $h_i(u, T)$ are concatenated for a linear layer to predict intent $Int(u, T)$ (u and T are omitted),

$$Int = \text{Softmax}(\mathbf{W}^I [c, h_s, h_i] + b^I) \quad (25)$$

Where \mathbf{W}^I and b^I are linear mapping parameters.

5.3 Design of Ensemble Module

The structure of the intent-aware ranking ensemble module is shown in Figure 3(a). Since the final weights $\{w_n^k\}$ should be learned from both behavior and item categories aware of user intents, the predicted intents, single-behavior-objective basic list scores, and categories of items in basic lists are adopted as inputs.

Firstly, lists of item scores $\{S_n^k | k \in \{1, 2, \dots, K\}, n \in \{1, 2, \dots, N\}\}$ are fed into a self-attention layer to represent the relationship among item scores in the same basic list. Item categories $\{I_n | n \in$

$\{1, 2, \dots, N\}$ are also encoded with a self-attention layer to capture the intra-list category distributions. The self-attention structure consists of a linear layer to embed scores $\{S_n^k\}$ (or categories $\{I_n\}$) into d_e -dimensional representations $S \in \mathbb{R}^{N \times d_e}$ (or $I \in \mathbb{R}^{N \times d_e}$) and T layers of multi-head attentions, which follow the cross-relation attention layer proposed by Wang et al. [40], as shown in Figure 3(b).

Secondly, user intent Int is embedded into d_{int} dimension with a linear projection $Int_d = W^i Int \in \mathbb{R}^{N \times d_{int}}$. Then the influences of user intent on representations of scores and features are obtained with cross-attention layers,

$$A_s = \text{Attention}(Q = W^Q Int_d, K = S, V = S) \quad (26)$$

$$A_i = \text{Attention}(Q = W^Q Int_d, K = I, V = I) \quad (27)$$

Where the projection matrix $W_Q \in \mathbb{R}^{d_e \times d_{int}}$ is shared between two intent-aware attention modules. Since behavior intents and category intents are associated when users interact with recommenders, we use the holistic user intents to guide the aggregations of both basic list scores and item categories rather than splitting the intents into two parts.

Finally, weights $\{w_n^k\}$ should be generated from all information. Intent-aware score embeddings A_s , intent-aware item category embeddings A_i , and intent embedding Int_d are concatenated and projected into space of \mathbb{R}^K to get the weight matrix $W \in \mathbb{R}^{N \times K}$,

$$W = W^w \cdot ([A_s, A_i, Int_d]) \quad (28)$$

Where $W^w \in \mathbb{R}^{K \times (2d_e + d_{int})}$ is a trainable projection matrix. The output matrix $W = \{w_n^k\}$ is used as the weights for summing basic model scores as in Eq. 1.

5.4 Model Learning Strategy

Since an end-to-end framework is to train the intent predictor module and intent-aware ranking ensemble module, joint learning of two modules is utilized for model optimization.

To optimize ranking ensemble results according to theorems via EA decomposition in Section 4, ensemble learning loss l_{el} consists of $l_{ens}(\pi, S^{ens})$ and A as in Eq. 24. Meanwhile, accurate user intents will guide the ranking ensemble, so an intent prediction loss is also used for model training. Since user intents are described by multi-dimensional distributions, KL-divergence [13] loss l_{int} is adopted to measure the distance between true intents Int and predicted intents \hat{Int} . The final recommendation loss l_{rec} is a weighted sum

$$l_{rec} = l_{el} + \gamma l_{int} = l_{ens}(\pi, S^{ens}) - \alpha A + \gamma l_{int} \quad (29)$$

Where $l_{ens}(\pi, S^{ens})$ is the ranking ensemble loss, A is the ambiguity term, and l_{int} is the intent prediction loss. α and γ are hyper-parameters to adjust the weights of ambiguity and intent loss, respectively.

6 EXPERIMENTS

6.1 Experimental Setup

6.1.1 Dataset. Experiments are conducted on a public online shopping recommendation dataset **Tmall**² and a private local life service recommendation dataset **LifeData**.

²<https://tianchi.aliyun.com/dataset/dataDetail?dataId=42>

Table 2: Datasets statistics in ranking ensemble experiments.

Dataset	#Item	#User	#Click	#Buy	#Fav.	#Session
Tmall	142.5k	148.4k	2.439m	185.5k	271.9k	743.3k
LifeData	165.5k	82.7k	819.4k	82.5k	-	559.2k

Tmall is a multi-behavior dataset from the IJCAI-15 competition, which contains half-year user logs with Click, Add-to-favorite (Fav.), and Buy interactions on Tmall online shopping platform. We employ the data in September for ensemble learning and exclude items and users with less than 3 positive interactions. Following the data processing strategy by Shen et al. [36], we treat a user's interactions within a day as a session. Three-week interactions before the ensemble dataset are used for the generation of basic-model scores, which will be discussed in Section 6.1.2.

LifeData comes from a local life service App, where the recommender provides users with nearby businesses such as restaurants or hotels. Users may click or buy items on the platform. One-month interactions of a group of anonymous users are sampled, and users and items with less than 3 positive interactions are filtered. A user's each visit (i.e., entering the App) is defined as a session, and sessions with positive interactions are retained.

Basic models are optimized for each of the behaviors, which will be introduced in Section 6.1.2. Ranking ensemble is conducted at session level, and interactions in most $t = 20$ historical sessions are considered in the intent predictor. Detailed statistics are shown in Table 2, which includes the dataset for ensemble learning only while excluding the data used for basic model generation. Moreover, training data for basic models have no overlap with ensemble learning data.

6.1.2 Basic-model Score and Intent Generation. In IntEL, basic scores are pre-generated and fixed during ranking ensemble. For **Tmall**, we adopted DeepFM [20] as basic models to train three models with Click, Fav., and Buy objectives separately. In each session, we select the top 30 items predicted by each basic model to construct three item sets, and take the union of them, plus positive interactions of the session, to form the basic item lists for reranking. Please refer to our public repository for details about basic model training strategy³. For **LifeData**, two basic score lists are used for ranking ensemble, which are sorted by predicted clicking probability and buying probability provided by the platform, respectively.

As for intents, in **Tmall**, $|B| = 3$, and we merge categories with less than 50 items, resulting in category $|I| = 357$. In **LifeData**, $|I| = 6$ and $|B| = 2$. Hence, the dimension for intent Int is 1071 for **Tmall** and 12 for **LifeData**. Intent ground truth Int probability is calculated from all positive interactions in each session.

6.1.3 Baseline Methods. We compare IntEL against basic models and several ranking ensemble baselines as follows,

1. **Single XXX**: Use one of the basic models' scores to rank the item list. XXX indicates Click, Fav., and Buy, respectively.
2. **RRA** [21]: An unsupervised ensemble method, where items are sorted with their significance in basic-model lists.
3. **Borda** [5]: An unsupervised ensemble method to take the average ranks of all basic models as the final ranking.

Table 3: Main differences between two datasets. Pos. indicates positive interactions.

Dataset	#Intent	Avg. Session Length	Avg. Pos./Session
Tmall	1,071	68.37	3.73
LifeData	12	32.78	1.47

4. **λ Rank** [7]: A gradient-based optimization method used for learning2rank task. We regard items as documents, basic-model scores and item categories as document features, and MLP as a backbone model.

5. **ERA** [30]: An evolutionary method to aggregate some basic-list features with Genetic Algorithm (GA), where fitness function is calculated on validation set.

6. **aWELv** [24]: A personalized ranking ensemble method to assign weights at basic model level, i.e., $w_n^k = w_m^k$ for any n, m . We adopt the list-wise training loss following [24].

7. **aWELv+Int/IntEL**: Two variations of aWELv considering user intents. Intents are predicted as a feature for aWELv+Int. The IntEL module is used for predicting list-level weights for aWELv+IntEL.

Our methods are shown as **IntEL-MSE**, **IntEL-BPR**, and **IntEL-PL** with three different kinds of loss functions.

6.1.4 Experimental settings. We split both datasets along time: the last week is the test set, and the last three days from the training set is the validation set. The priority for the multi-level ground truth π are Buy>Favorite>Click>Examine for **Tmall**, and Buy>Click>Examine for **LifeData**. As for evaluation, we adopt NDCG@3, 5, and 10 to evaluate the ensemble list S^{ens} on the multi-level ground truth π (i.e., all) and each behavior objective.

We implement IntEL model in *PyTorch*, and the code of IntEL and all baselines are released³. Each experiment is repeated with 5 different random seeds and average results are reported. All models are trained with Adam until convergence with a maximum of 100 epochs. For a fair comparison, the batch size is set to 512 for all models. We tuned the parameters of all methods over the validation set, where the learning rate are tuned in the range of $[1e-4, 1e-2]$ and all embedding size are tuned in $\{16, 32, 64\}$. Specifically, for IntEL, we found that it has stable performance when GRU [12] with embedding=128 is used for the intent predictor, and self-attention with $T = 2$ for Tmall and $T = 1$ for LifeData. The ambiguity loss weight α is set to $1e-5$, $1e-5$, and $1e-4$ for IntEL-MSE, IntEL-BPR, and IntEL-PL. Hyper-parameter details are released³.

6.2 Overall Performance

The overall performances on **Tmall** and **LifeData** are shown in Table 4 and Table 5, respectively. We divide all models into four parts: The first part evaluates on each single-objective basic model’s scores. The second is unpersonalized baseline ensemble models, and the third contains personalized baselines: aWELv and its two variants with user intents. The last part shows our method IntEL with three loss functions. From the results, we have several observations:

First, our proposed IntEL achieves the best performance on all behavior objectives in both datasets. IntEL with three loss functions, i.e., IntEL-MSE, IntEL-BPR, and IntEL-PL, outperform the best baselines on most metrics significantly. Although the two datasets are

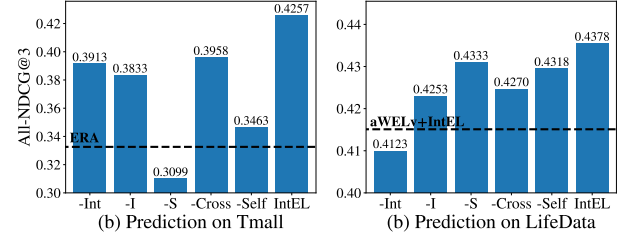


Figure 4: Ablation study. Performance comparison between IntEL and its variant, i.e., without: Intent modeling (-Int), item categories (-I), basic score lists (-S), cross-attention (-Cross), and self-attention (-Self).

quite different, as shown in Table 3, IntEL has stable, great ensemble results on both datasets.

Second, IntEL with different loss functions show different performances on two datasets. On Tmall, IntEL-MSE is better than IntEL-BPR and IntEL-PL. It is because there is four-level ground truth π (three behaviors), and ranking on such diverse item lists is close to rating prediction. Therefore, IntEL-MSE, which directly optimizes the ensemble scores, performs better than IntEL-BPR and IntEL-PL, which optimize the comparison between rankings. On LifeData, IntEL-PL and IntEL-BPR perform better since LifeData has shorter sessions with fewer positive interactions (as in Table 3). So comparison-based BPR and P-L achieve better performance.

Third, comparing different baselines, we find that supervised methods (λ Rank, ERA, and aWELv) outperform unsupervised RRA and Borda greatly on Tmall. It is because heterogeneous single-behavior objective models (Single XXX) have diverse performance, making rank aggregation difficult for unsupervised methods.

Lastly, aWELv and its variants perform well on LifeData but not on Tmall since session lists are generally longer (Table 3) for Tmall, and list-level weights of aWELv miss useful intra-list information. So item-level weights that consider item category intents are necessary. Nevertheless, aWELv is better than basic models in both datasets, which is consistent with the theory. Moreover, aWELv+Int/IntEL outperform aWELv on most metrics, indicating that user intents contributes to ranking ensemble learning.

6.3 Further Analysis

To further explore the performance of our ranking ensemble learning method, we conduct an ablation study, analysis of user intents, and hyper-parameters analysis on the best model for each dataset, i.e., IntEL-MSE for **Tmall** and IntEL-PL for **LifeData**.

6.3.1 Ablation Study. The main contributions of our proposed IntEL include adopting user intents for heterogeneous ranking ensemble and integration of basic-list scores, item categories, and user intents. We compare IntEL with five variants: Excluding one of the inputs: **-Int** (without intent), **-I** (without item categories), and **-S** (without basic-list scores). Replacing two main elements: **-Cross**, removing the intent-aware cross-attention layer; and **-Self**, replacing the self-attention layer with a direct connection.

NDCG@3 on the general multi-level ranking list of variants and IntEL are shown in Figure 4. Ranking performance drops on all five variants, indicating all inputs and two attention layers contribute to the performance improvement of IntEL. Removing scores

³<https://github.com/JiayuLi-997/IntEL-SIGIR2023>.

Table 4: Results of IntEL with three different loss functions and baseline methods on Tmall. Boldface shows the best result. Underline indicates the best baseline. Notation **/* demonstrates significantly better than the best baseline with $p < 0.05/0.01$.

Model	All-NDCG@K			Click-NDCG@K			Fav.-NDCG@K			Buy-NDCG@K		
	K=3	K=5	K=10	K=3	K=5	K=10	K=3	K=5	K=10	K=3	K=5	K=10
Single Click	0.1356	0.1473	0.1673	0.1435	0.1532	0.1721	0.0701	0.0829	0.1014	0.0918	0.1057	0.1243
Single Fav.	0.0752	0.0874	0.1066	0.0779	0.0894	0.1083	0.0630	0.0748	0.0920	0.0492	0.0607	0.0765
Single Buy	0.0572	0.0689	0.087	0.0587	0.0699	0.0878	0.0393	0.0489	0.0638	0.0632	0.0776	0.0974
RRA	0.0960	0.1093	0.1317	0.0998	0.1120	0.1341	0.0683	0.0813	0.1014	0.0753	0.0907	0.1122
Borda	0.1258	0.1398	0.1626	0.1317	0.1440	0.1660	0.0741	0.0880	0.1081	0.0830	0.0989	0.1218
λ Rank	0.2742	0.2797	0.3003	<u>0.3104</u>	<u>0.3064</u>	<u>0.3189</u>	0.1878	0.2122	0.2472	0.1376	0.1586	0.1913
ERA	<u>0.3325</u>	<u>0.3378</u>	<u>0.3623</u>	0.2301	0.2420	0.2716	<u>0.1933</u>	<u>0.2156</u>	<u>0.2502</u>	<u>0.1921</u>	<u>0.2163</u>	<u>0.2504</u>
aWELv	0.1387	0.1533	0.1770	0.1469	0.1584	0.1811	0.0837	0.0986	0.1197	0.1025	0.1198	0.1436
aWELv+Int	0.1398	0.1574	0.1784	0.1484	0.1592	0.1822	0.0903	0.1016	0.1183	0.1030	0.1259	0.1445
aWELv+IntEL	0.1427	0.1556	0.1774	0.1535	0.1620	0.1821	0.0906	0.0934	0.1120	0.1042	0.1263	0.1451
IntEL-MSE	0.4257**	0.4364**	0.4676**	0.4693**	0.4680**	0.4712**	0.2943**	0.3271**	0.3731**	0.2433*	0.2760**	0.3100**
IntEL-BPR	0.3992*	0.3859*	0.3755	0.4417**	0.4157**	0.3960*	0.2791**	0.2943**	0.3068*	0.2344*	0.2508*	0.2630
IntEL-PL	0.4041**	0.3865*	0.3678	0.4367**	0.4060**	0.3829**	0.2811**	0.2934**	0.3032*	0.2355*	0.2472*	0.2594

Table 5: Results of IntEL with three different loss functions and baseline methods on LifeData. Boldface shows the best result. Underline indicates the best baseline. Notation **/* demonstrates significantly better than the best baseline with $p < 0.05/0.01$.

Model	All-NDCG@K			Click-NDCG@K			Buy-NDCG@K		
	K=3	K=5	K=10	K=3	K=5	K=10	K=3	K=5	K=10
Single Click	0.4004	0.4443	0.4972	0.4009	0.4449	0.4980	0.6365	0.6665	0.6918
Single Buy	0.3102	0.3526	0.4070	0.3102	0.3528	0.4072	0.6893	0.7211	0.7438
RRA	0.3539	0.4020	0.4586	0.3540	0.4022	0.4590	0.6556	0.6865	0.7104
Borda	0.4094	0.4538	0.5061	0.4097	0.4541	0.5066	0.7030	0.7250	0.7447
λ Rank	0.4129	0.4487	0.4830	0.4133	0.4492	0.4835	0.6866	0.7083	0.7225
ERA	0.4063	0.4451	0.5053	0.4181	0.4579	0.5112	0.5782	0.6307	0.6764
aWELv	0.4074	0.4531	0.5033	0.4077	0.4535	0.5041	<u>0.7063</u>	<u>0.7339</u>	0.7466
aWELv+Int	0.4150	0.4607	0.5143	0.4151	0.4610	0.5147	0.6962	0.7271	0.7482
aWELv+IntEL	<u>0.4174</u>	<u>0.4663</u>	<u>0.5176</u>	<u>0.4189</u>	<u>0.4638</u>	<u>0.5171</u>	0.7036	0.7318	<u>0.7503</u>
IntEL-MSE	0.4253**	0.4695*	0.5211**	0.4257**	0.4700**	0.5217*	0.7096	0.7379	0.7498
IntEL-BPR	0.4308**	0.4752**	0.5268**	0.4312**	0.4757**	0.5275**	0.7115*	0.7390*	0.7609**
IntEL-PL	0.4378**	0.4819**	0.5332**	0.4382**	0.4825**	0.5339**	0.7093	0.7382*	0.7604**

Table 6: Intent prediction and ranking ensemble performance comparison with different treatments on user intents.

Dataset	Tmall			LifeData		
	-Int	His.Avg.	IntEL	-Int	His.Avg.	IntEL
I-Perform	-	0.1829	0.2347	-	0.2663	0.3298
E-NDCG@3	0.3913	0.4011	0.4257	0.4123	0.4265	0.4378

and the self-attention layer both lead to considerable performance decreases on Tmall, showing that intra-list basic scores information is essential, as long sessions are included in Tmall. Removing user intents leads to the most dramatic degradation on LifeData, which suggests it is important to adopt user intents for the multiple-objective ranking ensemble. Nevertheless, the ablation variants still outperform all basic lists (i.e., Single XXX), which aligns with our proof of loss reduction via EA ambiguity decomposition.

6.3.2 Influence of User Intent. Since intents are essential for ranking ensemble learning, we explore the influence of intent prediction

accuracy. Therefore, we compare IntEL with two variants: **-Int**, IntEL without user intents as input, and **His.Avg.**, predicting a user's current intents as her average historical session intents.

Since **Tmall** has 1071 intents and **LifeData** has 12 intents, we utilize NDCG@10 and Macro-F1 as intent performance (I-perform) indicators, respectively. And ensemble results (E-NDCG@3) are evaluated by All-NDCG@3 for both datasets. The results are shown in Table 6. It indicates that better performance of ranking ensemble is achieved by adding intent prediction and improving prediction accuracy. Therefore, predicting user intents is helpful for ranking ensemble in recommendation. On the other, *His.Avg.* works better than all baselines, providing an efficient and effective possible implementation in application.

6.3.3 Hyper-parameters Analysis. Since the construction of loss l_{rec} (Eq.29) is essential for our method, we analyze the influence of hyper-parameters during model optimization. Two hyper-parameters are considered in the optimization loss l_{rec} : α , the weight for basic list ambiguity A ; γ , the weight for intent prediction loss l_{int} . All-NDCG@3 with different hyper-parameters on two datasets are

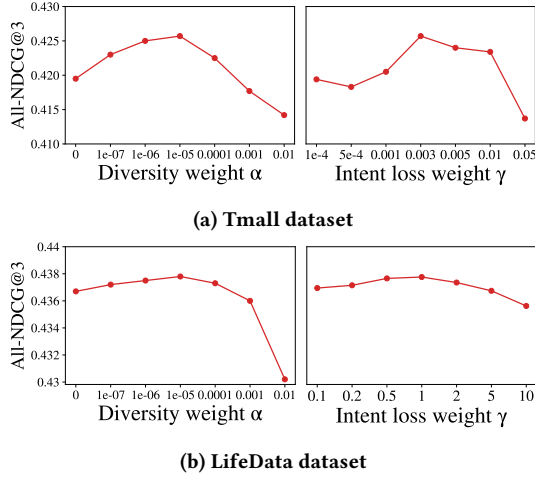


Figure 5: Ranking ensemble results of IntEL with different hyper-parameters.

shown in Figure 5. It illustrates that too large or small α will both lead to ranking ensemble performance decrease. Especially when α is too large, the model will focus on maximizing basic model ambiguity A to minimize l_{rec} , while ensemble learning loss l_{el} is less optimized. As for the intent loss weight γ , performance on Tmall shows fluctuation with γ , while performance on LifeData is relatively stable. It is because intent prediction difficulty differs on two datasets: Tmall contains 1071 types of intents, which is hard to predict accurately, so a proper intent loss weight is essential for predictor optimization, while LifeData has only 12 intents, which is easier to capture and model.

7 CONCLUSION

In this paper, we propose a novel ranking ensemble method IntEL for intent-aware single-objective ranking lists aggregation. To our knowledge, we are the first to generalize ranking ensemble learning with item-level weights on heterogeneous item lists. And we are also the first to integrate user intents into rank aggregation in recommendation. We generalize the ranking ensemble with item-level weights and prove its effectiveness with three representative loss functions via error-ambiguity decomposition theory. Based on the proof, we design an ensemble learning loss l_{el} to minimize ranking ensemble loss l_{ens} and maximize ambiguity A . Then we design an intent-aware ranking ensemble learning model, IntEL, to learn weights for heterogeneous lists' ensemble. In IntEL, a sequential intent predictor and a two-layer attention intent-aware ensemble module are adopted for learning the personalized and adaptive ensemble weights with user intents. Experiments on two large-scale datasets show that IntEL gains significant improvements on multiple optimization objectives simultaneously.

This study still has some limitations. For basic list generation, we only applied one classical method, DeepFM, for different behaviors separately. However, multi-behavior methods are also possible models to generate multiple basic lists simultaneously, which may lead to different performance for IntEL. Also, a straight-forward method was adopted to predict intents and incorporate intent prediction loss. In the future, we will investigate the possibility of

integrating more heterogeneous basic lists for other objectives in recommendation with IntEL. As we find that more accurate user intents will lead to better ranking ensemble performance, we will also try to design more sophisticated intent predictors to achieve better results.

ACKNOWLEDGMENTS

We sincerely thank our anonymous reviewers for their insightful feedback. This work is supported by the Natural Science Foundation of China (Grant No. U21B2026, 62002191) and Beijing Academy of Artificial Intelligence.

REFERENCES

- [1] Michał Bałchanowski and Urszula Boryczka. 2022. Aggregation of Rankings Using Metaheuristics in Recommendation Systems. *Electronics* 11, 3 (2022), 369.
- [2] Michał Bałchanowski and Urszula Boryczka. 2022. Collaborative Rank Aggregation in Recommendation Systems. *Procedia Computer Science* 207 (2022), 2213–2222.
- [3] Avradeep Bhowmik and Joydeep Ghosh. 2017. Letor methods for unsupervised rank aggregation. In *Proceedings of the 26th international conference on world wide web*. 1331–1340.
- [4] Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Yong-Nan Zhu, Zhangming Chan, Na Mou, et al. 2022. CAN: Feature Co-Action Network for Click-Through Rate Prediction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 57–65.
- [5] JC de Borda. 1784. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences pour 1781 (Paris, 1784)* (1784).
- [6] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. 2005. Diversity creation methods: a survey and categorisation. *Information fusion* 6, 1 (2005), 5–20.
- [7] Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems* 19 (2006).
- [8] Pablo Castells, Neil Hurley, and Saul Vargas. 2022. Novelty and diversity in recommender systems. In *Recommender systems handbook*. Springer, 603–646.
- [9] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2942–2951.
- [10] Tong Chen, Hongzhi Yin, Hongxu Chen, Rui Yan, Quoc Viet Hung Nguyen, and Xue Li. 2019. Air: Attentional intention-aware recommender systems. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 304–315.
- [11] Yongjun Chen, Zhiwei Liu, Jia Li, Julian McAuley, and Caiming Xiong. 2022. Intent Contrastive Learning for Sequential Recommendation. In *Proceedings of the ACM Web Conference 2022*. 2172–2182.
- [12] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
- [13] Imre Csiszár. 1975. I-divergence geometry of probability distributions and minimization problems. *The annals of probability* (1975), 146–158.
- [14] Quanyu Dai, Haoxuan Li, Peng Wu, Zhenhua Dong, Xiao-Hua Zhou, Rui Zhang, Rui Zhang, and Jie Sun. 2022. A generalized doubly robust learning framework for debiasing post-click conversion rate prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 252–262.
- [15] Weiqiang Di. 2022. A multi-intent based multi-policy relay contrastive learning for sequential recommendation. *PeerJ Computer Science* 8 (2022), e1088.
- [16] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*. 613–622.
- [17] Cynthia Dwork, Ravi Kumar, Moni Naor, and D Sivakumar. 2001. Rank aggregation revisited.
- [18] Ronald Fagin, Ravi Kumar, and Dandapani Sivakumar. 2003. Efficient similarity search and classification via rank aggregation. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. 301–312.
- [19] Mohamed Farah and Daniel Vanderpooten. 2007. An outranking approach for rank aggregation in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 591–598.
- [20] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).
- [21] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. 2012. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 28, 4 (2012), 573–580.
- [22] Anders Krogh and Jesper Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems* 7 (1994).
- [23] Shangsong Liang, Ilya Markov, Zhaochun Ren, and Maarten de Rijke. 2018. Manifold learning for rank aggregation. In *Proceedings of the 2018 World Wide Web Conference*. 1735–1744.
- [24] Hongzhi Liu, Yingpeng Du, and Zhonghai Wu. 2022. Generalized Ambiguity Decomposition for Ranking Ensemble Learning. *Journal of Machine Learning Research* 23, 88 (2022), 1–36.
- [25] Zhaoyang Liu, Haokun Chen, Fei Sun, Xu Xie, Jinyang Gao, Bolin Ding, and Yanyan Shen. 2021. Intent preference decoupling for user representation on online recommender system. In *Proceedings of the Twenty-Ninth International Conference on Artificial Intelligence*. 2575–2582.
- [26] Zhiwei Liu, Xiaohan Li, Ziwei Fan, Stephen Guo, Kannan Achan, and S Yu Philip. 2020. Basket recommendation with multi-intent translation graph neural network. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 728–737.
- [27] Hui Luo, Jingbo Zhou, Zhifeng Bao, Shuangli Li, J Shane Culpepper, Haochao Ying, Hao Liu, and Hui Xiong. 2020. Spatial object recommendation with hints: When spatial granularity matters. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 781–790.
- [28] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. 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*. 1930–1939.
- [29] Joao Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. 2012. Ensemble approaches for regression: A survey. *Acm computing surveys (csur)* 45, 1 (2012), 1–40.
- [30] Samuel Oliveira, Victor Diniz, Anisio Lacerda, and Gisele L Pappa. 2016. Evolutionary rank aggregation for recommender systems. In *2016 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 255–262.
- [31] Samuel EL Oliveira, Victor Diniz, Anisio Lacerda, Luiz Merschmann, and Gisele L Pappa. 2020. Is rank aggregation effective in recommender systems? an experimental analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 2 (2020), 1–26.
- [32] Xiaofeng Pan, Ming Li, Jing Zhang, Keren Yu, Hong Wen, Luping Wang, Chengjun Mao, and Bo Cao. 2022. MetaCVR: Conversion Rate Prediction via Meta Learning in Small-Scale Recommendation Scenarios. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2110–2114.
- [33] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).
- [34] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. *Advances in neural information processing systems* 30 (2017).
- [35] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1249.
- [36] Qi Shen, Lingfei Wu, Yitong Pang, Yiming Zhang, Zhihua Wei, Fangli Xu, and Bo Long. 2021. Multi-behavior graph contextual aware network for session-based recommendation. *arXiv preprint arXiv:2109.11903* (2021).
- [37] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.
- [38] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Fourteenth ACM Conference on Recommender Systems*. 269–278.
- [39] Aayushi Verma and Shikha Mehta. 2017. A comparative study of ensemble learning methods for classification in bioinformatics. In *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*. IEEE, 155–158.
- [40] Chenyang Wang, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. 2020. Toward dynamic user intention: Temporal evolutionary effects of item relations in sequential recommendation. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2020), 1–33.
- [41] Chenyang Wang, Zhefan Wang, Yankai Liu, Yang Ge, Weizhi Ma, Min Zhang, Yiqun Liu, Junlan Feng, Chao Deng, and Shaoping Ma. 2022. Target Interest Distillation for Multi-Interest Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2007–2016.
- [42] Gang Wang, Jianshan Sun, Jian Ma, Kaiquan Xu, and Jibao Gu. 2014. Convex classification: The contribution of ensemble learning. *Decision support systems* 57 (2014), 77–93.
- [43] Shoujin Wang, Liang Hu, Yan Wang, Quan Z Sheng, Mehmet Orgun, and Longbing Cao. 2020. Intention nets: psychology-inspired user choice behavior modeling for next-basket prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6259–6266.
- [44] Yifan Wang, Weizhi Ma, Min Zhang*, Yiqun Liu, and Shaoping Ma. 2022. A survey on the fairness of recommender systems. *ACM Journal of the ACM (JACM)* (2022).
- [45] Xu-Cheng Yin, Kaizhu Huang, Chun Yang, and Hong-Wei Hao. 2014. Convex ensemble learning with sparsity and diversity. *Information Fusion* 20 (2014), 49–59.
- [46] Enming Yuan, Wei Guo, Zhicheng He, Huifeng Guo, Chengkai Liu, and Ruiming Tang. 2022. Multi-Behavior Sequential Transformer Recommender. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1642–1652.
- [47] Qihua Zhang, Junning Liu, Yuzhuo Dai, Yiyan Qi, Yifan Yuan, Kunlun Zheng, Fan Huang, and Xianfeng Tan. 2022. Multi-Task Fusion via Reinforcement Learning for Long-Term User Satisfaction in Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4510–4520.
- [48] Jieming Zhu, Jinyang Liu, Shuai Yang, Qi Zhang, and Xiuqiang He. 2021. Open Benchmarking for Click-Through Rate Prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2759–2769.