AOTree: Aspect Order Tree-Based Model for Explainable Recommendation

Wenxin Zhao[1], Peng Zhang[1*], Hansu Gu[2], Dongsheng Li[3], Tun Lu[1*], Ning Gu [1]

[1] Fudan University, Shanghai, China
[2] Seattle, Washington, USA
[3] Microsoft Research Asia, Shanghai, China
zhaowx21@m.fudan.edu.cn, zhangpeng_@fudan.edu.cn, hansug@acm.org, dongshengli@fudan.edu.cn, lutun@fudan.edu.cn, ninggu@fudan.edu.cn

Abstract

Recent recommender systems aim to provide not only accurate recommendations but also explanations that help users understand them better. However, most existing explainable recommendations only consider the importance of content in reviews, such as words or aspects, and ignore the ordering relationship among them. This oversight neglects crucial ordering dimensions in the human decision-making process, leading to suboptimal performance. Therefore, in this paper, we propose Aspect Order Tree-based (AOTree) explainable recommendation method, inspired by the Order Effects Theory from cognitive and decision psychology, in order to capture the dependency relationships among decisive factors. We first validate the theory in the recommendation scenario by analyzing the reviews of the users. Then, according to the theory, the proposed AOTree expands the construction of the decision tree to capture aspect orders in users' decision-making processes, and use attention mechanisms to make predictions based on the aspect orders. Extensive experiments demonstrate our method's effectiveness on rating predictions, and our approach aligns more consistently with the user's decision-making process by displaying explanations in a particular order, thereby enhancing interpretability.

Introduction

With the overwhelming amount of information online, recommendations that aim to provide users with suitable items (products or services) by predicting a user's interest have been widely integrated into e-commerce, social network, and other web applications (Karn et al. 2023). These systems play a crucial role in alleviating information overload, making decision-making more user-friendly and enhancing user experience in online shopping, social interaction, and more. The dual objectives of recommender system encompass not only accurately predicting preferences but also enhancing user satisfaction and trust. This motivates research to emphasize both recommendation accuracy and user-centric metrics, with interpretability standing out as an important one. This involves creating explainable recommender systems capable of not just suggesting items but also providing users with explanations (Tan et al. 2021). For instance, in an online shopping scenario, a recommendation for a mobile phone could include an explanation like *"This phone is purchased by 80% of your peers."*

In the realm of explainable recommendations, reviews emerge as a valuable resource for generating explanations. Reviews offer insights from users' perspectives, serving a dual purpose: 1) user's reviews help characterize their preferences (Li et al. 2017), and 2) item's reviews can be leveraged to generate explicit explanations for its recommendation (Mcauley and Leskovec 2013). Many recent studies have focused on using techniques like multi-view deep learning approaches to analyze features from reviews, and integrate them into recommender systems (Fan et al. 2022). The key concept involves using attention mechanisms to capture importance of different aspects reflected in user preferences and item characteristics and generating explanations accordingly (Zhang et al. 2014b). For instance, considering a user's emphasis on the aspect of price in their reviews, the explanation could be framed as *"We recommend this phone because its price matches your taste."*

Despite the success of existing review-based explainable recommendations, a crucial factor often overlooked is the aspect order in users' decision-making processes. The "Order Effects Theory" suggests that humans tend to follow a specific order of factors when deciding (Anderson 1965), influencing outcomes. For instance, a mobile photography enthusiast may prioritize camera when buying a phone, while students might prioritize appearance and price. The different orders can significantly reveal users' preferences. Existing methods primarily focus on decisive factors but neglect the ordering relationship between these factors.

Incorporating aspect order into review-based explainable recommendations poses key challenges. Firstly, constructing aspect order is intricate due to its personalized and dynamic nature across diverse situations, presenting complexities in modeling. Users exhibit different decision orders, and likewise, various items entail distinct orders of consideration during selection. Furthermore, users' contemplation of each aspect is not predetermined but dynamically influenced by contextual factors such as item characteristics and prior decisions (Meshram, Gopalan, and Manjunath 2016). Secondly, integrating aspect orders from both users and items is non-trivial, given the inherent trade-off between explainability and effectiveness in optimization goals. Simultaneously optimizing these multiple aspects exacerbates the challenges

*Corresponding authors.

in implementing and evaluating recommendations.

To address these challenges, our focus is on introducing the effects of aspect order into explainable recommender systems. First, we validate the "Order Effects Theory" in the recommendation context by analyzing Yelp dataset*, delving into aspect orders embedded in users' decision-making sequences in online shopping. Second, to model and incorporate aspect orders, we propose an **A**spect **O**rder **Tree**-based (AOTree) explainable recommendation method. This three-stage process aims to enhance both recommendation accuracy and explanations. Specifically, we construct User-AOTree and Item-AOTree structures for users and items, capturing personalized and dynamic features of order effects. The final decisive sequence is derived by integrating aspect orders from user and item perspectives, facilitating predictions that ensure recommendation accuracy, transparency, and explainability. Extensive experiments validate our method's effectiveness, demonstrating higher accuracy compared to five state-of-the-art baselines. Additionally, our approach aligns with user's decision-making process, thereby enhancing interpretability. To conclude, the main contributions of this work are summarized as follows:

- We validate the order effects in recommendation scenario by using a well-known recommendation dataset.
- We propose an Aspect Order Tree-based (AOTree) explainable recommendation method by jointly characterizing order effects from perspectives of users and items.
- Extensive experiments demonstrate our method's better accuracy and interpretability compared with several state-of-the-art methods.

The remainder of this work is organized as follows. First, we provide the related work to introduce the context of the current research. Then, we delve into data analysis to assess the existence of the theory and its applicability, outlining our motivation. Next, we elaborate on the AOTree explainable recommendation method framework. The subsequent section describes the experiments, along with the corresponding results and discussions. Finally, the conclusion is given.

Related Works

Since our work is to apply the Order Effects Theory to the construction of explainable recommender systems, we divide the related work into two subsections: (a) *Explainable Recommender Systems*; and (b) *Order Effects Theory*.

Explainable Recommender Systems. In the domain of explainable recommender system research, there are two types of dominant models, namely model-agnostic and model-intrinsic approaches (Lipton 2018). Since there is generally a trade-off between performance and transparency in recommender systems, these two approaches have their own advantages and shortcomings regarding performance and transparency (Zhang, Chen et al. 2020).

The model-agnostic approach, also named post-hoc explanation approach (Peake and Wang 2018), allows recommendation model to be a black box and generates explanations after the results have been obtained. Aspect-aware techniques are commonly utilized to generate explanations, i.e., with corresponding aspects. Zhang et al. (2014a) proposed an Explicit Factor Model (EFM) by aligning latent factors with aspects, generating recommendations and explanations based on the matching degree of aspects between items and users. Chen et al. (2016) extended EFM to tensor factorization models and constructed user-item-feature cube for pair-wise learning, which provides personalized recommendations and feature-based explanations. Wang et al. (2018a) applied a joint factorization framework to integrate user preference and opinionated content for recommendations and aspect-level explanations. Although model-agnostic methods can achieve high accuracy, they are complex and lack transparency.

On the contrary, the model-intrinsic approach aims to develop an explainable model, i.e., interpretability is enhanced by improving the transparency (Zhang et al. 2014a). Decision tree is a basic transparent model, which characterizes each node by explicit criteria. Wang et al. (2018b) proposed a tree-enhanced embedding model (TEM) for explainable recommendation, combining the generalization ability of embedding-based models and explainability of tree-based models. Tao et al. (2019) introduced the FacT model to integrate regression trees to learn latent factors and use learnt tree structure to explain recommendations. Specifically, the model predefined aspects to construct template-based explanations. Bauman, Liu, and Tuzhilin (2017) proposed the Sentiment Utility Logistic Model (SULM), which extracted aspects and the corresponding sentiments based on user reviews. Besides, some works achieve interpretability by adopting a multi-rating approach to capture the relationship between various dimensions of criteria and final ratings. Zheng (2017) introduced "Criteria Chains" to establish the combination of criteria by contextual situations. Fan et al. (2023) proposed a collective model to predict final ratings by jointly learning users' sub-scores over multi-criterion (service, locations, etc.), which could explain user preference features more naturally by presenting sub-scores of various criteria. However, these models require more fine-grained information, such as the user sub-scores for multi-criterion. In addition, they impose higher demands on the accuracy of sub-scores, which means inaccuracy in any criterion will impact the final rating, resulting in overall poor performance (Anwar, Zafar, and Iqbal 2023).

Order Effects Theory. Order effects can be commonly observed and captured where information is presented as a sequence, such as the objects presented in psychology experiments (Anderson 1965), the evidence presented in court (Maegherman et al. 2020), and the items presented in recommender systems (Zhao et al. 2021). The Order Effects Theory claims that different sequence orders can cause different consequences (Petty et al. 2001). It has been commonly used in research, and the studies concentrate on two topics: 1) how people perceive and process information in order to make decisions; and 2) how to present information in order to get people's acceptance.

For the first research topic, several experimental psychology studies investigate the influence of order effects in human memory (Ebbinghaus 2013). The results show the im-

*https://www.yelp.com/dataset

portance of ordering in the decision-making process, and indicate that the first and the last items are easier to remember than the middle ones, recognized as Primacy and Recency (Gershberg and Shimamura 1994). Also, Wu et al. (2017) mentioned a specific form of interpretability known as human simulability, and used sequences as the input of recurrent neural network (RNN) to construct a human-simulatable model, simulating the decision-making process of humans. For the second topic, Maegherman et al. (2020) demonstrated that the order of evidence presentation affected the cognitive dissonance and confirmation bias, which in turn affects the ability to persuade. Petty et al. (2001) carried out experiments to illustrate that the participants with different levels of motivation showed various susceptibility to order sequence, especially for the Primacy Effects. However, the inherent logical mechanism and mode of action of such decision-making models are not transparent.

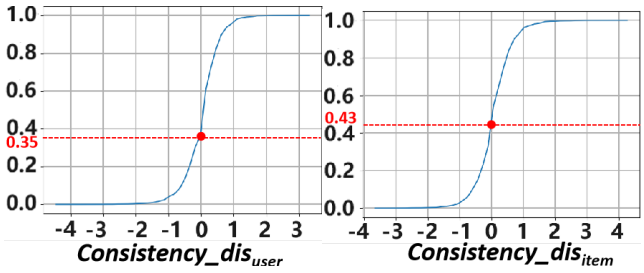
Furthermore, recent researches have indicated that the recommendation process is a way of persuasion (Gretzel and Fesenmaier 2006). Focusing on the influence of order effects for recommendation scenario, Felfernig et al. (2007) stated that the serial position could influence users' degree of acceptance, especially for the information at the beginning and the end of the sequence (Fogg 1998). Zhao et al. (2021) observed that users showed different interests for different orders of item sequence. Therefore, there could be an optimal order corresponding to the user's best acceptance.

To conclude, although Order Effects Theory has been investigated in several areas, the ordering relationship of attributes has not been considered in recommendations. To our knowledge, this is the first work to capture order effects within decision-making process in explainable recommender systems. By incorporating the Order Effects Theory into the construction of the model-intrinsic approach, we can enhance the interpretability and also improve performance.

Preliminary Analysis

In this section, we explore the applicability of the Order Effects Theory within the realm of recommendation, analyzing the ordering relationship among user concerns in reviews. The analysis is conducted on the Yelp dataset, which is a publicly available user review dataset and commonly used for recommendation tasks. The statistics of the dataset after preprocessing are shown in Table 1. Following the sentiment analysis toolkit, named Sentires, built by Zhang et al. (2014b), the triplets (Aspect, Opinion, Sentiment) were extracted from the reviews to represent each datum, wherein Aspect is the focus in our analysis. This phrase-level sentiment analysis toolkit is a widely used tool for aspect analysis in recommendation system due to its accuracy and ease of use in capturing aspect and sentiment (Zhang, Chen et al. 2020). Based on Yelp dataset, we present a scenario where users make decisions regarding restaurants, highlighting the *existence* of order effects and reveal its *difference* among users/items. This comprehensive analysis yields valuable insights into the applicability of the Order Effects Theory.

We show the following randomly picked review from the Yelp dataset (by user *Hie3_7_Nan3R3HaygoCFvA* for item *VitNqJm8DIjw5D-Q-aiENQ*). For brevity, we display only



(a) Among Users (b) Among Items

Figure 1: The $Consistency\_dis_{user/item}$ value among users/items. The horizontal axis represents the proportion of users/items under specific consistency difference.

key sentences containing the filtered extracted aspects (highlighted in bold). Aspect order can be considered as the order of the aspects that appeared as the sequence of the user's decision-making process (Arapoff 1967).

> *I've been here twice now, once for dinner and another time for the **lunch** buffet. The **food** is quite good, with nice hits of spice and savory **flavors**. The **prices** are average.*

As shown in the review, the user initially emphasizes the aspect of **lunch** for mealtime, followed by the restaurant's **food** and **flavor** to evaluate taste. The aspect of **price** is addressed later, leading to the final experience. The decisive chain can be identified as: {**lunch**, **food**, **flavor**, **price**} (denoted as O_{11}). Similarly, another review for the same above user can be randomly picked and displayed as: {**lunch**, **food**, **price**} (denoted as O_{12}). Further, for a different randomly selected user (user_id is *-_FCaLa5eYXedOotc7J18Q*), one aspect order can be extracted as {**food**, **service**, **food**, **chicken**, **service**, **table**, **spot**} (denoted as O_{21}). It is obvious to see that the consistency of the two aspect orders for the same user (O_{11} and O_{12}) is much higher than that for different users (O_{11} and O_{21}), suggesting that the intra-consistency (within one user) of aspect order is higher than the inter-consistency (between different users).

We further aim to generalize the above findings by randomly selecting 10,000 users from Yelp and constructing 10,000 pairs of intra-user reviews (two reviews from same user) and 10,000 pairs of inter-user reviews (two reviews from two different users). For each pair of intra-user or inter-user reviews, we utilize the NDCG metric (Kanoulas and Aslam 2009) (the details of NDCG are described in Experiment Section) to evaluate the consistency of aspect order between the two reviews (named $intra\_cons_{user}$ or $inter\_cons_{user}$, respectively). To highlight the difference between a user's $intra\_cons_{user}$ and $inter\_cons_{user}$, we construct a measurement named $consistency\_dis_{user}$ which is computed as $intra\_cons_{user}$ minus $inter\_cons_{user}$. It can reflect whether a user has a specific order preference. The CDF (Cumulative Distribution Function) of $consistency\_dis_{user}$ among the 10,000 users is shown in Figure 1 (a). From the figure, we can see the $intra\_cons_{user}$ of nearly 70% users is greater than their corresponding $inter\_cons_{user}$ ($p < 0.001$), indicating that most users have specific aspect order considera-

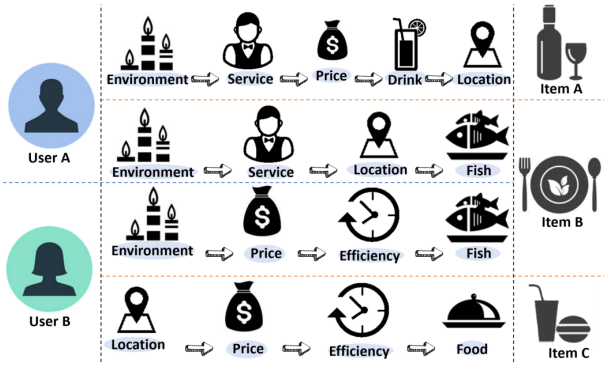


Figure 2: Example of existence and variance of order effects.

tions. We can define such users as *Sensitive Users*. Moreover, it suggests that a proportion of users are *Strong Sensitive Users*, i.e., the $intra\_cons_{user}$ is much higher than the corresponding $inter\_cons_{user}$. For example, about 20% of users' $consistency\_dis_{user}$ values are higher than 0.5 ($p < 0.001$). These results confirm our above observation, i.e., for most users, the $intra\_cons_{user}$ is higher than the $inter\_cons_{user}$, suggesting a user tend to have aspect order preference while different users exhibit some differences.

Similarly, we conduct an analysis from the perspective of items, i.e., randomly selecting 10,000 items and constructing 10,000 pairs of intra-item reviews (two reviews from the same item) and 10,000 pairs of inter-user reviews (two reviews from two different items). After that, $intra\_cons_{item}$ (the consistency of aspect order between two intra-item reviews), $inter\_cons_{item}$ (the consistency of aspect order between two inter-item reviews), and $consistency\_dis_{item}$ ($intra\_cons_{item}$ minus $intra\_cons_{item}$) are evaluated. The result is shown in Figure 1 (b). It shows that for over 60% items, $intra\_cons_{item}$ is overall higher than the corresponding $inter\_cons_{item}$, suggesting that an item tends to have a specific aspect order considered by users and different items have some differences ($p < 0.001$). Additionally, approximately 20% items are strong sensitive with the $consistency\_dis_{item}$ value higher than 0.5 ($p < 0.001$).

Based on the above analysis, we confirm the existence and variances of order effects. Specifically, each user prefers a specific aspect order, and each item also tends to have a specific order considered by users (Existence). However, a user's aspect order preference can vary across different items, and an item's aspect order considered by different users can also vary to some degree (Variance). We use an example to illustrate the conclusion (See Figure 2). When user A is selecting a restaurant for himself, he generally tends to prioritize **Environment** and **Service**, while other aspects like **Location** and **Price** are considered occasionally, which indicates his overall aspect order preference and some variance across items. Similarly, an item like item B tends to be judged by **Environment** due to its remarkable dining environment. Beside, different users may evaluate B in different aspect order. It suggests item B's overall specific aspect order considered by users and the variance across users.

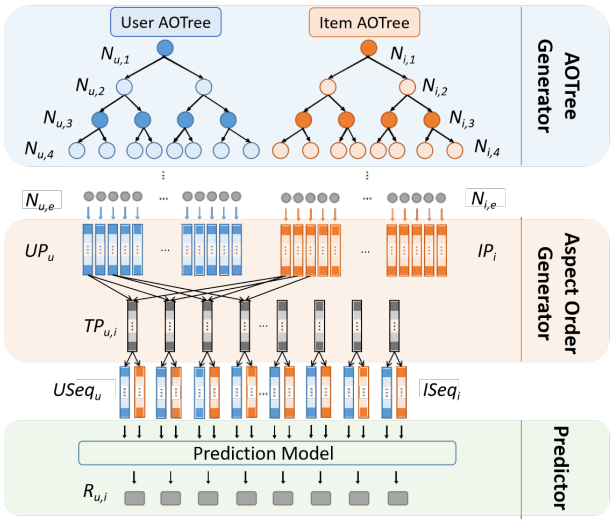


Figure 3: The algorithmic framework of AOTree.

Methodology

Then, we introduce our proposed Aspect Order Tree-based explainable recommendation method — AOTree, which aims to simulate human decisive behavior according to Order Effects Theory. We first present the Problem Formulation followed by the general architecture shown in Figure 3, which consists of three key modules: 1) the AOTree Generator, 2) the Aspect Order Generator, and 3) the final Predictor.

Problem Formulation

In this work, we target at the **rating prediction** task. Let $\mathcal{U} = \{u_1, u_2, ..., u_m\}$, $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ denote the user set and item set, respectively, where m and n represent the number of users and items. Then, $r_{i,j}$ is the rating value for a user u_i toward an item v_j. And a set of observed data is represented as $\mathcal{O} = \{(u, v) | u \in \mathcal{U}, v \in \mathcal{V}, u \textit{ has reviewed } v\}$.

The construction of sentiment set, represented by triplets (Aspect, Opinion, Sentiment), follows the same procedure in the Data Analysis. Suppose we finally get the aspect set $\mathcal{A} = \{a_1, a_2, ..., a_l\}$, where l denotes the number of aspects. Then, we construct the user aspect matrix $X \in \mathbb{R}^{m*l}$ and the item aspect matrix $Y \in \mathbb{R}^{n*l}$ as (Zhang et al. 2014a):

$$X_{i,k} = \begin{cases} 0, \text{ if } u_i \text{ did not mention aspect } a_k, \\ 1 + (N-1)(\frac{2}{1+exp(f_{-i,k})} - 1), \text{ otherwise.} \end{cases}$$

$$Y_{j,k} = \begin{cases} 0, \text{ if aspect } a_k \text{ was not mentioned for item } v_j, \\ 1 + \frac{(N-1)}{1+exp(-f_{j,k} \cdot s_{j,k})}, \text{ otherwise.} \end{cases}$$

N denotes the rating scale, which is 5 in our datasets. $f_{j,k}$ is the frequency of u_i mentioning aspect a_k, and $s_{j,k}$ is the average sentiment calculated from the sentiment set. Since $X_{i,k}$ and $Y_{j,k}$ measure the importance of aspect a_k to user u_i and item i_j, we thus call them "aspect importance" and X and Y are called **individual aspect importance matrices**.

Group Aspect

We first process the user/item aspect matrix into a general representation for all users/items rather than construct trees for each user/item in order to reduce the time complexity. We take the user AOTree generation as an example, the process of which is the same as item AOTree generation.

However, getting mean value of each aspect from individual user aspect importance matrix X ignores contribution proportions because the size and effectiveness of each review are different among users. Inspired by the weighted technique (Chen et al. 2018), we integrate aspect matrix by considering user contribution, weighing preference value for each aspect. The formulation can be described as follows:

$$\hat{X}_k = \sum_{i=1}^{m} X_{i,k} * W_i, \quad W_i = \frac{N_i}{\sum_{i=1}^{m} N_i}, \tag{1}$$

where $\hat{X}_k$ represents the general (globally weighted) aspect importance on aspect a_k over all users ($\hat{X}$ is called **general aspect importance vector**). And W_i is the portion of contribution from user u_i, calculated by N_i, the number of reviews written by user u_i. We could finally get the general user aspect importance vector $\hat{X} = \{\hat{X}_1, \hat{X}_2, ..., \hat{X}_l\}$ for all aspects. Similarly, the general item aspect importance vector $\hat{Y} = \{\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_l\}$ can be obtained.

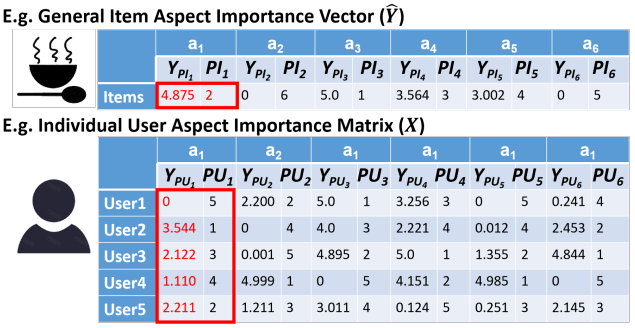
AOTree Generator

The tree-based model is considered to be explainable and transparent when applied to recommendation tasks (Zhang, Chen et al. 2020). Also, it is easy for humans to simulate and understand (Wu et al. 2017). Rather than learning features for prediction like other deep learning methods, the tree-based model learns decision rules from datasets. However, the traditional decision tree aims to solve classification problems, which is not applicable. So, we design the decision rules to split items/users according to the matching degree within aspect quality/preference to find an aspect order representing the considering decision process of users.

Due to the variances of aspect order effects suggested in the Preliminary Analysis, we personalize the order construction for users and items. Specifically, when building the User-AOTree, we could split user sets and get the considering aspect sequence for each user by matching the general item aspect importance vector $\hat{Y}$ with the individual user aspect importance matrix X, meaning how the sequence of general item quality is shown according to different users. The Item-AOTree can be constructed in the same way with the corresponding general user aspect importance vector $\hat{X}$ with the individual item aspect importance matrix Y.

In general, inspired by the decisive process of Markov Decision Processes (MDP), the goal is to maximize its reward, and thus, each decisive step is chosen based on the best current state (Shani et al. 2005). In our scenario, each step of our aspect order is defined as the chosen aspect, which is based on the minimal Split Expense for the optimal Split Value. Specifically, we follow three steps:

1. Calculate Split Value (SV) for each aspect, representing the optimal split criteria to match user preference and item characteristic corresponding to each aspect;



	a_1		a_2		a_3		a_4		a_5		a_6	
	Y_{PI_1}	PI_1	Y_{PI_2}	PI_2	Y_{PI_3}	PI_3	Y_{PI_4}	PI_4	Y_{PI_5}	PI_5	Y_{PI_6}	PI_6
Items	4.875	2	0	6	5.0	1	3.564	3	3.002	4	0	5

	a_1		a_2		a_1		a_1		a_1		a_1	
	Y_{PU_1}	PU_1	Y_{PU_2}	PU_2	Y_{PU_3}	PU_3	Y_{PU_4}	PU_4	Y_{PU_5}	PU_5	Y_{PU_6}	PU_6
User1	0	5	2.200	2	5.0	1	3.256	3	0	5	0.241	4
User2	3.544	1	0	4	4.0	3	2.221	4	0.012	4	2.453	2
User3	2.122	3	0.001	5	4.895	2	5.0	1	1.355	2	4.844	1
User4	1.110	4	4.999	1	0	5	4.151	2	4.985	1	0	5
User5	2.211	2	1.211	3	3.011	4	0.124	5	0.251	3	2.145	3

Figure 4: Example of calculating Split Value (SV).

2. Calculate Split Expense (SE) for each aspect with its corresponding SV, representing the reward for each aspect;
3. Choose aspect and corresponding SV with the minimal SE value, representing the optimal decision step.

For each node, we repeat the above three steps to split sets into different nodes until the node contains only one element, recognized as the leaf. However, in order to improve efficiency, we limit the depth of the tree as a pre-defined threshold, meaning that the leaf node may contain more than one element but with very similar characteristics.

Calculate Split Value (SV) We first traverse all aspects to calculate the split value for each aspect, representing the matching degree between user and item. Rather than directly using values in the general item aspect importance vector ($\hat{Y}$) to split nodes, which may cause biases among different users, we focus on the degree of ranking matching between users and items on each aspect (Nevo 1989).

In detail, for each aspect, we first get the item quality rank position ($PI_k = 1, 2, ..., l$) according to the general item aspect importance vector ($\hat{Y}$) and find the corresponding user rank position (PU_k) in the individual user aspect importance matrix (X) for the same aspect (a_k). PI_k and PU_k satisfy the Equation 2, which means the item's quality rank matches the user's preference rank.

$$PI_k/l = PU_k/m. \tag{2}$$

as the values of PI_k, m and l are already known, the corresponding user rank position (PU_k) could be calculated.

Then, the Split Value (SV), the value in the individual user aspect importance matrix (X) at the corresponding position, could be obtained by the following equation inspired by the matching-ranking technique (Nevo 1989):

$$\frac{PU_k - PU_{kl}}{PU_{kr} - PU_{kl}} = \frac{SV_k - X_{P_{kl},k}}{X_{PU_{kr},k} - X_{PU_{kl},k}}. \tag{3}$$

as PU_k is obtained as a decimal, PU_{kr} and PU_{kl} represent the right and left integer rank position index of PU_k ($PU_{kr} > PU_k > PU_{kl}$). And $X_{PU_{kr},k}$ and $X_{PU_{kl},k}$ are corresponding values in individual user aspect importance matrix (X) for aspect a_k. As SV_k is the only unknown value in Equation 3, it can be easily derived given other values.

The process can be illustrated as a specific example in Figure 4 for a better understanding. Take aspect a_1 as one of the traversals, so the k is 1 in this example. Then, the value

for PI_1, l and m are 2 (item quality rank position for a_1), 6 (number of aspects) and 5 (number of users), and we can obtain $PU_1 = m * PI_1/l = 5 * 2/6 = 1.667$ (according to Equation 2). The corresponding value for PU_{1r}, PU_{1l} are 2 and 1, then $X_{PU_{1r},k}$ and $X_{PU_{1l},k}$ are 3.544 and 2.211, respectively corresponding to $User2$ and $User5$. And the final SV value for aspect a_1 is $SV_1 = \frac{(1.667-1)}{(2-1)} * (3.544 - 2.211) + 2.211 = 4.433$ (according to Equation 3). We could get SV value for all aspects in the similar way.

Calculate Split Expense (SE) Then, we decide which aspect should be chosen as the node split criterion and split users by comparing the corresponding aspect value in the individual user aspect importance matrix (X) with SV based on the chosen aspect. Take a_k as an example, if the aspect importance value for the user (X_{ik}) is larger than SV_k, the user is split into the right node, otherwise, into the left node. Our goal is to find one aspect could make the best matching between user preference and item quality, simultaneously ensuring the robustness of the decision tree. Targeting the rationality and robustness of the decision tree construction, we specify the split principles into three components, SE_1, SE_2 and SE_3 aiming at child nodes after splitting the node. The division principles refer to classification and clustering methods, such as Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), aiming to enhance intra-class similarity while promoting inter-class dissimilarity (suppose a_k is the split criterion for the current node):

1) SE_1: users in the same child node should be similar on aspect a_k, i.e., with minimal distance:

$$\min SE_1 = \sum_{i=1}^{m} |X_{ik} - \bar{X}_k|, \tag{4}$$

where $\bar{X}_k$ and X_{ik} are the average aspect importance value in the individual user aspect importance matrix (X) and the specific value for user i, both on aspect a_k.

2) SE_2: users in the different child node groups should be different on aspect a_k, i.e., with maximal distance:

$$\max SE_2 = |\bar{X}_{right,k} - \bar{X}_{left,k}|, \tag{5}$$

where $\bar{X}_{right,k}$ and $\bar{X}_{left,k}$ are the average aspect importance values for aspect a_k in the right and left child node.

3) SE_3: users in the same child node group should be different on aspects except a_k, to continually split users:

$$\max SE_3 = \sum_{o=1}^{l-k} \sum_{i=1}^{m} |X_{io} - \bar{X}_o|, \tag{6}$$

where X_{io} is the aspect importance for user i on aspect except a_k, and $\bar{X}_o$ is the average aspect importance for all users on aspect a_o.

Then, we could get the Split Expense (SE) value by traverse all aspects according to Split Value (SV) as follows:

$$SE = NV * SE_l * SE_r, \tag{7}$$

where NV represents the normalization constant to normalize the value of SE, and we set it to $10^{N_r * N_l}$, where N_r and N_l are the number of users in each child node. The subscripts "r" and "l" denote the values of the right and left child nodes. SE_l and SE_r are Split Expense (SE) for the left and right nodes after splitting, which could be obtained by:

$$SE_l = \frac{SE_{l1}}{SE_{l2} * SE_{l3}}, \quad SE_r = \frac{SE_{r1}}{SE_{r2} * SE_{r3}}. \tag{8}$$

Choose Aspect and Split Value We finally choose aspect with minimal Split Expense (SE) in order to satisfy our goal:

$$k = \arg\min_k \ SE. \tag{9}$$

Then aspect a_k is the split aspect for this node, and corresponding Split Value (SV) can be obtained by Equation 3.

Build AOTree We continue the steps to decide on each node. The criteria for stopping the division is until each node has only one item or the tree depth has reached the predefined threshold. The obtained User-AOTree represents how the sequence of general item quality is shown according to different users. Then we could construct aspect order for certain user sets. In each node, users with similar aspect quality share aspect order (e.g., $UP_i = \{N_1, N_2, ..., N_e\}$ for user u_i, where N and e denote the aspect id and the path length).

Similarly, the Item-AOTree can be built following the same process (e.g., aspect order is $IP_j = \{N_1, N_2, ..., N_e\}$ for user v_j), representing how sequence of general user preference is shown according to different items.

Aspect Order Generator

For each interaction, we convert it into aspect order UP_i/IP_j for u_i/v_j according to the User-AOTree/Item-AOTree. The two orders are considered separately from user and item sides and we combine them by a simple average:

$$Ind_k = \frac{Ind_{UP,k} + Ind_{IP,k}}{2}, \tag{10}$$

where $Ind_{UP,k}$ and $Ind_{IP,k}$ denote the ranking index of aspect a_k in UP_i and IP_j. Then, the average aspect index Ind_k can be sorted to construct the final aspect sequence $TP = \{TP_1, TP_2, ..., TP_e\}$, and each element is id for aspect, representing the decisive aspect process for certain user toward certain item. We fill each sequence into a fixed-length e with random aspect id if the length is less than e.

Corresponding to the aspect id sequence TP, we could get the final user or item aspect importance sequence $USeq$ or $ISeq$ by picking the value of the corresponding aspect id from the original individual user/item aspect importance matrix (X/Y). For example, the value for $USeq = \{USeq_1, USeq_2, ..., USeq_e\}$ is picked from X_i based on the aspect id in TP, which means, $USeq_i$ is equal to X_{i,TP_i}.

Prediction with Aspect Order

The final rating prediction result is obtained mainly by two components: decision process information, which is the core part of our method, and user/item context information. In the learning process, we refer to the traditional latent factor method, but extend with our designed components: applying position embedding and self-attention layer to the sequence, and then final ratings are obtained with a linear model.

Position Embedding Layer Given the aspect order generated by AOTree (TP), which is represented by aspect id, we project each id into an embedding vector with size d. Then, we could obtain the path embedding $\hat{TP} = \{\hat{TP}_1, \hat{TP}_2, ..., \hat{TP}_e\} \in \mathbb{R}^{e \times d}$.

In order to highlight the position of each aspect in the sequence, we consider combining our sequence with a position

embedding vector. Inspired by Kang and McAuley (2018), we add a position embedding $E = \{E_1, E_2, ..., E_e\} \in \mathbb{R}^{e \times d}$ to the path embedding $\hat{TP}$ to get $\mathcal{N}$, described as:

$$\mathcal{N} = \begin{bmatrix} \hat{TP}_1 + E_1 \\ \hat{TP}_2 + E_2 \\ ... \\ \hat{TP}_e + E_e \end{bmatrix}. \quad (11)$$

Self-Attention Layer We use scaled dot-product attention to achieve self-attention as (Waswani et al. 2017):

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \quad (12)$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ denote queries, keys and values, respectively. The scaling factor $\sqrt{d}$ is to avoid overly large values of the inner product, especially when the dimensionality is high, and $\mathbf{K}$ and $\mathbf{V}$ are with the same values. In our method, $\mathcal{N}$ is converted to the three input for self-attention as:

$$Att = Attention(\mathcal{N}\mathbf{W}^Q, \mathcal{N}\mathbf{W}^K, \mathcal{N}\mathbf{W}^V), \quad (13)$$

where $\mathbf{W}^Q$, $\mathbf{W}^K$ and $\mathbf{W}^V$ are projection matrices. Then, the result is added to the normalized input $\mathcal{N}$ to obtained the Sequence Feature (SF) as:

$$SF = LayerNorm(\mathcal{N} + Att). \quad (14)$$

Due to the order consideration, we should mask the first $t-1$ aspects in the order when processing the t-th layer of aspect, so we add a triangular matrix for masking the unseen position. Layer normalization is used to normalize the inputs, which is beneficial for stabilizing and accelerating neural network training, represented as:

$$LayerNorm(x) = \alpha \odot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta, \quad (15)$$

where μ and σ represent mean and variance of $\mathcal{N}$, α is scaling factor, β is bias term, and $\odot$ is element-wise product.

Finally, we optimize aspect order for user ($USeq$) and item ($ISeq$) by element-wise product with SF as:

$$\begin{aligned} \hat{USeq} &= USeq \odot SF. \\ \hat{ISeq} &= ISeq \odot SF. \end{aligned} \quad (16)$$

Prediction Layer In the prediction layer, we apply a simple linear model to integrate all the components:

$$\hat{r}_{ij} = W_1^T(\hat{USeq}_i \odot \hat{ISeq}_j) + W_2^T(p_i \odot q_j) + b_u + b_i + \mu, \quad (17)$$

where $W_1 \in \mathbb{R}^d$ and $W_2 \in \mathbb{R}^d$ are weight matrices of the linear model, and p_i and q_j denote the ID embedding for user u_i and item v_j. $\hat{USeq}_i \odot \hat{ISeq}_j$ is the decision process information, and $p_i \odot q_j$ represents the interaction between user u_i and item v_j.

Model Training

We use Mean Square Error (MSE) as our objective function:

$$\textbf{MSE} = \frac{1}{N}\sum_{(i,j)\in\mathcal{O}}(r_{ij} - \hat{r}_{ij})^2, \quad (18)$$

where r_{ij} denotes the ground truth rating for user u_i on item v_j, and $\hat{r}_{ij}$ is the predicted rating from our method.

Experiment

In this section, we present experimental setup, comprehensive experimental results and detailed analysis, aiming at answering the following research questions (RQs):

- **RQ1:** Does AOTree outperform the state-of-the-art methods in terms of recommendation accuracy?
- **RQ2:** Does the considering aspect order obtained by AOTree help to improve recommendation accuracy?
- **RQ3:** Does the considering aspect order obtained from AOTree help to improve explainability?

Experimental Setup

Dataset. The experiments are conducted on five review datasets from Amazon and Yelp. For Amazon, we chose four common datasets: Cells Phones and Accessories, Office Product, Patio, Lawn and Garden and Digital Music.

We adopt the sentiment analysis toolkit (Zhang et al. 2014b) and keep users/items with more than T_u/T_i reviews due to dataset sparsity (Pan et al. 2022). As we focus on aspects extracted from reviews, we set T_a to the minimal number of occurrences for each aspect. The specific value of threshold and the statistics of the datasets after preprocessing are shown in Table 1. The Density value denotes the average number of aspects mentioned by users. It is worth noting that the T_u/T_i for Digital Music and Yelp are set differently in order to keep a relatively consistent density value. In the experiments, each dataset is divided into training set, validation set, and test set by 8:1:1.

Compared Methods. To validate the performance of our method, several baselines are selected. Specifically, we choose commonly-used traditional and state-of-the-art methods in the field of recommendations, such as the classic benchmark method (MF), tree-based methods (XGBoost and TEM), review-based method (NARRE, R3), and aspect-based methods (SULM, ANR and ERRA). All the compared methods are evaluated based on the original papers with careful hyper-parameter tuning. The compared methods are:

- **MF** (Koren, Bell, and Volinsky 2009): Matrix Factorization (MF) is a classic rating prediction method using bias terms and latent features for prediction.
- **XGBoost** (Chen and Guestrin 2016): XGBoost is the state-of-the-art tree-based method. The final tree captures complex feature dependencies, that is the cross feature combination for different paths.
- **TEM** (Wang et al. 2018b): Tree-enhanced Embedding Method (TEM) is also a tree-based model. It is constructed into two cascade parts with GBDT and an easy-to-interpret attention network, making the recommendation process fully transparent and explainable.
- **NARRE** (Chen et al. 2018): Neural Attentional Regression model with Review-level Explanations (NARRE) is a widely used state-of-the-art explainable recommendation method. This method focuses on the reviews usefulness and uses an attention mechanism to learn the importance weights over different reviews.

Dataset	#Aspect	#Users	#Items	#Reviews	T_u/T_i	T_a	Density
Cells Phones and Accessories	1,016	14,782	10,825	109,956	20	20	44.80
Office Product	433	2,433	2,074	27,091	20	20	44.63
Patio, Lawn and Garden	399	1,137	1,103	9,926	20	20	41.48
Digital Music	215	4,392	3,247	44,595	5	20	32.34
Yelp	28	19,225	13,083	81,777	3	5	5.05

Table 1: The statistics of the five datasets used in the experiments.

Model	Cells Phones and Accessories	Office Product	Patio, Lawn and Garden	Digital Music	Yelp
MF	1.2184	0.8716	1.1089	1.1851	1.3994
XGBoost	1.0127	0.6980	1.0101	0.8589	1.2736
NARRE	0.9967	0.6902	1.0292	0.8620	1.2535
R3	0.9847	0.6882	0.9760	0.8278	1.2498
ANR	0.9801	0.6810	1.0126	0.8189	1.2503
TEM	0.9837	0.6771	0.9939	0.8247	1.2686
ERRA	0.9811	0.7066	1.0270	0.8347	1.2364
AOTree	**0.9719***	**0.6729***	**0.9452****	**0.8050****	**1.2140***

Table 2: Overall Performance on MSE on benchmark datasets. The best performance is in boldface. * and ** represent significant difference at 0.05 and 0.01 level, respectively, compared with the best baseline.

- **R3** (Pan et al. 2022): Recommendation via Review Rationalization (R3) is a causal-aware explainable method that extracts rationales from reviews via a rationale generator to alleviate spurious correlations in recommendation. The final recommendation and causal-aware explanation can be generated according to the rationales.
- **SULM** (Bauman, Liu, and Tuzhilin 2017): Sentiment Utility Logistic Model (SULM) is a basic aspect-based model, which uses sentiment analysis of user reviews to identify the most valuable aspects for recommendation.
- **ANR** (Chin et al. 2018): Aspect-based Neural Recommender (ANR) is an end-to-end attention-based method, which performs aspect-based representation learning for both users and items via attention mechanism.
- **ERRA**(Cheng et al. (2023)): Explainable Recommendation by personalized Review retrieval and Aspect learning (ERRA) obtains recommendations and explanations by additional information from retrieval enhancement.

Evaluation Metrics. AOTree and all baselines can be evaluated by Mean Square Error (MSE), except SULM. For MSE, described as Equation (18), a lower value means better performance. To compare with the aspect-based algorithm, SULM, which converts the rating prediction problem into a preference classification task for 0 and 1, we convert our scenario into a ranking task. Specifically, we sort the recommended items into a list according to ratings for each user and evaluate the ranking performance using NDCG. NDCG, short for Normalized Discounted Cumulative Gain, is a metric commonly used in recommender systems to assess the quality of ranked recommendations (Kanoulas and Aslam 2009). The formulas are as follows:

$$NDCG@K = \frac{DCG@K}{IDCK@K} \tag{19}$$

$$DCG@K = \sum_{j=1}^{K} \frac{2^{rel_j}-1}{log_2(j+1)} \tag{20}$$

$$IDCG@K = \sum_{j=1}^{K} \frac{2^{rel_j}-1}{log_2(j+1)} \tag{21}$$

where K denotes the position up to which the relevance is accumulated, and it is omitted and default to 5. rel_j denotes the graded relevance at position j, and the set consisting of the top K results is taken. Then, the final $NDCG$ is calculated by normalizing DCG as Equation (19), and a larger value means better performance.

The goal of our method is to achieve human simulability, that is, to obtain considering aspect order as explanations. As the order of aspects in reviews appears as a decision process, we compare it with the predicted aspect order to evaluate the explanation. On the one hand, we verify aspect coverage to show performance from the importance dimension ($Num\%$). On the other hand, inspired by ranking metrics, we use $NDCG$ and $F1$ metrics to test ordering effectiveness. Specifically, we assess aspect composition order's quality at the aspect level using $NDCG$, and coverage using $F1$. According to the "Order Effects Theory" introduced above, the Primacy and Recency principles are crucial to order performance. However, due to the fixed tree depth, we only test the front parts of the sequence to illustrate the Primacy principle, focusing on the first five aspects.

Implementation Details. For the compared algorithms, we follow the corresponding papers to initialize hyperparameters and carefully tune them for optimal performance. We use Adaptive Moment Estimation (Adam) (Kinga, Adam et al. 2015) to optimize our method, with learning rate in {0.00001, 0.0005, 0.001, 0005}, L2 regularization coefficient in $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ and dropout rate in {0, 0.2, 0.4, 0.6, 0.8}. We use early stopping with 10 epochs and report results from the best-performing model on validation sets. Other hyperparameters are searched with batch size in {16, 32, 64, 128}, latent factors number in the range [5, 20]. To control the expense of constructing AOTree, the max tree depth is limited to {5, 10, 15,..., 100} and the construction process requires no training.

Model	Cells Phones and Accessories	Office Product	Patio, Lawn and Garden	Digital Music	Yelp
SULM	0.2576	0.3743	0.3456	0.3241	0.2363
AOTree	**0.2691***	**0.3771***	**0.3467***	**0.3317***	**0.2490***

Table 3: Overall Performance on NDCG on benchmark datasets. The best performance is in boldface. * and ** represent significant difference at 0.05 and 0.01 level, respectively, compared with the best baseline.

Dataset	Overall	Sensitive Users			Non-Sensitive Users		
		Basic	Shuffle	Top@5	Basic	Shuffle	Top@5
Cells Phones and Accessories	0.9719	0.7355	0.7532	0.7628	1.2042	1.1708	1.1977
Office Product	0.6729	0.5478	0.5763	0.5612	0.7924	0.7820	0.7901
Patio, Lawn and Garden	0.9452	0.8973	0.9536	0.9446	1.04564	1.0156	1.0382
Digital Music	0.8050	0.5888	0.6102	0.6379	0.8743	0.8691	0.8690
Yelp	1.2140	0.9774	0.9923	0.9879	1.4266	1.3234	1.3078

Table 4: Effectiveness of aspect order on the AOTree model measured by MSE. $Basic$ means the original results by only using (Non-) Sensitive Users. $Shuffle$ and $Top@5$ denote the two disturbance operations applied to the aspect order.

Results

Overview Performance (RQ1)

From Table 2 and Table 3, AOTree significantly outperforms all other competitors on all datasets ($p < 0.05$), which could answer RQ1. R3 ranks second due to the dependence on causal relations rather than spurious correlations. NARRE and ANR also show quite good performance, especially in Cell Phones and Accessories and Office Product datasets, demonstrating attention mechanism can capture implicit finer-grained properties from reviews. TEM captures explicit cross features but ignore relationships from "ordering" dimension, and only shows a slight advantage in Office Product. By exploring the "ordering" of aspects, AOTree leverages sequence property and shows promising performance. From NDCG results compared with SULM, the AOTree method still significantly outperforms on all datasets ($p < 0.05$), demonstrating its superiority tasks.

Besides, we have several findings from the perspective of datasets. For Patio, Lawn and Garden dataset, AOTree shows the best improvement, nearly 3.08% ($p < 0.01$) better than the second place method, TEM, which means most benefiting from order effects phenomenon. In other words, the reviews written by users in this dataset well display such writing rules, and the reason may be that these users take writing reviews more seriously, with a process of decision-making.

Effectiveness of the Aspect Order (RQ2)

We perturb the obtained aspect order to demonstrate the superiority of the current order. However, the results are not stable for the whole dataset, so we split users according to the MSE according to Table 2. To be specific, we focus on the samples with MSE less than the average value in the training set. If all interactions for one user meet the above criterion, the user will be selected. In fact, this is a very strict criterion, and the selected users should have strong order effect. So we mark them as *Identified Strong Sensitive Users* while others as *Identified non-Strong Sensitive Users*. Our analysis suggests the fraction of *Identified Strong Sensitive Users* reaches 20% to 30% in different datasets (20.37% in Cells Phones and Accessories, 26.61% in Office Product, 21.60% in Patio, Lawn and Garden, 21.05% in Digital Music, and 31.57% in Yelp). Such proportion is close to that of the *Strong Sensitive Users* with $consistency\_dis_{user}$ higher than 0.5 in our Preliminary Analysis, indicating our method's effectiveness in capturing order effect.

Then, for each user, we verify the effectiveness of the captured aspect order for each interaction by adopting two perturbation operations related to "ordering": 1) shuffling the whole order to verify whether the sequence generated by the model is effective, and 2) replacing the first five aspects with random ones to see whether the Primacy principle contributes. The final results are shown in Table 4. From the results, we observe the following findings. For the *Identified Strong Sensitive Users*, the results after perturbation show higher MSE values, suggesting the superiority of the aspect order obtained by our method. For the *Identified non-Strong Sensitive Users*, the conclusion seems to be dismissed since the results after perturbation show no significant difference.

Explainability of Aspect Order (RQ3)

As we seek for a specific form of interpretability known as human simulability, we finally verify the explainability of the AOTree model according to the simulability, that is, to attain the matching degree between the obtained aspect order and the original aspect sequence that appeared in the reviews. We compare the importance and order of the first five aspects to prove that our method could predict users' decision process displayed as reviews. The importance is represented by calculating the coverage of the aspects, and the order is verified by numerical metrics of NDCG and F1.

We choose TEM, SULM and ERRA as the baselines in the explanability evaluation because they also focus on using aspects as explanations. The aspect order of TEM and SULM is constructed according to the rank of attention weights for each aspect, while ERRA's aspect order is extracted from aspects appearing in generated explanations. The results are shown in Table 5. For $Num\%$ value, AOTree significantly outperforms the other three methods, indicating its advantage in more accurately discovering key factors, especially for Yelp ($p < 0.01$). And for $NDCG@5$ and $F1@5$ metrics, although TEM performs better on Cells Phones and

Model	Cells Phones and Accessories			Office Product			Patio, Lawn and Garden			Digital Music			Yelp		
	Import	Order		Import	Order		Import	Order		Import	Order		Import	Order	
	Num%	NDCG@5	F1@5	Num%	NDCG@5	F1@5	Num%	NDCG@5	F1@5	Num%	NDCG@5	F1@5	Num%	NDCG@5	F1@5
TEM	7.10	**0.0158**	0.0101	6.58	0.0224	0.0522	0.57	0.0026	0.0016	5.43	**0.0560**	0.0329	13.84	0.0010	0.0580
SULM	4.30	0.0002	0.0031	6.74	0.0006	0.0076	8.65	0.0029	0.0097	13.93	0.0051	0.0210	9.24	0.0002	0.0020
ERRA	10.61	0.0099	**0.3257**	13.87	0.0167	0.2477	8.48	0.0018	0.2420	11.30	0.0013	0.2793	19.06	0.0010	0.2230
AOTree	**10.87***	0.0013**	0.0587**	**24.04****	**0.0231***	**0.2816****	**12.10****	**0.0037***	**0.2578***	**33.62****	0.0025**	**0.3594****	**89.35****	**0.0069****	**0.7170****

Table 5: Explainability of the Aspect Order on benchmark datasets. $Num\%$ denotes the coverage ratio of the reviews. $NDCG@5$ and $F1@5$ are the verification of the aspect order explanation only on the first 5 aspects. * and ** represent significant difference at 0.05 and 0.01 level, respectively, compared with the best baseline.

Explanation	{**service**, lunch, location, tables, taste, **food**, staff, **prices**, flavor, **environment**, **service**, reservation, pizza}
Original	Great customer **service**, tasty **food** and drinks, and good **prices**. What more do I need to say? In a city like Vancouver, where most dishes are overpriced, with entitled servers (expecting 20% for barely passable **service**), and in a pretentious **environment**, you crave for something real: simply good **food**, **service**, and **prices**. That's it. And they hit all three notes, and that's how you create loyal customers who come back time and time again supporting your business. Thank you Suika, keep up the good work!
Groundtruth	{**service**, **food**, **prices**, **service**, **environment**, **food**, **service**, **prices**}

Table 6: An example of aspect order generated by AOTree. The words in boldface in the original review are the extracted aspects by the sentiment analysis toolkit.

	Cells Phones & Accessories	Patio, Lawn & Garden	Digital Music
w/o Tree	1.0130	0.9869	0.8135
w/o PEL	1.0175	0.9814	0.8130
w/o SAL	1.0116	1.0642	0.8220
w/o LN	1.0179	0.9946	0.8191
AOTree	**0.9719**	**0.9452**	**0.8050**

Table 7: Ablation study on MSE for AOTree.

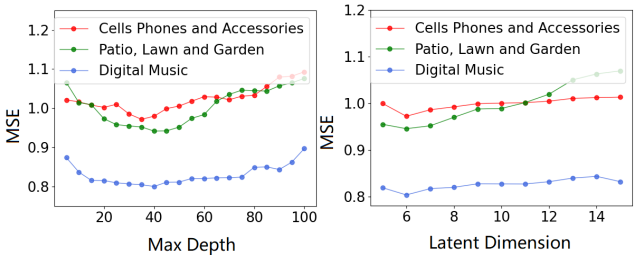


Figure 5: Hyperparameter analysis on three datasets.

Accessories and Digital Music, the coverage value is quite low (only 7.10% and 5.43%). The reason is that the proportion of sensitive users is the lowest in these two datasets (20.37% in Cells Phones and Accessories and 21.05% in Digital Music), resulting the order effects not fully matching the actual user reviews. Also, ERRA performs relatively good results on Cells Phones and Accessories, which show its advatange on aspect prediction due to the aspect enhancement component. In summary, we can conclude that AOTree outperforms TEM and SULM in terms of explainability.

Case Study

We use a case study (from a randomly picked review) to illustrate the explanations. The results are shown in Table 6. Considering the context of recommending a restaurant, our model presents why the restaurant is recommended as an aspect sequence shown in the first row. The original review is shown with extracted aspects in boldface and the ground-truth aspects in the last row. We can see that AOTree find the user may be concerned with aspects of **service**, **lunch**, **location** and so on for this item, which follows a specific decisive order, leading to the final decision. The aspect order can mimic the decision process as reviews and is mostly consistent with the ground-truth order. This qualitative study further confirms the effectiveness of our method.

Ablation and Hyper-parameter Study

Ablation Study The ablation study is considered to justify its validity. As our main goal is accuracy, we perform experiments on MSE. The two main components are the AOTree Generator (Tree) and the Prediction Generator. Furthermore, we split the Prediction Generator into the Position Embedding Layer (PEL), the Self-Attention Layer (SAL) and the Layer Normalization (LN). The MSE are shown in Table 7, where w/o means without the corresponding component. When replacing each components, the corresponding MSE increases, especially for SAL and LN. Although AOTree's design is not complicated and consists of independent components, the ablation results confirm the effectiveness.

Hyper-parameter Study We analyze the impacts of two key hyper-parameters: the max depth of the tree and the size of the latent dimensions. The results are shown in Figure 5. For the max depth, it limits the length of the aspect order, which is related to users' considering process. The figure presents the performance w.r.t. MSE loss over the varying max depth. Although the optimal values are not the same for different datasets, they show a similar trend. With increasing max depth, the MSE loss of AOTree gradually decreases and then increases. Generally, when the max depth is set in the range of $[35, 45]$, we observe the best performance. For the value of the latent dimension, embedding component is related to user/item/aspect id, which shows the ability to represent each variable. Also, the best performance is achieved when the latent dimension is set in the range between $[5, 10]$.

Time Complexity Analysis

We separate complexity analysis into two parts since there are two phases in the learning process of our method. For the AOTree building phase, the time complexity is $O(m \cdot e \cdot l \cdot \log l)$ or $O(n \cdot e \cdot l \cdot \log l)$, where m/n represents the number of users/items for User-AOTree/Item-AOTree, e is the maximum depth of trees, and l is the number of extracted aspects. For the embedding phase, the time complexity is $O(2 \cdot a \cdot d \cdot N)$, where a is the attention size, d is the embedding size and N is the number of training instances. Above all, the overall time complexity is $O(l \cdot e \cdot m \cdot \log l + l \cdot e \cdot n \cdot \log l + 2 \cdot a \cdot d \cdot N)$. The time complexity of the comparison algorithm TEM is $O(S \cdot e \cdot \|x\|_0 \cdot logN + 2 \cdot a \cdot d \cdot S \cdot N)$, where S and $\|x\|_0$ denote the number of trees and the average number of non-zero entries in training instances. Although our method introduces aspect-level complexity (l), we simplify the procedure of feature fusion from multiple trees, thereby reducing overall complexity and making it acceptable.

Discussion

In this work, we confirm both the existence and variance of order effects within reviews and design AOTree to capture the order effects. Extensive experiments exhibit AOTree's strengths like more accurate rating predictions and better aspect order explanations. However, we identify that aspect order mining and modeling is a challenging task and our model still has several limitations. The following discusses the limitations and the potential problems of AOTree, shedding light on future investigation and improvements.

Our preliminary analysis suggests although most users (nearly 70%) and most items (over 60%) have specific aspect order considerations, the order effect seems strong for just a small proportion of users and items. For example, only about 20% users' $consistency\_dis_{user}$ and 20% items' $consistency\_dis_{item}$ values are higher than 0.5, making mining aspect order challenging. As a result, the proportion of *Identified Strong Sensitive Users* uncovered in our experiments is about 20% to 30% in different datasets. We thought data sparsity is the primary reason for this problem, i.e., 73% users only contribute less than 5 reviews, providing limited corpus for aspect mining. Data sparsity is an essential problem in recommender systems, and there have been abundant advanced techniques like auxiliary information integration (Cheng et al. 2023). Future research can investigate how to combine these techniques with our model to strengthen the effectiveness of aspect order mining, especially for users and items with limited reviews.

While our method provides promising solutions, it might aggravate problems embedded in recommender systems. Our work focuses on utilizing users' decision aspect order preference, which may inadvertently exacerbate filter bubbles by reinforcing users' preferences on aspect order, potentially leading to less exposure to diverse perspectives and aggravating polarization. So further research can explore integrating diversity-enhancing mechanisms (e.g., incorporating new items) when leveraging a user's aspect order preference, promoting exposure to different perspectives.

Conclusion

This work focuses on the Order Effects Theory of aspects in explainable recommendation and seeks a specific form of interpretability, known as human simulability. We first verify the applicability of Order Effects Theory by analyzing the Yelp dataset. Then, we proposed Aspect Order Tree-based (AOTree) explainable recommendation based on the theory, which is achieved to capture the relationships among decisive factors. Experiment results show that the proposed method can perform well on public datasets, achieving higher accuracy as well as better interpretability.

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) under the Grant No. 62172106 and 61932007. Tun Lu is also a faculty of Shanghai Key Laboratory of Data Science, Fudan Institute on Aging, MOE Laboratory for National Development and Intelligent Governance, and Shanghai Institute of Intelligent Electronics & Systems, Fudan University.

References

Anderson, N. H. 1965. Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology*, 2(1): 1.

Anwar, K.; Zafar, A.; and Iqbal, A. 2023. An efficient approach for improving the predictive accuracy of multi-criteria recommender system. *International Journal of Information Technology*, 1–8.

Arapoff, N. 1967. Writing: A thinking process. *Tesol Quarterly*, 1(2): 33–39.

Bauman, K.; Liu, B.; and Tuzhilin, A. 2017. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 717–725.

Chen, C.; Zhang, M.; Liu, Y.; and Ma, S. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*, 1583–1592.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, X.; Qin, Z.; Zhang, Y.; and Xu, T. 2016. Learning to rank features for recommendation over multiple categories. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 305–314.

Cheng, H.; Wang, S.; Lu, W.; Zhang, W.; Zhou, M.; Lu, K.; and Liao, H. 2023. Explainable recommendation with personalized review retrieval and aspect learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, 51–64.

Chin, J. Y.; Zhao, K.; Joty, S.; and Cong, G. 2018. ANR: Aspect-based neural recommender. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 147–156.

Ebbinghaus, H. 2013. Memory: A contribution to experimental psychology. *Annals of Neurosciences*, 20(4): 155–156.

Fan, G.; Zhang, C.; Chen, J.; Li, P.; Li, Y.; and Leung, V. C. M. 2023. Improving rating prediction in multi-criteria recommender systems via a collective factor model. *IEEE Transactions on Network Science and Engineering*, 10(6): 3633–3643.

Fan, G.; Zhang, C.; Wang, K.; and Chen, J. 2022. MV-HAN: A hybrid attentive networks based multi-view learning model for large-scale contents recommendation. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 1–5. Association for Computing Machinery.

Felfernig, A.; Friedrich, G.; Gula, B.; Hitz, M.; Kruggel, T.; Leitner, G.; Melcher, R.; Riepan, D.; Strauss, S.; Teppan, E.; et al. 2007. Persuasive recommendation: Serial position effects in knowledge-based recommender systems. In *International Conference on Persuasive Technology*, 283–294. Springer.

Fogg, B. J. 1998. Persuasive computers: Perspectives and research directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 225–232.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gershberg, F. B.; and Shimamura, A. P. 1994. Serial position effects in implicit and explicit tests of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6): 1370–1378.

Gretzel, U.; and Fesenmaier, D. R. 2006. Persuasion in recommender systems. *International Journal of Electronic Commerce*, 11(2): 81–100.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, 197–206.

Kanoulas, E.; and Aslam, J. A. 2009. Empirical justification of the gain and discount function for nDCG. In *Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, 611–620.

Karn, A. L.; Karna, R. K.; Kondamudi, B. R.; Bagale, G.; Pustokhin, D. A.; Pustokhina, I. V.; and Sengan, S. 2023. Customer centric hybrid recommendation system for E-Commerce applications by integrating hybrid sentiment analysis. *Electronic Commerce Research*, 23(1): 279–314.

Kinga, D.; Adam, J. B.; et al. 2015. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, volume 5, 6. San Diego, California;.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37.

Li, P.; Wang, Z.; Ren, Z.; Bing, L.; and Lam, W. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 345–354.

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57.

Maegherman, E.; Ask, K.; Horselenberg, R.; and van Koppen, P. J. 2020. Law and order effects: on cognitive dissonance and belief perseverance. *Psychiatry, Psychology and Law*, 1–20.

Mcauley, J.; and Leskovec, J. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Acm Conference on Recommender Systems*, 165–172.

Meshram, R.; Gopalan, A.; and Manjunath, D. 2016. Optimal recommendation to users that react: Online learning for a class of POMDPs. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 7210–7215.

Nevo, B. 1989. Validation of graphology through use of a matching method based on ranking. *Perceptual and Motor Skills*, 69(3_suppl): 1331–1336.

Pan, S.; Li, D.; Gu, H.; Lu, T.; Luo, X.; and Gu, N. 2022. Accurate and explainable recommendation via review rationalization. In *Proceedings of the ACM Web Conference 2022*, 3092–3101.

Peake, G.; and Wang, J. 2018. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2060–2069.

Petty, R. E.; Tormala, Z. L.; Hawkins, C.; and Wegener, D. T. 2001. Motivation to think and order effects in persuasion: The moderating role of chunking. *Personality and Social Psychology Bulletin*, 27(3): 332–344.

Shani, G.; Heckerman, D.; Brafman, R. I.; and Boutilier, C. 2005. An MDP-based recommender system. *Journal of Machine Learning Research*, 6(9).

Tan, J.; Xu, S.; Ge, Y.; Li, Y.; Chen, X.; and Zhang, Y. 2021. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 1784–1793.

Tao, Y.; Jia, Y.; Wang, N.; and Wang, H. 2019. The fact: Taming latent factor models for explainability with factorization trees. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 295–304.

Wang, N.; Wang, H.; Jia, Y.; and Yin, Y. 2018a. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, 165–174.

Wang, X.; He, X.; Feng, F.; Nie, L.; and Chua, T.-S. 2018b. Tem: Tree-enhanced embedding model for explainable recommendation. In *Proceedings of the 2018 World Wide Web Conference*, 1543–1552.

Waswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Wu, M.; Hughes, M.; Parbhoo, S.; Zazzi, M.; Roth, V.; and Doshi velez, F. 2017. Beyond sparsity: Tree regularization of deep models for interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Zhang, Y.; Chen, X.; et al. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1): 1–101.

Zhang, Y.; Lai, G.; Zhang, M.; Zhang, Y.; Liu, Y.; and Ma, S. 2014a. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval*, 83–92.

Zhang, Y.; Zhang, H.; Zhang, M.; Liu, Y.; and Ma, S. 2014b. Do users rate or review? Boost phrase-level sentiment labeling with review-level sentiment classification. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1027–1030.

Zhao, P.; Shui, T.; Zhang, Y.; Xiao, K.; and Bian, K. 2021. Adversarial oracular seq2seq learning for sequential recommendation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 1905–1911.

Zheng, Y. 2017. Criteria chains: A novel multi-criteria recommendation approach. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 29–33.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, addressing this research question aims to advance science while adhering to social contracts.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the contributions are addressed in the Abstract and the Introduction section.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we illustrate the applicability of the method in the Preliminary Analysis and the Experiment Section.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No, because the datasets used in the experiment are commonly used public datasets in the field, the excessive declaration of their usage is not necessary.
   (e) Did you describe the limitations of your work? Yes, we describe the limitations of our work in the Conclusion section, providing insights into the study's scope and applicability.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, there is a brief discussion in the Ethical Statement section.
   (g) Did you discuss any potential misuse of your work? Yes, there is a brief discussion in the Ethical Statement section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, there is a brief description in the Ethical Statement section.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we have read the ethics review guidelines and ensured that our paper conforms to them. Details are in the Ethical Statement section.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA.
   (b) Have you provided justifications for all theoretical results? NA.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA.
   (e) Did you address potential biases or limitations in your theoretical framework? NA.
   (f) Have you related your theoretical results to the existing literature in social science? NA.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA.
   (b) Did you include complete proofs of all theoretical results? NA.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, the code, data and instructions will be public if the paper is accepted.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, all the details are included in the Experimental Setup section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, the parameters not mentioned in the paper use default values, so there are not too many declarations.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, the details are included in the Experimental Setup section.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, the evaluation metrics are commonly used and related to our work and claims.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, because the discussion might not be relevant to the specific objectives of the work.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, the corresponding references are noted in the sentences.
   (b) Did you mention the license of the assets? Yes, the license of the assets are mentioned in the Ethical Statement section.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, the URL is attached at the corresponding footnote.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, because the datasets used in the experiment are commonly used public datasets in the field.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, because the datasets used in the experiment are commonly used public datasets in the field, they have already undergone some processing.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.
   (d) Did you discuss how data is stored, shared, and deidentified? NA.

Ethical Statement

The study was conducted on commonly used public datasets. All the data is anonymous and allowed to use. The result of our research conducted positive outcome for effective recommendation. In our perspective, our work does not exists negative society impacts or potential misuse, also, the data collection process and modeling process comply with the procedure without ethical implications.