Journal of Data Intelligence, Vol. 2, No.4 (2021) 434-447

MAKING USE OF MORE REVIEWS SKILLFULLY IN EXPLAINABLE RECOMMENDATION GENERATION[a]

SHUNSUKE KIDO, RYUJI SAKAMOTO, MASAYOSHI ARITSUGI[b]

Kumamoto University, Japan

{souen0823, 131t6802}@gmail.com, aritsugi@cs.kumamoto-u.ac.jp

There are a lot of reviews in the Internet, and existing explainable recommendation techniques use them. However, how to use reviews has not been so far adequately addressed. This paper proposes a new exploiting method of reviews in explainable recommendation generation. Our new method makes use of not only reviews written but also those referred to by users. This paper adopts two state-of-the-art explainable recommendation approaches and shows how to apply our method to them. Moreover, our method in this paper considers the possibility of making use of reviews which do not provide detailed review utilization. Our proposal can be applied to different explainable recommendation approaches, which is shown by adopting the two approaches, with reviews that do not necessarily provide their detailed utilization data. The evaluation with using Amazon reviews shows an improvement of the two explainable recommendation approaches. Our proposal is the first attempt to make use of reviews which are written or referred to by users in generating explainable recommendation. Particularly, this study does not suppose that reviews provide their detailed utilization data.

Key words: Recommender systems, sentiment analysis, review utilization, recommendation quality

1 Introduction

Recently, explainable recommendation has emerged in recommender systems [21]. Conventional collaborative filtering approaches, especially matrix factorization (MF)-based ones, are able to give good recommendation of items to users [8]. However, just presenting recommendation to users may not be good enough. Explainable recommender systems therefore attempt to show to users not only recommendation but also its explanation. Such explanations can serve seven aims, namely, effectiveness, satisfaction, transparency, scrutability, trust, persuasiveness, and efficiency [17, 18]. Note that it is generally difficult for MF-based techniques to understand and explain recommendation generated by them due to MF's concept.

---

[a] This paper is based on a previous paper published in the proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services (iiWAS), 2020 [6]

[b] Corresponding author.

Conventional explainable recommendation approaches have tried to obtain user preferences from review texts [10, 22, 3, 1] and achieved explainable recommendation with improved recommendation accuracy. However, there is room to improve in treating reviews, that is, it is hard to obtain a user's preferences correctly if the number of reviews written by the user is small. For example, 49% of users in the Yelp dataset used in [22] wrote one review only and thus such review texts were not exploited in recommendation generation. Similarly, many users in the Epinion dataset used in [11] wrote few reviews. We should also note that there can be reviews of low quality; since review texts are often input in an unstructured format, reviews of various qualities can exist. It is difficult to obtain user preferences correctly with using review texts of various qualities. Therefore, making use of more reviews skillfully in explainable recommendation generation is required.

In this paper, we focus on a new exploiting method of more reviews borrowed from [6]. The method makes use of not only reviews written but also those referred to by users. Note that it can be supposed to appear that who wrote a given review but not that who referred to a given review generally. We therefore consider the possibility of making use of reviews which do not provide detailed review utilization in this study. Our method with the consideration is examined empirically with real review data by applying it to two different state-of-the-art explainable recommendation approaches.

The main contributions of this paper are as follows:

- We focus on a new method [6] of making use of reviews both written and referred to by users to infer user preferences more precisely. We show how to apply the method to existing explainable recommendation approaches by adopting two state-of-the-art ones, that is, explicit factor model (EFM) [22] and multi-task explainable recommendation (MTER) [19].
- We consider the possibility of making use of reviews which do not provide detailed review utilization in the method. This consideration can make the value of the new method higher.
- We evaluate the method with the consideration to show how it can be of benefit to existing explainable recommendation approaches.

The remainder of this paper is organized as follows. Section 2 mentions related work. Section 3 introduces our proposal and considers how we make use of reviews without their detailed utilization in our proposal. Section 4 discusses the results of empirical evaluation. Finally, we conclude this paper with future research directions in Section 5.

2 Related Work

There have been studies of using reviews to obtain user preference for generating recommendations [10, 22, 19]. Since review data include user opinions on products and services, we can infer user preference using them. McAuley and Leskovec [10] proposed a model called Hidden Factors as Topics (HFT) which discovers topics correlating with the hidden factors of products and users based on review texts. HFT also used ratings attached to reviews. Zhang et al. [22] proposed a model called explicit factor model (EFM) which extracts explicit product features and user opinions from user reviews based on phrase-level sentiment analysis and integrates them in a matrix factorization algorithm to generate explainable recommendations. They showed EFM could handle more explicit features of items than HFT and generate both recommendation of high accuracy and its good

explanation. Wang et al. [19] exploited opinionated review text data in generating recommendation and its explanation. With Tucker decomposition of tensors [7, 5] they represented users, items, features, and opinion phrases as four non-negative matrices and developed multi-task explainable recommendation (MTER). In this study, we adopt EFM and MTER as existing explainable recommendation approaches.

Reviews used in these studies were those written by users. Note that while users do not so often write reviews but that they refer to reviews written by others frequently, the studies focused on reviews written by users only. Kido et al. [6] developed a method that can use not only reviews written but also those referred to by users in generating explainable recommendation. Note that they assumed that review data to be used provided their detailed utilization in their study. However, there have been few such data in the real world. In this paper, we focus on the method proposed by Kido et al. [6] but do not assume that review data provide their detailed utilization; we consider how to apply the method to general review data, which do not provide their utilization histories. We evaluate the consideration with using real review data empirically.

3 Proposal

Lu et al. [9] said that reviews play a central role in the decision-making process of online users, and the role of online reviews is expected to become increasingly important as online commerce activity continues to grow. Hong et al. [4] hypothesized that the basic user intention is to acquire the product information from reviews, by which to support the purchase decisions. We therefore assume that item features mentioned in reviews that users referred to can be useful in inferring user preferences. In this section, we first assume that reference histories of users are available and attempt to take account of reviews in the histories in the process of generating explainable recommendation in addition to reviews written by users, by borrowing discussions in [6]. We then consider removing the assumption.

3.1. Making Use of Actually Utilized Reviews

Let $\mathrm{F}=\{F_1, F_2, \cdots, F_p\}$ and $\mathrm{U}=\{u_1, u_2, \cdots, u_m\}$ be the sets of p product features and m users, respectively. The existing explainable recommendation approaches focused on two types of product features, namely, (A) those appearing in reviews written by a user, and (B) the others. Let F_i^+ be the set of features mentioned in the reviews written by user u_i, then data $Data$ learned by the approaches can be expressed as follows:

$$Data = \{(u_i, f_A, f_B) \mid u_i \in \mathrm{U} \wedge f_A \in \mathrm{F}_i^+ \wedge f_B \in \mathrm{F} \backslash \mathrm{F}_i^+\} \quad (1)$$

We add two more types of product features, namely, (C) those appearing in reviews referred to by a user, and (D) the others. Let G be the set of features mentioned in reviews written by user *u'* and referred to by user *u*, that is,

$$G = \{(u, u', f) \mid u \in \mathrm{U}, u' \in \mathrm{U}, f \in \mathrm{F}\} \quad (2)$$

where user *u* referred to feature *f* mentioned in reviews written by user *u'*. It is generally hard to distinguish which features are really referred to. We here assume that every feature written in a review is referred to whenever the review is referred to.

Let $\mathrm{F}_i'^+ = \{f_C \mid (u, u', f_C) \in G \wedge f_C \in \mathrm{F} \backslash \mathrm{F}_i^+\}$, then data $Data'$ learned by our proposal can be expressed as follows:

$$Data' = \{(u_i, f_A, f_B, f_C, f_D) \mid u_i \in \mathrm{U} \wedge f_A \in \mathrm{F}_i^+ \wedge f_B \in \mathrm{F} \backslash (\mathrm{F}_i^+ \cup \mathrm{F}_i'^+) \wedge f_C \in \mathrm{F}_i'^+ \wedge f_D \in \mathrm{F} \backslash (\mathrm{F}_i^+ \cup \mathrm{F}_i'^+)\} \quad (3)$$

Rendle et al. [14] assumed that users prefer the observed items to the unobserved ones and proposed Bayesian Personalized Ranking (BPR). Similarly, Yu et al. [20] proposed a Multiple Pairwise Ranking (MPR), where more types of items than BPR were introduced, and showed that their MPR was superior to BPR. Based on these studies, we focus on treating the four types as "preference degrees of f_A and f_C are larger than those of f_B and f_D, respectively, and preference degree of f_A is larger than that of f_C". For making discussions concrete, we adopt EFM [22] and MTER [19] and extend them with the method in this section.

3.1.1. EFM with Actually Utilized Reviews

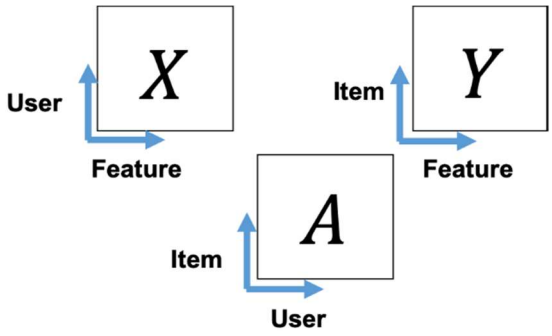

Figure 1 Matrixes in EFM.

In EFM [22], product features and user opinions are extracted from reviews, and user-item rating matrix A, user-feature attention matrix X, and item-feature quality matrix Y are constructed (Figure 1). X is constructed as follows:

$$X_{ij} = \begin{cases} 0, (if\ user\ u_i\ did\ not\ mention\ feature\ F_j) \\ 1 + (N-1)\left(\frac{2}{1+e^{-t_{ij}}} - 1\right), (else) \end{cases} \quad (4)$$

where N is the maximum rating value and user u_i mentioned feature F_j in t_{ij} times in u_i's reviews. Let $\mathcal{P} = \{p_1, p_2, \cdots, p_n\}$ be the set of n items/products. Y is constructed as follows:

$$Y_{ij} = \begin{cases} 0, (if\ item\ p_i is\ not\ reviewed\ on\ feature\ F_j) \\ 1 + \frac{N-1}{1+e^{-k \cdot s_{ij}}}, (else) \end{cases} \quad (5)$$

where feature F_j was mentioned k times on item p_i and the average of sentiment of feature F_j in those k mentions was s_{ij}.

The following optimization task is performed in EFM:

$$\underset{U_1,U_2,V,H_1,H_2}{\text{minimize}} \{\|PQ^T - A\|_F^2 + \lambda_x\|U_1V^T - X\|_F^2 + \lambda_y\|U_2V^T - Y\|_F^2 + \lambda_u(\|U_1\|_F^2 + \|U_2\|_F^2) + \lambda_h(\|H_1\|_F^2 + \|H_2\|_F^2) + \lambda_v\|V\|_F^2\}$$
$$s.t.\ U_1 \in \mathbb{R}_+^{m\times r}, U_2 \in \mathbb{R}_+^{n\times r}, V \in \mathbb{R}_+^{p\times r}, H_1 \in \mathbb{R}_+^{m\times r'}, H_2 \in \mathbb{R}_+^{n\times r'}\ and\ P = [U_1\ H_1], Q = [U_2\ H_2] \quad (6)$$

where U_1, U_2, V, H_1 and H_2 are implicit features.

In LRPPM-CF (Learn to Rank user Preferences based on Phrase-level sentiment analysis across Multiple categories-Collaborative Filtering) [2], ranking-based optimization target on implicit feedback is implemented with using the ranking-based criterion of BRP [14]. Let $\hat{X}_{u_i f_A}$ and $\hat{X}_{u_i f_B}$ be predicted preference degrees of features f_A and f_B of user u_i, then following the way, user-feature attention differences between the two types of product features are expressed as

$$\hat{X}_{u_i f_A f_B} = \hat{X}_{u_i f_A} - \hat{X}_{u_i f_B} = \boldsymbol{u}_{1i}^T \boldsymbol{v}_{f_A} - \boldsymbol{u}_{1i}^T \boldsymbol{v}_{f_B}. \tag{7}$$

To replace $Data$ in Eq.(1) with $Data'$ in Eq.(3) for making use of utilized reviews, we consider the following user-feature attention differences as in MPR [20]:

$$\hat{X}_{u_i f_A f_B f_C f_D} = \left(\hat{X}_{u_i f_A} - \hat{X}_{u_i f_B}\right) - \left(\hat{X}_{u_i f_C} - \hat{X}_{u_i f_D}\right) = \left(\boldsymbol{u}_{1i}^T \boldsymbol{v}_{f_A} - \boldsymbol{u}_{1i}^T \boldsymbol{v}_{f_B}\right) - \left(\boldsymbol{u}_{1i}^T \boldsymbol{v}_{f_C} - \boldsymbol{u}_{1i}^T \boldsymbol{v}_{f_D}\right). \tag{8}$$

When user u_i did not write a review but referred to some reviews written by other users, Eq.(8) is replaced with the following:

$$\hat{X}_{u_i f_A f_B f_C f_D} = \hat{X}_{u_i f_C} - \hat{X}_{u_i f_D} = \boldsymbol{u}_{1i}^T \boldsymbol{v}_{f_C} - \boldsymbol{u}_{1i}^T \boldsymbol{v}_{f_D}. \tag{9}$$

By taking Eq.(8) into account, we enhance Eq.(6) and perform the following optimization task in the method:

$$\begin{aligned}
\underset{U_1, U_2, V, H_1, H_2}{\text{minimize}} \Bigg\{ & \sum_{(i,j)\in K} \left(a_{ij} - \boldsymbol{h}_{1j}^T \boldsymbol{h}_{2j} - \boldsymbol{u}_{1i}^T \boldsymbol{u}_{2j}\right)^2 \\
& - \lambda_x \sum_{(i,j)\in K} \sum_{f_A \in F_i^+} \sum_{f_B, f_D \in F \setminus \left(F_i^+ \cup F_i'^+\right)} \sum_{f_C \in F_i'^+} \ln \sigma\left(\hat{X}_{u_i f_A f_B f_C f_D}\right) \\
& + \lambda_y \sum_{(i,j)\in K} \sum_{f_Q \in F_j} \left(y_{j f_Q} - \boldsymbol{u}_{2j}^T \boldsymbol{v}_{f_Q}\right)^2 + \lambda_u \sum_{r}^{R} \sum_{(i,j)\in K} \left(u_{1ir}^2 + u_{2jr}^2\right) \\
& + \lambda_h \sum_{r'}^{R'} \sum_{(i,j)\in K} \left(h_{1ir'}^2 + h_{2jr'}^2\right) \\
& + \lambda_v \sum_{r}^{R} \left(\sum_{f_A \in F_i^+} v_{f_A r}^2 + \sum_{f_B, f_D \in F \setminus \left(F_i^+ \cup F_i'^+\right)} \left(v_{f_B r}^2 + v_{f_D r}^2\right) + \sum_{f_C \in F_i'^+} v_{f_C r}^2 \right) \Bigg\} \\
s.t.\ & U_1 \in \mathbb{R}_+^{m \times r}, U_2 \in \mathbb{R}_+^{n \times r}, V \in \mathbb{R}_+^{p \times r}, H_1 \in \mathbb{R}_+^{m \times r'}, H_2 \in \mathbb{R}_+^{n \times r'} \text{ and } P = [U_1\ H_1], Q = [U_2\ H_2].
\end{aligned} \tag{10}$$

We pay attention on user-item ratings and user preferences to features mentioned in written and referred reviews in calculating ranking scores. We agree an assumption in [22, 2] that a user's decision about whether to make a purchase is based on several important product features to the user, and we thus use the most cared k largest values in row vector X_u, of user u. Let $C_u = \{c_{u1}, c_{u2}, \cdots, c_{uk}\}$ be the set of k most cared features, which are actually the column indices of the k largest values in row vector X_u, of user u. We then express the aspect of user u's preferences to features of item i as follows:

$$R_{ui}^{feature} = \frac{\sum_{c \in C_u} \hat{X}_{uc} \cdot \hat{Y}_{ic}}{kN}. \tag{11}$$

We calculate the ranking score of item i for user u as follows:

$$RS_{ui} = \alpha \cdot R_{ui}^{feature} + (1 - \alpha)\hat{A}_{ui} \tag{12}$$

where α is a scaling factor for the trade-off between feature-based scores and direct user-item ratings.

3.1.2 MTER with Actually Utilized Reviews

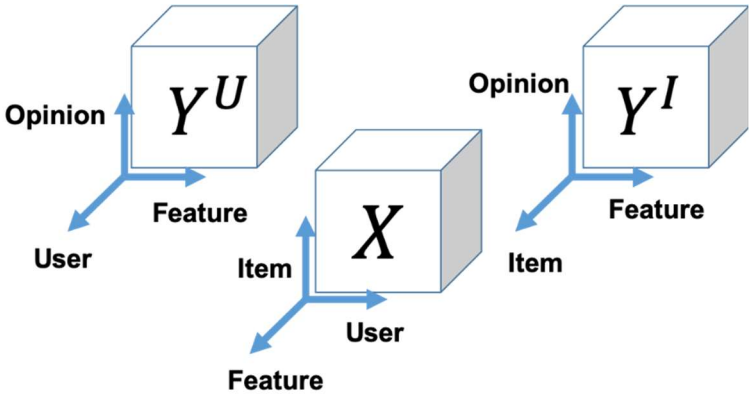

Figure 2 Tensors in MTER.

There are tensors X over feature, user, and item, Y^U over user, feature, and opinion, and Y^I over item, feature, and opinion in MTER [19] (Figure 2). X is defined like Eq.(5), and the others are defined like Eq.(4). The following optimization task is performed in MTER:

$$\begin{aligned}\min_{\hat{X},\hat{Y}^U,\hat{Y}^I} &\left\|\hat{X} - \tilde{X}\right\|_F + \left\|\hat{Y}^U - Y^U\right\|_F + \left\|\hat{Y}^I - Y^I\right\|_F \\ &- \lambda_B \sum_{i=1}^{m} \sum_{(j,l)\in D_i^S} \ln \sigma\left(\hat{x}_{ij(p+1)} - \hat{x}_{il(p+1)}\right) + \lambda_F(\|U\|^2 + \|I\|^2 + \|F\|^2 + \|O\|^2) \\ &+ \lambda_G \left(\|G_1\|^2 + \|G_2\|^2 + \|G_3\|^2\right) \\ s.t.\, &\hat{X} = G_1 \times_a U \times_b I \times_c \tilde{F}, \hat{Y}^U = G_2 \times_a U \times_c F \times_d O, \hat{Y}^I = G_3 \times_b I \times_c F \times_d O, \\ &U \geq 0, I \geq 0, F \geq 0, O \geq 0, G_1 \geq 0, G_2 \geq 0, G_3 \geq 0\end{aligned} \tag{13}$$

where $\hat{X}$ is a learned tensor with training data, $\tilde{X}$ is an observed tensor, missing data of $\tilde{X}$'s are predicted by Tucker decomposition [7, 5], D_i^S is a pairwise order set of user u_i, $U \in \mathbb{R}_+^{m\times a}$, $I \in \mathbb{R}_+^{n\times b}$, $F \in \mathbb{R}_+^{p\times c}$, and $O \in \mathbb{R}_+^{q\times d}$ are non-negative matrices in the latent factor space, $G \in \mathbb{R}_+^{a\times b\times c}$ is a core tensor, and $G \times_n M$ denotes the n-mode product between tensor G and matrix M.

To extend data for learning from $Data$ in Eq.(1) to $Data'$ in Eq.(3), we express user-feature attention difference as follows:

$$\hat{X}_{u_i f_A f_B f_C f_D} = \left(\hat{X}_{u_i f_A} - \hat{X}_{u_i f_B}\right) - \left(\hat{X}_{u_i f_C} - \hat{X}_{u_i f_D}\right). \tag{14}$$

In the framework of MTER, we express the relation between users and features as the average value simple. Let $\mathrm{I} = \{i_1, i_2, \cdots, i_n\}$ be the set of n items/products. Then, we consider the following user-feature attention differences:

$$\begin{aligned}\hat{X}_{u_i f_A f_B f_C f_D} &= \left(\hat{X}_{u_i f_A} - \hat{X}_{u_i f_B}\right) - \left(\hat{X}_{u_i f_C} - \hat{X}_{u_i f_D}\right) \\ &= \left(\frac{\sum_{j=1}^{n} g_1 \boldsymbol{u}_i \boldsymbol{i}_j \boldsymbol{f}_A}{n} - \frac{\sum_{j=1}^{n} g_1 \boldsymbol{u}_i \boldsymbol{i}_j \boldsymbol{f}_B}{n}\right) - \left(\frac{\sum_{j=1}^{n} g_1 \boldsymbol{u}_i \boldsymbol{i}_j \boldsymbol{f}_C}{n} - \frac{\sum_{j=1}^{n} g_1 \boldsymbol{u}_i \boldsymbol{i}_j \boldsymbol{f}_D}{n}\right).\end{aligned} \tag{15}$$

As Eq.(9), when user u_i did not write a review but referred to some reviews written by other users, Eq.(15) is replaced with the following:

$$\hat{X}_{u_i f_A f_B f_C f_D} = \hat{X}_{u_i f_C} - \hat{X}_{u_i f_D} = \frac{\sum_{j=1}^{n} g_1 \boldsymbol{u}_i \boldsymbol{i}_j \boldsymbol{f}_C}{n} - \frac{\sum_{j=1}^{n} g_1 \boldsymbol{u}_i \boldsymbol{i}_j \boldsymbol{f}_D}{n}. \tag{16}$$

Then, we add this to the optimization task as follows:

$$\begin{aligned}\min_{\hat{X}, \hat{Y}^U, \hat{Y}^I} & \left\|\hat{X} - \tilde{X}\right\|_F + \left\|\hat{Y}^U - Y^U\right\|_F + \left\|\hat{Y}^I - Y^I\right\|_F \\ & - \lambda_B \sum_{i=1}^{m} \sum_{(j,l) \in D_i^S} \ln \sigma\left(\hat{x}_{ij(p+1)} - \hat{x}_{il(p+1)}\right) \\ & - \lambda_C \sum_{(i,j) \in K} \sum_{f_A \in F_i^+} \sum_{f_B, f_D \in F \setminus (F_i^+ \cup F_i'^+)} \sum_{f_C \in F_i'^+} \ln \sigma\left(\hat{X}_{u_i f_A f_B f_C f_D}\right) \\ & + \lambda_F\left(\|U\|^2 + \|I\|^2 + \|F\|^2 + \|O\|^2\right) + \lambda_G \left(\|G_1\|^2 + \|G_2\|^2 + \|G_3\|^2\right) \\ s.t. \hat{X} &= G_1 \times_a U \times_b I \times_c \tilde{F}, \hat{Y}^U = G_2 \times_a U \times_c F \times_d O, \hat{Y}^I = G_3 \times_b I \times_c F \times_d O, \\ & U \geq 0, I \geq 0, F \geq 0, O \geq 0, G_1 \geq 0, G_2 \geq 0, G_3 \geq 0.\end{aligned} \tag{17}$$

Similar to the case of EFM, we calculate ranking scores with user preferences to features mentioned in written and referred to reviews. Let $\left\{s_{ij1}, s_{ij2}, \cdots, s_{ijn_j}\right\}$ be the set of sentiment values of item i's feature j which was mentioned n_j times, then we define the quality of feature j of item i as follows:

$$q_{ij} = \begin{cases} 0, (if\ item\ i\ is\ not\ reviewed\ on\ feature\ j) \\ \dfrac{1}{1 + e^{-\sum_{l=1}^{n_j} s_{ijl}}}, (else) \end{cases}. \tag{18}$$

Let $C_{ui} = \{c_{ui1}, c_{ui2}, \cdots, c_{uik}\}$ be the set of the k largest values in tensor X for the user-item (u, i) pair, then we express the aspect of user preferences to features as follows:

$$R_{ui}^{feature} = \frac{\sum_{c \in C_{ui}} \hat{X}_{uic} \cdot q_{ic}}{\pi k} \tag{19}$$

where π is a rescaling parameter. The ranking score of item i for user u is calculated by Eq.(12) with Eq.(19).

3.2. In Cases where Detailed Utilization Histories are Unavailable in Review Data

Ciao [16] provides detailed utilization of data, that is, who referred to which reviews. Kido et al. [6] used the data in evaluating their method of using written and referred reviews in explainable recommendation generation and showed that their method can improve existing explainable recommendation approaches as regards both recommendation and its explanation qualities. However, we cannot assume we can use detailed utilization of reviews generally. In this paper, we do not assume that each review does not know who referred to it.

Instead, we assume that each review knows how many times it is referred to and how many times it is rated by users. Such data can be generally available in many review data, such as Amazon and TripAdvisor. To infer review utilization, we assume that reviews with high ratings and high referred numbers are utilized by all users in our proposal. This does not allow us to reflect personalized interest precisely but general interest to recommendation, thereby making use of many reviews as much as possible. Note that all we have to do is to perform the optimization tasks of Eqs.(10) and (17) under the assumption.

A problem of the assumption is that every review could have high referred numbers if it has been referred to for a long time. We propose putting the following score to each review and use reviews with high ratings and high scores in our proposal:

$$score = \frac{the\ referred\ number}{\ln(shown\ periods\ in\ years) + 1}. \tag{20}$$

Another idea is that we take account of similarities between users based on their reviews. We calculate the similarities with using a user-feature matrix (e.g., Eq.(4) in EFM) and users having 0.5 or more cosine similarity are supposed to refer to reviews written by them.

4 Experiments

4.1. Data

We collected data of mobile phones and peripheral equipment from Amazon. As in [6], we filtered out the users with less than five item reviews and the items with less than two reviews. The statistics of the data are shown in Table I.

Table I: Statistics of data

Amazon data	#users	#items	#reviews
data collected	27,879	10,429	194,439
data used	60	598	2,705

4.2. Top-10 Recommendation

We did top-10 recommendation generation and compared results of 5-cross validation in terms of normalized discounted cumulative gain (NDCG) of recommended ranked items. We fixed the number of most cared features 25. We denote the idea that we assume top 10% of highly referred reviews are referred to by all users as count, and the two additional proposals are denoted as time and time+sim, respectively. Figure 3 shows the results where EFM+ and MTER+ denote the enhancements of EFM

and MTER by our proposal, respectively. Although the improvements shown in the figure were small compared to those shown in [6], we observe that our proposal was able to slightly improve EFM and MTER.

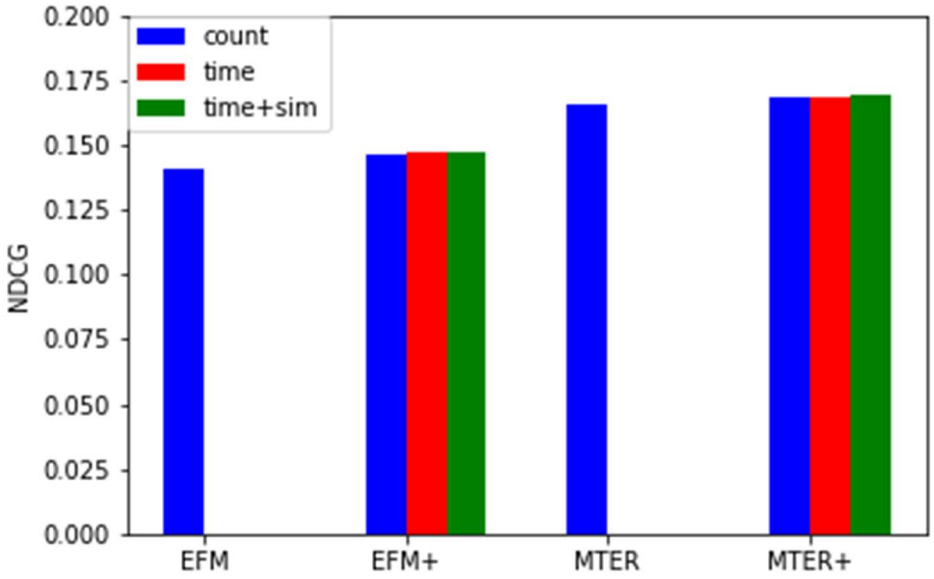

Figure 3 Top-10 recommendation results.

4.3. Explanations

Based on the results shown in Figure 3, we use time+sim as EFM+ and MTER+ in the following. Table II shows explanation examples generated in the experiments. Our proposals generated different explanations from EFM and MTER to user A, as shown in the table, while there were cases where there was no difference between the original methods and ours, as to user B in the table. In the experiments, we used 73 reviews that were supposed to be referred to by user A, while we used only 20 reviews for user B case. Also, the item recommended to user A had 39 reviews, while that to user B had 6 reviews only. The results indicate that the number of reviews we could use in explainable recommendation generation has big effect on generated recommendation explanations.

Table II: Explanation examples

user	method	explanation
A	EFM	You might be interested in [camera], on which this product performs well.
	EFM+	You might be interested in [battery], on which this product performs well.
	MTER	Its [performance] is [good] [compare] [enhances].
	MTER+	Its [capacity] is [extra] [large] [comparable].
B	EFM	You might be interested in [sound], on which this product performs well.
	EFM+	You might be interested in [sound], on which this product performs well.
	MTER	Its [speaker] is [great] [clear] [portable].
	MTER+	Its [speaker] is [great] [clear] [portable].

The results also indicate that our proposal could give same or better explanations to users, which is evaluated quantitatively in the following. We compare the qualities of the methods in quantity in two aspects, as in [6]. One is based on the concept of explanation continuity [12]. We generated two sets of reviews, one was the whole and the other was removed 1/3 randomly selected reviews from the whole. We then supposed them as two nearly equivalent data points in the concept, and got NDCG of top-10 features. We expect that good explanation generation methods have high NDCG. The other is about the quality of feature prediction. This was performed as follows: we did 5-cross validation where we generated recommendations and their explanations for test data. We supposed that features used in the explanations would have relatively higher rankings in the model constructed in training. We evaluated this by comparing the distributions of rankings of features in test data. We fixed the number of most cared features 25 as in the previous.

Table III: Explanation continuity in terms of NDCG

EFM	EFM+	MTER	MTER+
0.7927	0.8193	0.8344	0.8452

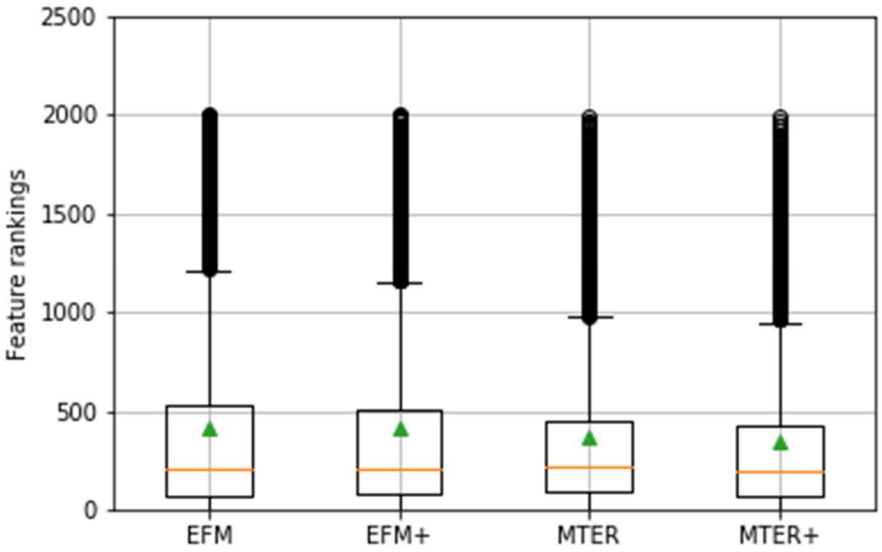

Figure 4 Feature prediction.

Table III and Figure 4 show the results of the two aspects, respectively. As shown in them, the performances of EFM+ and MTER+ were slightly better than their original methods. If a quite larger number of data were available, the amount of improvement could become larger. According to the results, we can say that our proposal could improve the performance of explainable recommendation by means of making use of reviews, which have not been used for inferring preferences of users who did not write the reviews in the existing approaches.

5 Conclusions and Future Work

In this paper, we focused on a method of making use of both reviews written and those referred to by users. The method can be applied to existing explainable recommendation approaches. The experimental results indicated that the method could improve the performance of explainable recommendation approaches in terms of both recommendation and explanation with using many reviews, which have not been used so far for inferring preferences of users who did not write the reviews.

Our current algorithms need long calculation time because our proposal increases the numbers and variations of reviews used in generating explainable recommendations. We would like to develop fast algorithms for the calculation in the future. We did not consider bad reviews, such as opinion spams [13] and deceptive reviews [15] in this paper. We would also like to handle such bad reviews in explainable recommendation generation.

References

1. Baral, R., Zhu, X., Iyengar, S. S. and Li, T., ReEL: review aware explanation of location recommendation. in Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, Singapore (UMAP '18), Association for Computing Machinery, New York, NY, USA, (2018), 23–32. https://doi.org/10.1145/3209219.3209237

2. Chen, X., Qin, Z., Zhang, Y. and Xu, T., 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, Pisa, Italy (SIGIR '16), Association for Computing Machinery, New York, NY, USA, (2016), 305–314. https://doi.org/10.1145/2911451.2911549

3. He, X., Chen, T., Kan, M.-Y. and Chen, X., TriRank: review aware explainable recommendation by modeling aspects. in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, Melbourne, Australia (CIKM '15), Association for Computing Machinery, New York, NY, USA, (2015), 1661–1670. https://doi.org/10.1145/2806416.2806504

4. Hong, Y., Lu, J., Yao, J., Zhu, Q. and Zhou, G., What reviews are satisfactory: novel features for automatic helpfulness voting. in Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, Oregon, USA (SIGIR '12), Association for Computing Machinery, New York, NY, USA, (2012), 495–504. https://doi.org/10.1145/2348283.2348351

5. Karatzoglou, A., Amatriain, X., Baltrunas, L. and Oliver, N., Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. in Proceedings of the Fourth ACM Conference on Recommender Systems, Barcelona, Spain (RecSys '10), Association for Computing Machinery, New York, NY, USA, (2010), 79–86. https://doi.org/10.1145/1864708.1864727

6. Kido, S., Sakamoto, R. and Aritsugi, M., Making use of reviews for good explainable recommendation. in Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services (iiWAS '20), Association for Computing Machinery, New York, NY, USA, (2020), 234-240. https://doi.org/10.1145/3428757.3429125

7. Kolda, T. G. and Bader, B. W., Tensor decompositions and applications. SIAM Review, 51 (3), (2009), 455–500. https://doi.org/10.1137/07070111X

8. Koren, Y., Bell, R. and Volinsky, C., Matrix factorization techniques for recommender systems. Computer, 42 (8), (2009), 30–37. https://doi.org/10.1109/MC.2009.263

9. Lu, Y., Tsaparas, P., Ntoulas, A. and Polanyi, L., Exploiting social context for review quality prediction. in Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA (WWW '10), Association for Computing Machinery, New York, NY, USA, (2010), 691–700. https://doi.org/10.1145/1772690.1772761

10. McAuley, J. and Leskovec, J., Hidden factors and hidden topics: understanding rating dimensions with review text. in Proceedings of the 7th ACM Conference on Recommender Systems, Hong

Kong, China (RecSys '13), Association for Computing Machinery, New York, NY, USA, (2013), 165–172. https://doi.org/10.1145/2507157.2507163

11. Moghaddam, S., Jamali, M. and Ester, M., ETF: extended factorization model for personalizing prediction of review helpfulness. in Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, Seattle, Washington, USA (WSDM '12), Association for Computing Machinery, New York, NY, USA, (2012), 163–172. https://doi.org/10.1145/2124295.2124316

12. Montavon, G., Samek, W. and Müller, K.-R., Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, (2018), 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

13. Rayana, S. and Akoglu, L., Collective opinion spam detection: bridging review networks and metadata. in Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia (KDD '15), Association for Computing Machinery, New York, NY, USA, (2015), 985–994. https://doi.org/10.1145/2783258.2783370

14. Rendle, S., Freudenthaler, C., Gantner, Z. and Schmidt-Thieme, L., BPR: Bayesian personalized ranking from implicit feedback. in Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada (UAI '09), AUAI Press, Arlington, Virginia, USA, (2009), 452–461. https://arxiv.org/abs/1205.2618 https://www.cs.mcgill.ca/~uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf

15. Siagian, A. H. A. M. and Aritsugi, M., Robustness of word and character N-gram combinations in detecting deceptive and truthful opinions. Journal of Data and Information Quality, 12 (1), Article 5, (2020). https://doi.org/10.1145/3349536

16. Tang, J., Gao, H., Hu, X. and Liu, H., Context-aware review helpfulness rating prediction. in Proceedings of the 7th ACM Conference on Recommender Systems, Hong Kong, China (RecSys '13), Association for Computing Machinery, New York, NY, USA, (2013), 1–8. https://doi.org/10.1145/2507157.2507183

17. Tintarev, N. and Masthoff, J., Effective explanations of recommendations: user-centered design. in Proceedings of the 2007 ACM Conference on Recommender Systems, Minneapolis, MN, USA (RecSys '07), Association for Computing Machinery, New York, NY, USA, (2007), 153–156. https://doi.org/10.1145/1297231.1297259

18. Tintarev, N. and Masthoff, J., Evaluating the effectiveness of explanations for recommender systems - methodological issues and empirical studies on the impact of personalization. User Modeling and User-Adapted Interaction, 22, (2012), 399–439. https://doi.org/10.1007/s11257-011-9117-5

19. Wang, N., Wang, H., Jia, Y. and Yin, Y., Explainable recommendation via multi-task learning in opinionated text data. in Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA (SIGIR '18),

Association for Computing Machinery, New York, NY, USA, (2018), 165–174. https://doi.org/10.1145/3209978.3210010

20. Yu, R., Zhang, Y., Ye, Y., Wu, L., Wang, C., Liu, Q. and Chen, E., Multiple pairwise ranking with implicit feedback. in Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy (CIKM '18), Association for Computing Machinery, New York, NY, USA, (2018), 1727–1730. https://doi.org/10.1145/3269206.3269283

21. Zhang, Y., and Chen, X., Explainable recommendation: a survey and new perspectives. Foundations and Trends® in Information Retrieval, 14 (1), (2020), 1–101. https://doi.org/10.1561/1500000066

22. Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y. and Ma, S., Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. in Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, Queensland, Australia (SIGIR '14), Association for Computing Machinery, New York, NY, USA, (2014), 83–92. https://doi.org/10.1145/2600428.2609579