Efficiency Improvement of Neutrality-Enhanced Recommendation

Toshihiro Kamishima, Shotaro Akaho,
and Hideki Asoh
National Institute of Advanced Industrial Science
and Technology (AIST)
AIST Tsukuba Central 2, Umezono 1-1-1,
Tsukuba, Ibaraki, 305-8568 Japan
mail@kamishima.net,
s.akaho@aist.go.jp, h.asoh@aist.go.jp

Jun Sakuma
University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305-8577 Japan
jun@cs.tsukuba.ac.jp

ABSTRACT

This paper proposes an algorithm for making recommendations so that neutrality from a viewpoint specified by the user is enhanced. This algorithm is useful for avoiding decisions based on biased information. Such a problem is pointed out as the filter bubble, which is the influence in social decisions biased by personalization technologies. To provide a neutrality-enhanced recommendation, we must first assume that a user can specify a particular viewpoint from which the neutrality can be applied, because a recommendation that is neutral from all viewpoints is no longer a recommendation. Given such a target viewpoint, we implement an information-neutral recommendation algorithm by introducing a penalty term to enforce statistical independence between the target viewpoint and a rating. We empirically show that our algorithm enhances the independence from the specified viewpoint.

Keywords

recommender system, neutrality, fairness, filter bubble, collaborative filtering, matrix factorization, information theory

1. INTRODUCTION

A recommender system searches for items or information that is estimated to be useful to a user based on the user's prior behaviors and the features of items. Over the past decade, such recommender systems have been introduced and managed at many e-commerce sites to promote items sold at those sites. The influence of personalization technologies such as recommender systems or personalized search engines on people's decision making is considerable. For example, at a shopping site, if a customer checks a recommendation list and finds five-star-rated items, he/she will more seriously consider buying these strongly recommended items. These technologies have thus become an indispensable tool for users. However, the problem of *filter bubble*, which is the unintentional bias or the limited diversity of information provided to users, has accompanied the growing influence of personalization algorithms.

The term filter bubble was recently coined by Pariser [12]. Due to the strong influence of personalized technologies, the topics of information provided to users are becoming restricted to those originally preferred by them, and this restriction is not perceived by users. In this way, each individual is metaphorically enclosed in his/her own separate *bubble*. Pariser claimed that users lose the opportunity to find new interests because of the limitations of the bubbles created around their original interests, and that sharing reasonable yet opposing viewpoints on public issues affecting our society is thus becoming more difficult. To discuss this filter bubble problem, a panel discussion was held at the RecSys 2011 conference [14].

During the RecSys panel discussion, panelists made the following assertions about the filter bubble problem. The diversity of topics is certainly biased by the influence of personalization. At the same time, it is impossible to make recommendations that are absolutely neutral from any viewpoint, and thus there is a trade-off between focusing on topics that better fit users' interests or needs and enhancing the varieties of provided topics. To address this problem, the panelists also pointed out several possible directions: taking into account users' immediate needs as well as their long-term needs; optimizing a recommendation list as a whole; and providing tools for perspective-taking.

To our knowledge, there is no major tool that enables users to control their perspective to address this filter bubble problem. We therefore advocate a new *information-neutral recommender system* that guarantees the neutrality of recommendations. As pointed out during the RecSys 2011 panel discussion, it is impossible to make a recommendation that is absolutely neutral from all viewpoints, and we therefore focus on neutrality from a viewpoint or type of information specified by the user. For example, users can specify a feature of an item, such as a brand, or a user feature, such as a gender or an age, as a viewpoint. An information-neutral recommender system is designed so that these specified features will not influence the recommendation results. This system can also be used to ensure fair treatment of content providers or product suppliers or to avoid the use of infor-

RecSys'13, October 12–16, 2013, Hong Kong, China.
Paper presented at the 2013 Decisions@RecSys workshop in conjunction with the 7th ACM conference on Recommender Systems.

mation that is restricted by law or regulation.

Last year at this Decisions workshop, we borrowed the idea of fairness-aware data mining, which we had proposed earlier [8], to build an information-neutral recommender system of the type described above [7]. To enhance neutrality or independence in recommendations, we introduced a constraint term that represents the mutual information between a recommendation result and a specified viewpoint. The naive implementation of this constraint term did indeed enhance the neutrality of recommendations, but there remained serious shortcomings in its scalability. In this paper, therefore, we advocate several new formulations of this constraint term that are more scalable.

Our contributions are as follows. First, we present a definition of neutrality in recommendation based on the consideration of why it is impossible to achieve an absolutely neutral recommendation. Second, we propose a method to enhance the neutrality of a probabilistic matrix factorization model. Finally, we demonstrate that the neutrality of a recommender system can be enhanced.

In section 2, we discuss the filter bubble problem and the concept of neutrality in recommendation, and define the goal of an information-neutral recommendation task. An information-neutral recommender system is proposed in section 3, and the experimental results of its application are shown in section 4. Sections 5 and 6 cover related work and our conclusion, respectively.

2. INFORMATION NEUTRALITY

In this section, we discuss information neutrality in recommendation based on an examination of the filter bubble problem and the ugly duckling theorem.

2.1 The Filter Bubble Problem

We will first summarize the filter bubble problem posed by Pariser and the panel discussion about this problem held at the RecSys 2011 conference. The *Filter Bubble* problem is the concern that personalization technologies narrow and bias the topics of information provided to people, who do not notice this phenomenon [12].

Pariser demonstrated the following examples in a TED talk about this problem [11]. Users of the social network service Facebook specify other users as *friends* with whom they then can chat, have private discussions, and share information. To help users find their friends, Facebook provides a recommendation list of others who are expected to be related to a user. When Pariser started to use Facebook, the system showed a friend recommendation list that consisted of both conservative and progressive people. However, because he more frequently selected progressive people as friends, conservative people were increasingly excluded from his recommendation list by a personalization functionality. Pariser claimed that, in this way, the system excluded conservative people without his permission and that he lost the opportunity to be exposed to a wide variety of opinions.

Pariser's claims can be summarized as follows. First, personalization technologies restrict an individual's opportunities to obtain information about a wide variety of topics. The chance to gain knowledge that could ultimately enhance an individual's life is lessened. Second, the individual obtains information that is too personalized; thus, the amount of shared information and shared debate in our society is decreased. Pariser asserts that the loss of shared information is a serious obstacle for building social consensus. He claimed that the personalization of information thereby becomes a serious obstacle for building consensus.

RecSys 2011 featured a panel discussion on this filter bubble problem [14]. The panel concentrated on the following three points: (a) Are there filter bubbles? (b) To what degree is personalized filtering a problem? and (c) What should we as a community do to address the filter bubble problem? Among these points, we focus on the point (c). The panelists presented several directions to explore in addressing the filter bubble problem. First, a system could consider users' immediate needs as well as their long-term needs. Second, instead of selecting individual items separately, a recommendation list or portfolio could be optimized as a whole. And Finally a system could provide tools for perspective-taking to see the world through other viewpoints.

2.2 Neutrality in Recommendation

Among the directions for addressing the filter bubble, we here take the approach of providing a tool for perspective-taking. Before presenting this tool, we explored the notion of neutrality based on the ugly duckling theorem. The *ugly duckling theorem* is a classical theorem in pattern recognition literature that asserts the impossibility of classification without weighing certain features or aspects of objects as more important than others [17]. Consider a case in which 2^n ducklings are represented by n binary features and are classified into positive or negative classes based on these features. It is easy to show that the number of possible decision rules based on these features to discriminate an ugly duckling and a normal duckling is equal to the number of patterns to discriminate any pair of normal ducklings. In other words, every duckling resembles a normal duckling and an ugly duckling equally. This counterintuitive conclusion is deduced from the premise that all features are treated equally. Attention to an arbitrary feature such as black feathers makes an ugly duckling ugly. When we classify something, we of necessity weigh certain features, aspects, or viewpoints of classified objects. Because recommendation is considered a task for classifying whether items are interesting or not, certain features or viewpoints inevitably must be weighed when making a recommendation. Consequently, the absolutely neutral recommendation is impossible, as pointed out in the RecSys panel.

We propose a neutral recommendation framework other than the absolutely neutral recommendation. Recalling the ugly duckling theorem, we must focus on certain features or viewpoints in classification. This fact indicates that it is feasible to make a recommendation that is neutral from a specific viewpoint instead of all viewpoints. We hence advocate an *information-neutral recommender system* (INRS) that enhances the neutrality in recommendation from the viewpoint specified by a user. In Pariser's Facebook example, a system could enhance the neutrality so that recommended friends are both conservative and progressive, but the system would be allowed to make biased decisions in terms of the other viewpoints, e.g., the birthplace or age of friends.

We formally model this neutrality by the statistical independence between recommendation results and viewpoint values, i.e., $\Pr[R|V] = \Pr[R]$. This means that the same recommendations are made for the cases where all conditions are the same except for the viewpoint values. In other words, no information of viewpoint features influences the recommendation results according to the information theory. An INRS hence tends to be less accurate, because useable information is decreased. In the example of a friend recommendation, no matter what a user's political conviction is, the conviction is ignored and excluded in the process of making a recommendation.

We wish to emphasize that neutrality is distinct from recommendation diversity, which is the attempt to recommend items that are mutually less similar. Topic diversification is one of the proposed techniques for enhancing diversity by excluding similar items from a recommendation list [20]. The constraint term in [19] is designed to exclude similar items from a final list. Therefore, while neutrality involves the relation between recommendations and single viewpoint features, diversity concerns the mutual relation among recommendations. Inversely, enhancing the diversity cannot suppress the use of specific information, and an INRS is allowed to offer mutually similar items. In the case of the friend recommendation, if a progressive person is recommended as a friend, the INRS will recommend another person whose conditions other than political convictions are the same. In the case of the diversified recommendation, one of two persons would not be recommended because the two persons are very similar.

The INRS is beneficial not only for users but also for system managers. It can be used to ensure the fair treatment of content providers or product suppliers. The federal trade commission has been investigating Google to determine whether the search engine ranks its own services higher than those of competitors [3]. E-commerce sites want to treat their product suppliers fairly when making recommendations to their customers. If a brand of providers or suppliers is specified as a viewpoint, a system can make recommendations that are neutral in terms of the items' brands. An information-neutral recommendation is also helpful for avoiding the use of information that is restricted by law or regulation. For example, the use of some information is prohibited for the purpose of making recommendations by privacy policies. In this case, by treating the prohibited information as a viewpoint, recommendations can be neutral in terms of the prohibited information.

3. THE INFORMATION-NEUTRAL RECOMMENDER SYSTEM

We formalize the task of information-neutral recommendation and present an algorithm for performing this task.

3.1 Task Formalization

Recommendation tasks can be classified into three types: *recommending good items* that meet a user's interest, *optimizing the utility* of users, and *predicting item ratings* of items for a user [5]. Among these tasks, we here concentrate on the task of predicting ratings. $X \in \{1, \ldots, n\}$ and $Y \in \{1, \ldots, m\}$ denote random variables for the user and item, respectively. An event (x, y) is an instance of a pair (X, Y). R denotes a random variable for the rating of Y as given by X, and its instance is denoted by r. We here assume that the domain of ratings is the set of real values. These variables are in common with an original predicting ratings task.

To enhance information neutrality in recommendation, we additionally introduced a viewpoint random variable, V, which indicates the viewpoint feature from which the neutrality is enhanced. This variable is specified by a user, and its value depends on various aspects of an event. Possible examples of viewpoint variables are a user's gender, which is part of the user component of an event, a movie's release year, which is part of the item component of an event, and the timestamp when a user rates an item, which would belong to both elements in an event. In this paper, we restrict the domain of a viewpoint variable to a binary type, $\{0, 1\}$, for simplicity. A training sample consists of an event, (x, y), a viewpoint value for the event, v, and a rating value for the event, r. A training set is a set of N training samples, $\mathcal{D} = \{(x_i, y_i, v_i, r_i)\},\ i = 1, \ldots, N$.

Given a new event, (x, y), and its corresponding viewpoint value, v, a rating prediction function, $\hat{r}(x, y, v)$, predicts a rating of the item y by the user x, and satisfies $\hat{r}(x, y, v) = \mathbb{E}_{\Pr[R|x,y,v]}[R]$. This rating prediction function is estimated by optimizing an objective function having three components: a loss function, $\mathrm{loss}(r^*, \hat{r})$, a neutrality term, $\mathrm{neutral}(R, V)$, and a regularization term, reg. The loss function represents the dissimilarity between a true rating value, r^*, and a predicted rating value, $\hat{r}$. The neutrality term quantifies the expected degree of neutrality of the predicted rating values from a viewpoint expressed by a viewpoint feature, and its larger value indicates the higher level of neutrality. The aim of the regularization term is to avoid over-fitting. Given a training sample set, $\mathcal{D}$, the goal of the information-neutral recommendation (predicting rating case) is to acquire a rating prediction function, $\hat{r}(x, y, v)$, so that the expected value of the loss function is as small as possible and the neutral term is as large as possible. We formulate this goal by finding a rating prediction function, $\hat{r}$, so as to minimize the following objective function:

$$\sum_{\mathcal{D}} \mathrm{loss}(r, \hat{r}(x, y, v)) + \eta\, \mathrm{neutral}(R, V) + \lambda\, \mathrm{reg}(\mathbf{\Theta}), \qquad (1)$$

where $\eta > 0$ is a neutrality parameter to balance between the loss and the neutrality, $\lambda > 0$ is a regularization parameter, and $\mathbf{\Theta}$ is a set of model parameters.

3.2 Probabilistic Matrix Factorization Model

In this paper, we adopt a probabilistic matrix factorization model [15] to predict ratings, because this model is highly effective in its prediction accuracy as well as efficient in its scalability. Though there are several minor variants of this model, we here use the following model defined as equation (3) in [9]:

$$\hat{r}(x, y) = \mu + b_x + c_y + \mathbf{p}_x^\top \mathbf{q}_y, \qquad (2)$$

where μ, b_x, and c_y are global, per-user, and per-item bias parameters, respectively, and $\mathbf{p}_x$ and $\mathbf{q}_y$ are K-dimensional parameter vectors, which represent the cross effects between users and items. We then adopt the following squared loss with a regularization term:

$$\sum_{(x_i, y_i, r_i) \in \mathcal{D}} (r_i - \hat{r}(x_i, y_i))^2 + \lambda\, \mathrm{reg}(\mathbf{\Theta}). \qquad (3)$$

This model is proved to be equivalent to assuming that true rating values are generated from a normal distribution whose mean is equation (2). If all samples over all X and Y are available, the objective function is convex; and thereby globally optimal parameters can be derived by a simple gradient descent method. Unfortunately, because not all samples are observed, the loss function (3) is non-convex, and only local optima can be found. However, it is empirically known that a simple gradient method succeeds in finding a good solution in most cases [9].

We then extend this model to enhance the information neutrality. First, we modify the model of equation (2) so that it is dependent on the viewpoint value, v. For each value of V, 0 and 1, we prepare a parameter set, $\mu^{(v)}$, $b_x^{(v)}$, $c_y^{(v)}$, $\mathbf{p}_x^{(v)}$, and $\mathbf{q}_y^{(v)}$. One of the parameter sets is chosen according to the viewpoint value, and we get the rating prediction function:

$$\hat{r}(x,y,v) = \mu^{(v)} + b_x^{(v)} + c_y^{(v)} + \mathbf{p}_x^{(v)\top}\mathbf{q}_y^{(v)}. \qquad (4)$$

By substituting equations (4) into equation (1) and adopting a squared loss function as in the original probabilistic matrix factorization case, we obtain an objective function of an information-neutral recommendation model:

$$\sum_{(x_i,y_i,r_i,v_i)\in\mathcal{D}} (r_i - \hat{r}(x_i,y_i,v_i))^2 + \eta\,\mathrm{neutral}(R,V) + \lambda\,\mathrm{reg}(\mathbf{\Theta}), \quad (5)$$

where the regularization term is a sum of L_2 regularizers of parameter sets for each value of v except for global biases, $\mu^{(v)}$. Model parameters, $\mathbf{\Theta}^{(v)} = \{\mu^{(v)}, b_x^{(v)}, c_y^{(v)}, \mathbf{p}_x^{(v)}, \mathbf{q}_y^{(v)}\}$, for $v \in \{0,1\}$, are estimated so as to minimize this objective. Once we learn the parameters of the rating prediction function, we can predict a rating value for any event by applying equation (4).

3.3 Neutrality Term

Now, all that remains is to define a neutrality term. As described in section 2.2, we formalize the neutrality as the statistical independence between a recommendation result and a viewpoint feature. We propose neutrality terms that are based on mutual information and Calders-Verwer's discrimination score, both of which quantify the degree of independence between R and V.

3.3.1 Mutual Information

We first use the same idea as in [8] and quantify the degree of the neutrality by negative mutual information under the assumption that neutrality can be regarded as statistical independence. Negative mutual information between R and V is defined as:

$$\begin{aligned} -\mathrm{I}(R;V) &= -\sum_V \int \Pr[R,V] \log \frac{\Pr[R|V]}{\Pr[R]} dR \\ &\approx -\frac{1}{N} \sum_{(x_i,y_i,v_i)\in\mathcal{D}} \log \frac{\Pr[\hat{r}_i|v_i]}{\Pr[\hat{r}_i]} \\ &= -\frac{1}{N} \sum_{(x_i,y_i,v_i)\in\mathcal{D}} \log \frac{\Pr[\hat{r}_i|v_i]}{\sum_{v\in\{0,1\}} \Pr[\hat{r}_i|v]\Pr[v]}, \end{aligned} \qquad (6)$$

where $\hat{r}_i$ is derived by applying $(x_i,y_i,v_i) \in \mathcal{D}$ to equation (4). The marginalization over R and V is approximated by the sample mean over $\mathcal{D}$ in the second line, and we use a sample mass function as $\Pr[V]$. $\Pr[R|V]$ can be derived by marginalizing $\Pr[R|X,Y,V]\Pr[X,Y]$ over X and Y. We again approximate this marginalization by the sample mean and get:

$$\Pr[r|v] \approx \frac{1}{|\mathcal{D}^{(v)}|} \sum_{(x_i,y_i)\in\mathcal{D}^{(v)}} \mathrm{Normal}\left(r;\, \hat{r}(x_i,y_i,v), \mathbb{V}_{\mathcal{D}^{(v)}}(R)\right), \quad (7)$$

where $\mathrm{Normal}(\cdot)$ is a pdf of normal distribution, $\mathcal{D}^{(v)}$ consists of all training samples whose viewpoint values are equal to v, and $\mathbb{V}_{\mathcal{D}^{(v)}}(R)$ is a sample variance,

$$\tfrac{1}{|\mathcal{D}^{(v)}|} \textstyle\sum_{r_i\in\mathcal{D}^{(v)}} (r_i - \mathbb{M}_{\mathcal{D}^{(v)}}(\{\hat{r}\}))^2,$$

where $\mathbb{M}_{\mathcal{D}}(\{\hat{r}\})$ is

$$\mathbb{M}_{\mathcal{D}}(\{\hat{r}\}) = \tfrac{1}{|\mathcal{D}|} \textstyle\sum_{(x_i,y_i,v_i)\in\mathcal{D}} \hat{r}(x_i,y_i,v_i).$$

This is very hard to manipulate because this is a mixture distribution with an enormous number of components. We hence took an approach of directly modeling $\Pr[r|v]$, and used two types of models.

The first one is a histogram model, which was proposed in our preliminary work [7]. Though rating values are treated as real values, they are originally discrete scores. Therefore, a set of predicted ratings, $\{\hat{r}_i\}$, are divided into bins. Given a set of intervals, $\{\mathrm{Int}_j\}$, for example $\{(-\infty, 1.5], (1.5, 2.5], \ldots, (4.5, \infty)\}$ in a five-point-scale case, predicted ratings are placed into the bins corresponding these intervals. By using these bins, $\Pr[r|v]$ is modeled by a multinomial distribution:

$$\Pr[\hat{r}|v] \approx \prod_{j=1}^{\#\mathrm{Int}} \left[\frac{\sum_{(x_i,y_i)\in\mathcal{D}^{(v)}} \mathbb{I}[\hat{r}(x_i,y_i,v)\in\mathrm{Int}_j]}{|\mathcal{D}^{(v)}|} \right]^{\mathbb{I}[r\in\mathrm{Int}_j]}, \quad (8)$$

where $\mathbb{I}[r \in \mathrm{Int}]$ is an indicator function and $\#\mathrm{Int}$ is the number of intervals. We refer to this model as mi-hist, which is an abbreviation of *mutual information modeled by a histogram model.*

However, because this model has discontinuous points, we develop a second new approach, which is to model $\Pr[\hat{r}|v]$ by a single normal distribution, which is continuous and easy to handle. Formally,

$$\Pr[\hat{r}|v] \approx \mathrm{Normal}\left(\hat{r};\, \mathbb{M}_{\mathcal{D}^{(v)}}(\{\hat{r}\}), \mathbb{V}_{\mathcal{D}^{(v)}}(\{\hat{r}\})\right), \qquad (9)$$

where $\mathbb{V}_{\mathcal{D}}(\{\hat{r}\})$ is a sample variance over predicted ratings $\hat{r}_i$ from samples in $\mathcal{D}$. We refer to this model as mi-normal, which is an abbreviation of *mutual information modeled by a normal distribution model.*

Unfortunately, it is not easy to derive an analytical form of gradients for these neutrality terms. This is because the discretization is a discontinuous transformation in the mi-hist case, and $\Pr[\hat{r}]$ is a normal mixture, which is not a member of an exponential family, in a mi-normal. We therefore adopt the Powell optimization method for this class of neutrality terms, because it can be applied without computing gradients. However, this optimization method is too slow to apply to a large data set, and its lack of scalability is a serious deficit. In our implementation, these methods failed to complete the processing of 100k data in several days, whereas the methods described in the next section could process this dataset in minutes.

3.3.2 Calders-Verwer's Discrimination Score

To develop a neutrality term whose gradients can be derived in analytical form, we borrowed an idea in discrimination-aware data mining [13]. We here introduce Calders and Verwer's approach used in [2]. They proposed a score to measure the degree of socially discriminative decision, which is here referred by a *CV score*. This CV score is defined as the difference between distributions of target variable given $V = 0$ and $V = 1$.

$$\Pr[R|V = 0] - \Pr[R|V = 1]. \tag{10}$$

To reduce the influence of V on R, they tried to learn a classification model that would make the two distributions, $\Pr[R|V = 0]$ and $\Pr[R|V = 1]$, similar by causing the CV score to approach zero. It is easy to show that this process enforces the statistical independence between V and R [6]. Based on this idea, we design two types of neutrality terms the would make the two distributions $\Pr[R|V = 0]$ and $\Pr[R|V = 1]$ similar.

We design the first type of neutrality term so as to match the first-order moment of the two distributions, i.e., the means. It is formally defined as

$$-(\mathbb{M}_{\mathcal{D}^{(0)}}(\{\hat{r}\}) - \mathbb{M}_{\mathcal{D}^{(1)}}(\{\hat{r}\}))^2. \tag{11}$$

We refer to this neutrality term as m-match, which is an abbreviation of *mean matching.* The second type is designed to constrain so that the same ratings are predicted for the same value pair, x and y, irrelevant of the viewpoint values. This neutrality term is formally defined as

$$-\sum_{(x_i, y_i) \in \mathcal{D}} (\hat{r}(x_i, y_i, 0) - \hat{r}(x_i, y_i, 1))^2. \tag{12}$$

We refer to this neutrality term as r-match, which is an abbreviation of *rating matching.*

Because both types of neutrality terms are simple quadratic polynomials, it is very easy to derive analytical forms of their derivatives. We hence used a conjugate gradient method for these neutrality terms in optimization, which is much more efficient than the Powell method. Even if the size of data set becomes larger, more scalable optimizers, e.g., a stochastic gradient method, can be used because the gradients can be analytically calculated.

These terms have the additional merit of being less frequently trapped by local minima, because they are simple quadratic formulae. Conversely, it is not straightforward to extend these CV-score-based neutrality terms so that they are applicable to the case in which a viewpoint variable is multivariate discrete or continuous, as in mutual-information-based neutrality terms. When comparing m-match and r-match, the computation time for r-match is roughly twice that for m-match, because a rating prediction function must be evaluated for the cases of both $V = 0$ and $V = 1$ to compute r-match. r-match more strictly formulates neutrality than m-match. In the case of m-match, because the neutrality is enhanced on average over the user population, the neutrality of one user might be greatly enhanced, but that of the other might not. On the other hand, r-match is designed so that neutrality is uniformly enhanced almost everywhere over the domain of users and items. Unlike m-match, r-match treats counterfactual cases. For example, when the gender of a user is a viewpoint, even though the gender does not change, ratings in such a counterfactual case must be computed for using the r-match term. This fact may be semantically improper.

4. EXPERIMENTS

We implemented our information-neutral recommender system and applied it to a benchmark data set. We examined the four types of neutrality terms proposed in section 3.3.

4.1 Data Set

We used a Movielens 100k data set [4] in our experiments. Unfortunately, neither of the mutual-information-based methods in section 3.3.1, mi-hist and mi-normal, were able to process this entire data set. Therefore, we shrank the Movielens data set by extracting events whose user ID and item ID were less than or equal to 200 and 300, respectively. For scalable m-match and r-match methods, we applied a larger data set as described in section 4.4. This shrunken data set contained $9,409$ events, 200 users, and 300 items. The purpose of experiments on this small set was to compare the characteristics of all four neutrality terms. The mutual-information-based methods more strictly modeled the distribution over R and V than the CV-score-based methods described in section 3.3.2, m-match and r-match. If the CV-score-based methods behaved similarly to the mutual-information-based methods, we would be able to conclude that CV-score-based methods can enhance the neutrality and are scalable.

We tested the following two types of viewpoint variable. The first type of variable, Year, represents whether a movie's release year is newer than 1990, which is part of the item component of an event. In [10], Koren reported that older movies have a tendency to be rated more highly, perhaps because masterpieces are more likely to survive, and thus the set of older movies has more masterpieces. When adopting Year as a viewpoint variable, our recommender enhances the neutrality from this masterpiece bias. The second type of variable, Gender, represents the user's gender, which is part of the user component of an event. We expect that the movie ratings would depend on the user's gender.

4.2 Experimental Conditions

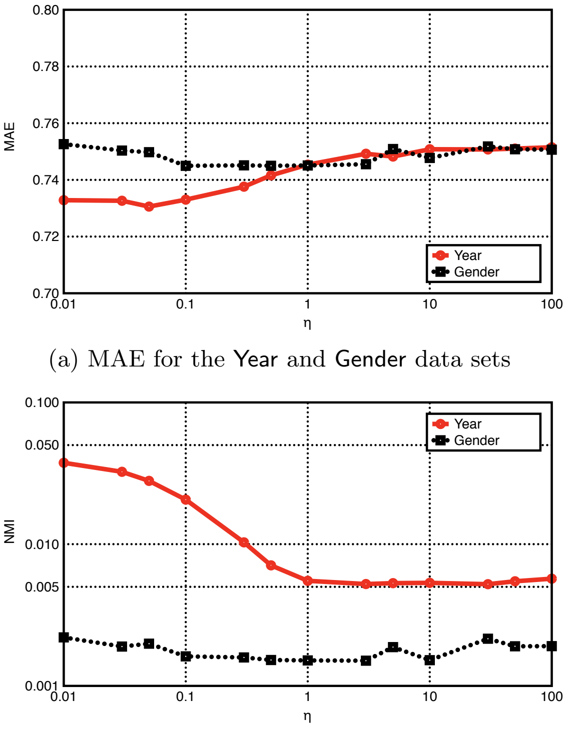
We optimized an objective function (5) with neutrality terms mi-hist or mi-normal by the Powell method, and that with terms m-match or r-match by the conjugate gradient method implemented in the SciPy package [16]. To initialize the model parameters, events in a training set, $\mathcal{D}$, were first divided into two sets according to their viewpoint values. For each value of a viewpoint variable, the parameters were initialized by minimizing an objective function of an original probabilistic matrix factorization model (equation (3)). For convenience in implementation, a loss term of an objective was re-scaled by dividing it by the number of training examples, and an L_2 regularizer was scaled by dividing it by the number of parameters. The four types of neutrality terms were re-scaled so that the magnitudes of these terms became roughly equal. Because the original rating values are $1, 2, \ldots, 5$, we adopted five bins $(-\infty, 1.5], (1.5, 2.5], \ldots, (4.5, \infty)$ for the mi-hist term. We use a regularization parameter $\lambda = 0.01$ and the number of latent factors, $K = 1$, which is the size of vectors $\mathbf{p}^{(v)}$ or $\mathbf{q}^{(v)}$. It should be notice that this data set was so small that the prediction performance was degraded if $K > 1$. Though in the case without cross term, i.e., $K = 0$, the performance was better than the case where $K = 1$, but we tested the model having the minimum cross terms. Our experimental codes are available at `http://www.kamishima.net/inrs/`.

We evaluated our experimental results in terms of prediction errors and the degree of neutrality. Prediction errors were measured by the mean absolute error (MAE) [5]. This index was defined as the mean of the absolute difference between the observed rating values and predicted rating values. A smaller value of this index indicates better prediction accuracy. To measure the degree of neutrality, we adopted mutual information between the predicted ratings and viewpoint values. The smaller mutual information indicates a higher level of neutrality. Mutual information is normalized into the range $[0, 1]$ by employing the geometrical mean as described in [6]. Note that the distribution $\Pr[\hat{r}|v]$ is required to compute this mutual information, and we used the same histogram model as in equation (8). We performed a five-fold cross-validation procedure to obtain evaluation indices of the prediction errors and the neutrality measures.

4.3 Experimental Results

Experimental results for the four types of neutrality terms are shown in Figure 1. The MAE was 0.903, when the rating being offered was held constant at 3.74, which is the mean rating over all ratings in the training data. This approximately simulates the case of randomly recommending items, and can be considered as the most unbiased and unintentional recommendation. We call this case *random prediction*. On the other hand, when applying the original probabilistic matrix factorization model (equation (2)), the MAE was 0.759. Because the trade-off for enhancing the neutrality generally worsens the prediction accuracy the accuracy as discussed in section 2.2, this error level can be considered as the lower bound. We call this case *basic prediction*.

In Figures 1(a) and (c), the prediction errors were better than random predictions. Overall, the increase of MAEs as the neutrality parameter, η, was not very great in any of the neutrality terms. The errors for the r-match term sometimes decreased even if η was increased. As described in section 4.2, the model without cross terms better performed. We think that the cross term effects would be eliminated by the strong restriction of the r-match terms, and MAEs were improved. Turning to Figure 1(b) and (d), the results obtained with the r-match term and with the other three terms were clearly contrasted. The three terms, mi-hist, mi-normal, and m-match, yielded successfully enhanced neutrality for the Year data, but less enhanced neutrality for the Gender data. Conversely, the r-match term was able to enhance neutrality for the Gender data, but it failed to do so for the Year data. We expected that this distinction was caused by the original independence between predicted ratings and viewpoint values. By comparing the NMIs at $\eta = 0.01$ of Figures 1(b) and (d), it was found that the dependence between ratings and viewpoint values for the Year data was larger that for the Gender data. Additionally, as described in section 3.3.2, while the r-match term is designed so that neutrality is uniformly enhanced over the domain of users and items, the other three terms are designed so as to enhance neutrality on average. In the case of the Year data, the three terms could enhance neutrality on average because the neutrality was low when η was small. However, the restriction of the r-match term was expected to be too strong for this data set. On the other hand, because the averaged neutrality for the Gender data was high at the beginning, the three terms failed to improve the neutrality, but the stronger neutrality-enhancement ability of the r-match would be effective in this case.



(a) MAE for the Year and Gender data sets

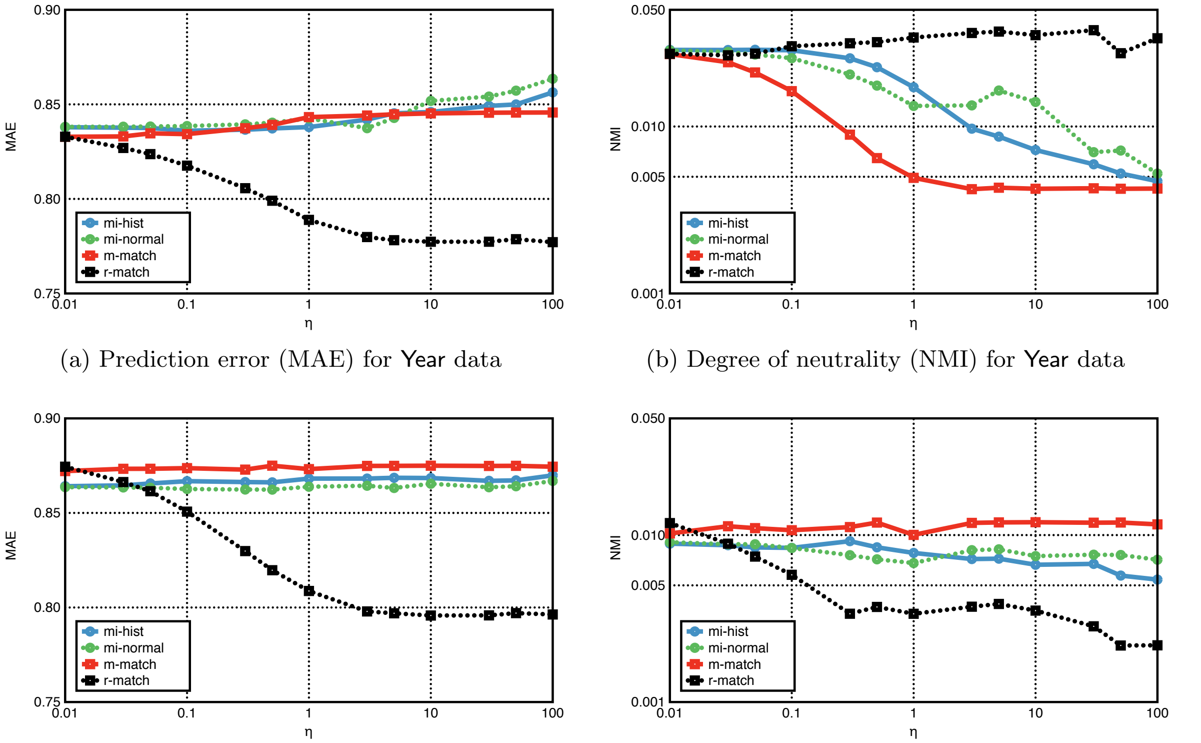
(b) NMI for the Year and Gender data sets

Figure 3: Changes of prediction errors and neutrality measures

To further investigate this phenomenon, we show the changes of mean predicted ratings in Figure 2. Two types of neutrality terms, m-match and r-match, were examined. First, we focus on the case where $\eta = 0.01$, in which the neutrality term was less influenced. By comparing Figures 2(a) and (b), the difference between the mean ratings for old and new movies was much larger than the difference between the mean ratings rated by male and female users. In particular, while the former difference was 0.36, the latter difference was 0.024. This result again indicates that a higher level of neutrality is achieved for the Gender data than for the Year data. For the Year data, the m-match term successfully reduced the difference of two means as the increase of η, but the r-match term failed to do so. For the Gender data, both terms failed to reduce the difference between the two means, because the difference was already small and constraint terms were not effective.

4.4 Experiments on a Larger Data set

To show that our new neutrality terms are applicable to larger data sets, we made an INRS on the entire Movielens 100k data set, which contains 10 times as much data as the data set in the previous section. In our preliminary work [7], a data set of this size could not be processed. MAE of random and basic predictions for this data set were 0.945 and 0.750, respectively. We adopted the m-match neutrality term, and the other conditions were set as in section 4.2 except for $K = 3$. Figure 3 shows the changes of the MAE and NMI according to the increase of η. Trends similar to those in Figure 1 were observed. While neutrality was successfully enhanced without sacrificing prediction errors for the Year data, the m-match term was not effective for



(a) Prediction error (MAE) for Year data

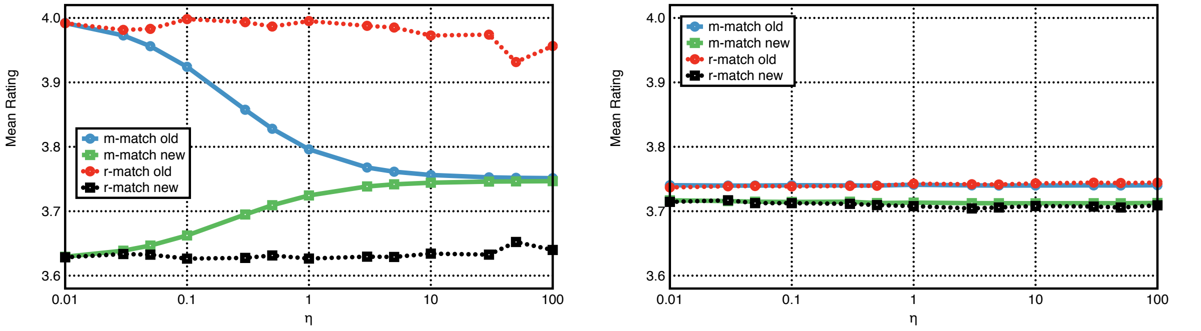
(b) Degree of neutrality (NMI) for Year data

(c) Prediction error (MAE) for Gender data

(d) Degree of neutrality (NMI) for Gender data

Figure 1: Changes of the degrees of neutrality accompanying the increase of a neutrality parameter

NOTE : Subfigures (a) and (b) are results on the Year data set, and Subfigures (c) and (d) are results on the Gender data set. Subfigures (a) and (c) show the changes of prediction errors measured by the mean absolute error (MAE in a linear scale). A smaller value of this index indicates better prediction accuracy. Subfigures (b) and (d) show the changes of the normalized mutual information (NMI in a log scale). A smaller NMI indicates a higher level of neutrality. The X-axes (log-scale) of these figures represent the values of a neutrality parameter, η, which balance the prediction accuracy and the neutrality. These parameters were changed from 0.01, at which the neutrality term was almost completely ignored, to 100, at which the neutrality was strongly enhanced.



(a) Year

(b) Gender

Figure 2: Changes of mean predicted ratings accompanying the increase of a neutrality parameter

NOTE : In both figures, the X-axes (log-scale) represent the values of a neutrality parameter, η, and the Y-axes represent mean predicted ratings for each case with a different viewpoint value. Subfigure (a) shows mean the predicted ratings when the viewpoint variable is Year. Means for the movies before 1990 were designated as “old,” and those after 1991 were designated as “new.” Subfigure (b) shows the mean predicted ratings when the viewpoint variable is Gender. Means of the ratings given by males and females were represented by “M” and “F,” respectively.

the Gender data.

Finally, we should comment on the computational time. Generally, terms based on mutual information were much slower than those based on CV score. This is because analytical forms of gradients can be derived for the m-match and r-match. In comparing the two terms, m-match and r-match, the former is found to be faster, as described in section 3.3.2. Empirically, as η increased, the convergence of optimizers became slower, because the neutrality terms were not smooth compared to the loss term and harder to optimize. The influence of the increase of η was more serious for the r-match than for the m-match.

5. RELATED WORK

We adopted techniques for fairness-aware or discrimination-aware data mining to enhance the neutrality. Fairness-aware data mining is a general term for mining techniques designed so that sensitive information does not influence the mining results. Pedreschi et al. first advocated such mining techniques, which emphasized the unfairness in association rules whose consequents include serious determinations [13]. Another technique of fairness-aware data mining focuses on classification designed so that the influence of sensitive information on classification results is reduced [8, 2]. These techniques would be directly useful in the development of an information-neutral variant of content-based recommender systems, because content-based recommenders can be implemented by standard classifiers.

Because information-neutral recommenders can be used to avoid the exploitation of private information, these techniques are related to privacy-preserving data mining [1]. To protect private information contained in rating information, dummy ratings were added [18].

6. CONCLUSION

In this paper, we proposed an information-neutral recommender system that enhances neutrality from the viewpoint specified by a user. This system is useful for alleviating the filter bubble problem. We then developed an information-neutral recommendation algorithm by introducing several types of neutrality terms. Because the neutrality term in our preliminary work had poor scalability, we proposed a new and more efficient neutrality term. Finally, we demonstrated that neutrality in recommendation could be enhanced by our algorithm without sacrificing the prediction accuracy.

There are many functionalities required for this information-neutral recommender system. We plan to explore the other types of neutrality terms that can more exactly evaluate the independence between a target variable and a viewpoint variable while maintaining efficiency. Because viewpoint variables are currently restricted to binary type, we also try to develop a neutrality term that can deal with a viewpoint variable that is multivariate discrete or continuous. Though our current technique is mainly applicable to the task of predicting ratings, we will develop another algorithm for the task of recommending good items.

7. ACKNOWLEDGMENTS

We would like to thank for providing a data set for the Grouplens research lab. This work is supported by MEXT/JSPS KAKENHI Grant Number 16700157, 21500154, 23240043, 24500194, and 25540094.

8. REFERENCES

[1] C. C. Aggarwal and P. S. Yu, editors. *Privacy-Preserving Data Mining: Models and Algorithms*. Springer, 2008.

[2] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21:277–292, 2010.

[3] S. Forden. Google said to face ultimatum from FTC in antitrust talks. Bloomberg, Nov. 13 2012. ⟨http://bloom.bg/PPNEaS⟩.

[4] Grouplens research lab, university of minnesota. ⟨http://www.grouplens.org/⟩.

[5] A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962, 2009.

[6] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Considerations on fairness-aware data mining. In *Proc. of the IEEE Int'l Workshop on Discrimination and Privacy-Aware Data Mining*, pages 378–385, 2012.

[7] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Enhancement of the neutrality in recommendation. In *Proc. of the 2nd Workshop on Human Decision Making in Recommender Systems*, pages 8–14, 2012.

[8] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Proc. of the ECML PKDD 2012, Part II*, pages 35–50, 2012. [LNCS 7524].

[9] Y. Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 426–434, 2008.

[10] Y. Koren. Collaborative filtering with temporal dynamics. In *Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 447–455, 2009.

[11] E. Pariser. The filter bubble. ⟨http://www.thefilterbubble.com/⟩.

[12] E. Pariser. *The Filter Bubble: What The Internet Is Hiding From You.* Viking, 2011.

[13] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 560–568, 2008.

[14] P. Resnick, J. Konstan, and A. Jameson. Panel on the filter bubble. The 5th ACM Conf. on Recommender Systems, 2011. ⟨http://acmrecsys.wordpress.com/2011/10/25/panel-on-the-filter-bubble/⟩.

[15] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems 20*, pages 1257–1264, 2008.

[16] Scipy.org. ⟨http://www.scipy.org/⟩.

[17] S. Watanabe. *Knowing and Guessing – Quantitative Study of Inference and Information.* John Wiley & Sons, 1969.

[18] U. Weinsberg, S. Bhagat, S. Ioannidis, and N. Taft. Blurme: Inferring and obfuscating user gender based on ratings. In *Proc. of the 6th ACM Conf. on Recommender Systems*, pages 195–202, 2012.

[19] M. Zhang and N. Hurley. Avoiding monotony: Improving the diversity of recommendation lists. In *Proc. of the 2nd ACM Conf. on Recommender Systems*, pages 123–130, 2008.

[20] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proc. of the 14th Int'l Conf. on World Wide Web*, pages 22–32, 2005.