Uncovering Response Biases in Recommendation

Kyung-Wha Park¹*, Byoung-Hee Kim²*, Tae-Suh Park³, Byoung-Tak Zhang¹,²,³

¹Brain Science Program, ²Department of Computer Science and Engineering, ³Cognitive Science Program, Seoul National University, Seoul 151-742, Republic of Korea
{kwpark, bhkim, tspark, btzhang}@bi.snu.ac.kr

Abstract
An user-specific tendency of biased movie rating is investigated, leading six identified types of rating pattern in a massive movie rating dataset. Based on the observed bias assumption, we propose a rescaling method of preferential scores by considering the rating types. Experimental results show significant enhancement for movie recommendation systems.

Introduction
Recommender systems in this information-overloading age have become mandatory tool for various online services on music, books, movies, etc. Automatic recommendation methods and systems are now hot research topics in various fields throughout machine learning, data mining, cognitive science, marketing, and artificial intelligence.

Recommender systems analyze patterns of user interest in products or items to build users’ preference models and provide personalized recommendations that suit a user’s taste (Koren, Bell, & Volinsky, 2009). Widely used materials on user preferences are users’ explicit rating values on items, for example 5-star ratings. Large databases on rating values are used to build automated recommender systems based on various computational analytic methods, i.e. machine learning and statistical pattern mining. These rating values are produced by complex interactions between users with various tastes and vast selection of items. Learning personal preferences from these ratings requires multidisciplinary studies.

Rating values contain inherent biases caused by users, items and interaction between user and item factors. For example, critical users give low ratings than normal raters and the popularity of an item causes a higher rating than average items (Shan & Banerjee, 2010). Although recommender systems have achieved highly personalized performances with vast explanation power, they mostly overlook these biases (Koenigstein, Dror, & Koren, 2011). For real personalization and fast adaptation for personal context, these biases should be uncovered and explained.

Previous studies exist on modeling biases of ratings in the recommendation field. Pennock et al. (2000) propose the personality diagnosis method for collaborative filtering, in which internal preferences for items are described as the personality type and encoded as a vector of user’s “true” ratings for all the rated items. They express all the bias factors as Gaussian noise. Shan & Banerjee (2010) model user biases and item biases using residual approach, and take off these biases before applying their latent factor based collaborative filtering approaches. They model user biases and item biases as Gaussian noises per each user and item. Koenigstein et al. (2011) point out that previous approaches like Shan & Banerjee (2010) do not give personalized ranking results. They identify biases that related and unrelated to personal features, and propose models to clean signals unrelated to personalization purposes, such as item taxonomy, user’s rating in a consecutive session, and item’s temporal dynamics. In these studies, researchers assume that rating values are in ratio scale (Stevens, 1946) even though the true property of ratings may be in interval or ordinal scale. Some preprocessing should be applied which convert rating values into clear interval scales or ratio scales to properly apply statistical tools for recommendations.

As a mandatory step to deal with rating data, we focus on cognitive biases for which cognitive science, social psychology, and behavioral economics studies have revealed, especially studies on response styles or response biases in surveys and questionnaires (Cannell, Miller, & Oksenberg, 1981; Cronbach, 1946; Krosnick, 1991; Lentz, 1938). From these studies, various types of responses are revealed, and in some following studies researchers attempt to detect and correct biases by developing computational models. Greenleaf (1992) focus on detecting and correcting biased components without attitude
information. Nowlis et al. (2002) aim to make data away from the midpoint response bias, excluding the neutral option. More recently, Rosmalen et al. (2010) suggested a latent-class-based regression model to identify biases from different types of response style and content of items. Despite notable differences between responses of questionnaires and online ratings (explained in the discussion section in detail), we suggest that a similar set of rating types for recommendations can be developed.

In this paper, we propose that i) rating values in recommender systems is a combined output of users' intrinsic preferences and exposed attitudes, ii) rating patterns by active users can be categorized into six subtypes, iii) user biases can be modeled based on these types, and iv) reducing user biases reveals user preferences on an objective scale, hence enhances recommendation. We present a model for six rating types, called RBIAS, and apply the model to movie recommendation datasets to show efficiencies.

The rest of the paper is organized as follows. Our model of RBIAS is defined and a rescaling method for bias reduction and normalization is presented. Utilities of RBIAS for user profiling and item profiling are presented within the movie recommendation domain. We further show that the rescaling method results in improvements in movie recommendations. Discussions follow including the similarities and differences of response values in recommendations and surveys, the relation between RBIAS and response styles, and relation of RBIAS with inter-rater reliability approaches. We conclude this paper with a summary and future work.

**RBIAS: the Model of Rating Types**

In most recommender systems, active users who join the system and usually choose what they want to rate give explicit ratings. Various biases intrude in this stage including user interfaces, user's affection states, or attributes of items. However, these biases are hidden and we need to analyze the observed patterns first to estimate the biases. By observing distributions of rating values of users, we can derive various types of rating patterns as shown in Figure 1. Two major rating types are like-biased (LB) and dislike-biased (DB). LB-type users may rate movies they like or tend to give higher ratings to most movies. They may want the movie to receive rewards from their ratings. Users of the DB type may mainly rate movies they hate and may feel it duty to prevent other movie-watchers from movies they think are awful. This LB-DB pair is the counterpart of the well-known pairs in consumer surveys, i.e. the yeasaying and naysaying pair (Greenleaf, 1992) or the acquiescent and disacquiescent pair (Rasmussen et al., 2010). Some users show both LB and DB actions and these users can be categorized as bi-polar (BP). Users with extreme indecisive responses are assigned to the neutral-biased (NB) type. Many statistical methods for recommendation assume that a user's rating patterns follow the normal (Gaussian) distribution. We set the type of these users as normal (N). Finally, we design a vague rating type (V) for those users whose ratings do not show any specific patterns described above.

With these observations, we introduce the model of rating types in recommendation, which we call RBIAS, based on users' intrinsic preferences and exposed attitudes (analogous with the intrinsic / extrinsic preference setting by (von Wright G.H., 1963)). Intrinsic preferences are modeled as having bi-level states which may be described as generous and stubborn. Exposed attitude in ratings can be intuitively categorized into three levels and we describe these states as picky, indecisive, and dry-minded (mechanical response, either too sincere or indifferent). On this preference-attitude plane, six rating types are depicted in Figure 2. Relation of proposed rating types and response styles in customer surveys are explained in Discussion.

![Figure 1: Conceptual diagram of the six rating patterns defined in the RBIAS model (x-axis: ratings in ordinal or interval scale, y-axis: frequencies of ratings).](image)

**Figure 2: Diagram of six rating types in RBIAS – LB, DB, BP, NB, V, and N, as the combination of two personal factors, i.e. intrinsic preference and exposed attitude. Response styles revealed from studies on customer surveys are added in smaller texts for comparison.**
Table 1: Deterministic modeling rules for LB, DB, BP, and NB rating types in RBIAS. For each column, four statistical conditions should be satisfied (* means ‘don’t care’). Normal and vague types are determined in different ways described in the text.

<table>
<thead>
<tr>
<th>Types</th>
<th>LB</th>
<th>DB</th>
<th>BP</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>&gt;= 3.5</td>
<td>&lt;= 2.5</td>
<td>*</td>
<td>&gt;= 2.5 &amp; &lt;= 3.5</td>
</tr>
<tr>
<td>Mean</td>
<td>&gt;= 3.5</td>
<td>&lt;= 2.5</td>
<td>&gt;= 2.5 &amp; &lt;= 3.5</td>
<td>*</td>
</tr>
<tr>
<td>Stdev</td>
<td>&gt; 0</td>
<td>&lt; 0</td>
<td>&gt;= 1.43</td>
<td>&lt; 1.3</td>
</tr>
<tr>
<td>Skewness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As a concrete tool for the RBIAS model, we present a simple set of deterministic rules to determine users’ rating types. Given a set of observed ratings of one user, we first check if the rating type is normal using any normality test (in this paper we apply the Lilliefors test). If the normality test fails, we proceed to apply the rule set in Table 1. If the user does not belong to any type in this stage, the user is labeled with the vague rating type.

Bias Reduction by Rescaling

We proceed to deal with biases hidden in rating values based on the RBIAS model. We first elicit rating-type-dependent biases and propose a rescaling method to reduce these biases. The objective of rescaling is to reduce biases and to get comparable rating values among users.

We set a simple ‘generative’ model for biases. Each type is represented as one hidden variable and one observed variable. Scales of observed variables are fixed, but hidden variables for rating types have their own ranges. During the generation process from hidden to observed, specific biases are added. For the 5 star rating system (Likert scale for questionnaires), we propose a hidden variable model with the exact interval scale as described in Figure 3. For comparison between the unbiased hidden representation of ratings and observed ones, we fix the minimum and maximum ratings on the observed scale, i.e. 1 and 5. For example, the rating value 2 by an LB-type user is transformed into 1.67 and rating value 4 by a NB-type user becomes 3.67.

The intuition behind the bias reduction rules in Figure 3 is as follows. For an LB-type user, we assume that observed rating values less than or equal to 3 are positively biased compared to the normal type and we remove these biases by translating them to the lower scale. In case of a DB-type user, observed rating values over 3 are assumed to be negatively biased and we translate them toward the higher scale. Rating values 1 and 5 for BP and NB types are translated toward extreme directions with the similar assumption.

Experiments

We apply the RBIAS model to the movie recommendation domain. Three public datasets of movie ratings are analyzed based on the model. We show how RBIAS reveals the structure behind the popularity of movies. Finally, we show how the bias-reduced rating dataset results in improvements of recommender systems.

Datasets

We choose three widely used datasets on movie recommendation that contain 5-star ratings of which statistics are summarized in Table 2.

A MovieLens 100K set (denoted as ML-100K) from September 1997 to April 1998 was collected through the MovieLens website. Users login to the website and choose movies to rate from one star (Awful) to five stars (Must See). A MovieLens 1M set (ML-1M) was collected from MovieLens users who joined and used MovieLens in 2000.

Netflix small dataset (NetflixS) is a subset of the Netflix dataset for the Netflix Prize which was collected during October 1998 to December 2005. We use the dataset provided with PREA toolkit (Lee, Sun, & Lebanon, 2012).

Table 2: Basic statistics of the datasets for movie recommendation: MovieLens 100K set, MovieLens 1M set, and Netflix small set

<table>
<thead>
<tr>
<th></th>
<th>ML-100K</th>
<th>ML-1M</th>
<th>NetflixS</th>
</tr>
</thead>
<tbody>
<tr>
<td># of users</td>
<td>943</td>
<td>6,040</td>
<td>4,102</td>
</tr>
<tr>
<td># of items</td>
<td>1,682</td>
<td>3,952</td>
<td>3,000</td>
</tr>
<tr>
<td># of ratings per user</td>
<td>106.04</td>
<td>165.60</td>
<td>33.86</td>
</tr>
<tr>
<td># of ratings per item</td>
<td>59.45</td>
<td>253.09</td>
<td>97.77</td>
</tr>
</tbody>
</table>
Rating Type Analysis on Movie Datasets

We apply the set of deterministic rules to the movie rating datasets to demonstrate the successful performance. We first analyze the distribution of rating scores for each rating type. Figure 4 shows the result from the MovieLens 100K dataset. As can be seen in the figure, the distribution matches well to the conceptual definition of rating types in Figure 1. For example, almost 70% of LB-type users’ ratings are over 3, which means that users that are assigned as LB really tend to give higher ratings to movies. On the other hand, DB-type users give almost 70% of their ratings under 3. BP users rate 3 less often than 4, 5, and 1. The proportion of 3 is the largest in NB.

By aggregating users’ response types, we can analyze the rating tendencies of group of users or the whole user in a specific domain. For illustration, the distribution of six rating types for each dataset is summarized in Table 3. Note that we can assume all the users in the three datasets are exclusive. The majority of users show like-biased (LB) rating styles, meaning that many participants are very picky and their intrinsic preferences may be generous. The second largest group is those who show indecisive rating behaviors and categorized as neutral-based (NB) types. Users with vague rating types follow. The ratios of N, DB, and BP type users are considerably minor and this trend is remarkable in the NetflixS dataset.

### Table 3: Relative frequencies of rating types in various datasets for movie recommendation (%).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>LB</th>
<th>DB</th>
<th>BP</th>
<th>NB</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML-100K</td>
<td>3.5</td>
<td>53.0</td>
<td>1.5</td>
<td>2.0</td>
<td>22.0</td>
<td>18.0</td>
</tr>
<tr>
<td>ML-1M</td>
<td>0.6</td>
<td>65.8</td>
<td>0.9</td>
<td>2.0</td>
<td>17.2</td>
<td>13.5</td>
</tr>
<tr>
<td>NetflixS</td>
<td>0.8</td>
<td>56.8</td>
<td>0.5</td>
<td>1.2</td>
<td>26.2</td>
<td>14.5</td>
</tr>
</tbody>
</table>

RBIAS-based Item Profiling

RBIAS is a user-bias oriented model. The consideration that user factors and item factors interact in rating scores, application of the rating type detection rules to users and inspection of the aggregate pattern for specific item may present us semantic attributes such as popularity.

As an illustrating example, we select two famous movies – Toy Story and L.A Confidential which are referred to as masterpieces of the 1990s. Both of their average ratings in IMDb, one of the most famous internet movie database site, are the same – 8.3.

![Figure 5: Proportion of rating types per each rating score for Toy Story in ML-100K set](image)

As depicted in Figure 5, ratings by the LB-type take up a relatively large proportion in the scores 4 and 5. NB types are largely observed around scores 3 and 4, which shows their natural traits. Note that DB-type users take 25% in score 1. Considering that DB types are rare, only 1.5%, it is relatively larger proportion than other types in the lower scores. It is interesting to note that, in spite of their penalty-minded traits, 2 out of 4 DB-type users rated 3 for this movie. This kind of rule-breaking case is rarely observed and we conjecture that this may be an evidence of the popularity.

![Figure 6: Proportion of rating types per each rating score for L.A Confidential in ML-100K set](image)

Figure 6 shows another rule-breaking case that where the DB-type rated this movie with high scores, i.e. 4 out of 7. The DB-type users gave scores 4 and 5.

![Figure 7: Proportion of rating types per each rating score for Batman Forever in ML-100K set](image)

In Figure 7, we can find the opposite case in which many LB-type users gave low scores. For this movie, LB users didn’t hesitate to “punish” this movie, *Batman Forever*, of which IMDb score is 5.4.
In summary, if a movie acquires mainly high scores regardless of the rating types, as in the case of LA Confidential, this movie can be said “popular”. By checking the rating patterns of LB and DB-types, the popularity of movies may be evaluated. For example, although the IMDb scores are the same, LA Confidential might be more popular than Toy Story. Batman Forever for which even LB-type users tend to give low scores may be labeled as “not popular”.

**RBIAS-based Rating Bias Correction**

In this section, we show that RBIAS-based correction of rating bias can improve performances of recommender systems. For comparison, we prepared two ML 100K datasets, original and RBIAS-based rescaled ones. Two modes of recommendation are tested. The first mode is the binary preference prediction in which predicted rating score larger than 3 is determined as positive and other cases are labeled as negative. Usual metrics for this setting are checked – precision, recall, and F1 measure. The second mode is regression-based approaches for rating score prediction. Mean absolute errors (MAE) and root-mean-squared errors (RMSE) are compared. We additionally check ranking-based metrics such as half-life utility (HLU) and normalized discounted cumulative gain (NDCG). PREA recommendation toolkit (Lee et al., 2012) is used for fair comparison. For more information on algorithms and evaluation metrics in this experiment, refer to the PREA website (prea.gatech.edu). State-of-the-art collaborative filtering methods show similar or higher performances when RBIAS is applied (Figure 8). Reduction of MAE and RMSE are results of type-dependent bias reduction which drives ratings in various ranges align into a common, fair scale. There is another factor of rescaling from 5-point scale to 7-point scale that may affect the evaluation results. Improvements in NDCG, i.e., improved detection of preferences, show that alignment by RBIAS-based bias correction may be the more prominent factor than rescaling to the finer range.

**Discussion**

RBIAS is a model of rating types in recommendation that account for user biases. As presented in previous sections, rating types are closely related to ‘response styles’ that have been studied in social psychology and marketing research fields. The response style is defined as the systematic tendency to respond to a range of questionnaire items with some biases, not mis-understanding or misdetermining, other than the specific item content (Paulhus, 1991). Commonly known response styles are acquiescent response styles (ARS) that tend to agree with items regardless of the contents, disacquiescent response style (DARS) which is the opposite type of ARS, midpoint responding (MPR) that tend to use the middle scale category regardless of contents, extreme response styles.

![Figure 8: Recommendation performance comparison with the MovieLens 100K dataset under 5-fold cross validation.](image)

Left, white bar and right, black bars represent performances on the original and RBIAS-based rescaled data, respectively. Precision, Recall, and F1 measure are evaluated under binary preference classification setting. MAE and RMSE are regression-based error measures. Finally, HLU and NDCG are rank-based measures (the larger, the better). Precision, F1 measure, and HLU does not show significant differences (p-value >0.05 under paired t-test). For MAE, RMSE and NDCG, the rescaled case show significant improvements (p-value ≪ 0.05)
(ERS) that tend to endorse the most extreme response categories regardless of the contents, rating range (RR) that tend to use a narrow or wide range of response around the mean response, and noncontingent responding (NCR) that tend to respond to items carelessly, randomly, or non-purposefully (Baumgartner & Steenkamp, 2001). As depicted in Figure 2, six rating types in the RBIAS model have corresponding response styles that result in similar observed rating patterns.

Despite the common characteristics among rating types and response styles, they have apparent differences and these differences illuminate the novelty and importance of RBIAS rating types. Acquiescent action is very common in survey tasks such that a responder tends to accept or allow what other people want or demand. However, rating in recommender system is not acquiescent, because the rater actively selects items and there are not explicit demands. This difference is well exposed when we consider NCR values. Rating types in RBIAS are expected to have higher NCR values compared to that of the response styles. The relation between two response types are not one-to-one and one response type is related to a few response styles.

Inter-rater reliability and related random effect models (Hallgren, 2012) are closely related to the motivation behind the RBIAS model. However, because users give ratings for different sets of items, measures and models from inter-rater reliability studies cannot be applied directly.

In the current RBIAS model, some ambiguities among types exist. BP type may be a combination of LB and DB types. N and NB may not be clearly separated conceptually. V itself is not clearly defined itself, but defined as a complementary of whole the other types. Further study for elaboration is required.

**Conclusion**

We proposed the RBIAS model for recommender systems in which six subtypes of rating pattern are defined: normal, like-biased, dislike-biased, bipolar, neutral-biased, and vague. Based on this set of rating types, a bias-reduction model and a rescaling method are suggested. Experimental results show that proposed methods significantly reduces RMSE and acquires higher rank-based metrics for various collaborative filtering methods in the movie recommendation domain.

Further studies are ongoing for the following issues: alternative rules for categorizing raters by their types such as in (Koenigstein et al., 2011; Rosmalen et al., 2010); feasibility in different scales to the current 5-point system; and, application to another domain such as book or music recommendations.

**Acknowledgement**

This work was supported by the NRF grant funded by the Korea government (MSIP) (NRF-2010-0017734), supported in part by KEIT grant funded by the Korea government (MKE) (KEIT-10035348, KEIT-10044009).

**References**


