The Impact of Rating Scales on User's Rating Behavior

The Impact of Rating Scales on User's Rating Behavior

Cristina Gena, Roberto Brogi, Federica Cena, and Fabiana Vernero

Dipartimento di Informatica, Università di Torino
Corso Svizzera 185; 10149 Torino, Italy
{gena,cena,vernerof}@di.unito.it

Abstract. As showed in a previous work, different users show different preferences with respect to the rating scales to use for evaluating items in recommender systems. Thus in order to promote users' participation and satisfaction with recommender systems, we propose to allow users to choose the rating scales to use. Thus, recommender systems should be able to deal with ratings coming from heterogeneous scales in order to produce correct recommendations. In this paper we present two user studies that investigate the role of rating scales on user's rating behavior, showing that the rating scales have their own "personality" and mathematical normalization is not enough to cope with mapping among different rating scales.

1 Introduction

Recommender systems help users overcome the information overload by automatically selecting potentially relevant items, based either on their similarity with items users liked in the past (content-based approach) or on the preferences of people with similar tastes (collaborative filtering approach). Recommender systems usually collect user preferences by means of "rating scales", i.e. graphical widgets that allow a user to express her preferences by means of a numerical score. According to [5], rating scales should ideally be devised so that users can express their preferences in an easy and meaningful way, and a smooth translation should be possible from the granularity of true user preferences, i.e., the number of levels among which users wish to distinguish, to the range and granularity provided by rating scales themselves [9].

Recommender systems usually provide the same rating scale to all their users. However, in a user experiment we carried out [3], we found that users have different preferences with respect to the rating scales to use for the topic they are evaluating, and that they prefer different rating scales for evaluating different topics. Thus, in order to improve users' satisfaction and promote their participation, we proposed to allow users to choose the rating scales to use in recommender systems.

This opportunity presents some challenges. In fact, recommender systems must be able to deal with ratings expressed by means of heterogeneous scales, mapping them to an internal representation, in order to generate correct recommendations. [5] found a high correlation among ratings for the same items given by means of rating scales which differ for their granularity, numbering, and visual metaphor. Consequently, they concluded that designers can safely allow users to choose any scale they prefer, since they only need to compute the ratings to use in the recommendation process by means

of a normalisation based on mathematical proportion. This is in contradiction with our findings in a similar experiment [3], where we observed that ratings expressed on different rating scales may depart considerably from mathematical proportion, motivating the idea that rating scales actually have an influence on user ratings. This insight is also confirmed by related work in the field of survey design, which reports the effect of scales on user ratings [7, 6, 2].

Given the importance of rating scales for recommender systems, and considering the controversial results reported in the state of the art and previous work by the authors, in this paper we decided to further investigate the issue of normalizing ratings given on heterogeneous rating scales.

As a first step, we aimed at experimentally confirming our past observation that rating scales actually have an influence on user ratings, and pure mathematical normalisation is not enough. To this purpose, we chose the gastronomy domain as a use case and carried out an experiment where users were asked to repeatedly assess a set of N recipes, using N different rating scales. We then confronted average user ratings on each rating scale, and we correlated all the ratings. We actually found that some rating scales are characterized by higher or lower than average ratings. This allowed to calculate a coefficient for each scale, that filters out the effects due to the use of a specific rating scale. This can be used to capture the actual meaning of user ratings, and to accurately represent user preferences.

As a second step, we aimed at confirming the results of the first experiment in a more realistic setting, i.e. in the context of use of a real recommender system. Thus, we wanted to validate i) that mathematical normalization is not sufficient, and ii) the rating scale coefficients we calculated. Therefore we carried out a controlled experiment wherein users were asked to rate a number of recipes they liked with different rating scales. We have contextualized this experiment using I-Cook, a recommender system in the gastronomy domain which builds user models based on user ratings of system-provided recipes and which offers customizable rating scales. We should notice that I-Cook currently manages rating coming from different rating scale using mathematical proportion. The inferences based on this process were not relevant for the presented experiments. Non-mathematical proportions will be managed and implemented considering the results presented in this paper.

The paper is organized as follows. We start by presenting the state of the art of rating scales studies in Section 2. Then, in Section 3 we present our definition of the main concepts we refer to in the paper. We present our experiments in Section 4: the first one is presented in Section 4.1, while the second one in Section 4.2. Finally, section 5 concludes the paper.

2 State of the Art

The role of rating scales is crucial in recommender systems where suggestions are generated by predicting ratings for items users are unaware of, based on ratings users explicitly provided for other items. It is commonly accepted that different users may use rating scales differently, and some sort of average adjusting is usually adopted in order to compensate for such an idiosyncratic behaviour (see for example [9, 1, 8, 11]). On

the other hand, relevant work in the area of recommender systems also focused on the choice of appropriate scales for collecting user ratings ([12, 9, 13])

Referring to the design of the rating process as a whole, in [12] the authors suggested to adopt a mix of different types of questions (e.g., expressing binary liking versus rating items on a Liker-like scale) and provide constant feedback on user contributions in order to keep users from getting bored or frustrated.

Distinguishing between domain features (which refer to the content being recommended) and inherent features of recommender systems, [9] points out that the granularity of true user preferences with respect to recommended contents may be different from the range and granularity of user ratings which are managed by a specific recommender system. An appropriate rating scale for a certain domain should allow users to distinguish among exactly as many levels of liking as it makes sense to them.

In [13], the authors defined the main elements that determine the design of interface aspects (corresponding to rating scales, according to our framework) aimed at presenting system predictions and at collecting explicit user feedback in the context of a TV recommender system: 1) presentation form (which quite closely corresponds to what we will call the "visual metaphor" in the rest of the paper); 2) scale of the prediction or rating (including range, precision, symmetric versus asymmetric and continuous versus discrete); 3) visual symmetry or asymmetry; and 4) use of colour. They also found that most users prefer to have predictions presented by means of five-star interfaces, while they are less in agreement as far as interfaces to provide feedback are concerned, consistently with our findings [3].

Differently from our approach, however, these works do not focus on the possible effect of different rating scales on user ratings and on ways to deal with it. Instead, starting from the consideration that a good rating scale should support users in expressing their preferences in a meaningful way and without much effort, in [5] the authors explicitly investigated the effect of rating scales on user ratings. More specifically, they asked their experimental subjects to re-rate each of three sets of movies they had already evaluated by means of the original MovieLens five-position rating scale on one of the following rating scales: a binary scale providing only thumbs up or down, a no-zero scale ranging from -3 to +3, and a half-star scale ranging from 0.5 to 5. Notice that, the authors did not explicitly focus on the possible effects of numbering and visual metaphors (unlike our case, as it will be seen later on in the paper), although they did use rating scales which differ with respect to these aspects. The authors found that ratings on all three scales correlated well with original user ratings, with no need for specific countermeasures, and suggested that designers might allow users to choose their favourite rating scale and compute recommendations by means of mathematically normalized scores. However, they also observed that users tended to give higher mean ratings on the binary and on the no-zero scales, and that new ratings on the binary scale correlated less strongly with original ratings (r = 0,706) than new ratings on the no-zero and half-star scales (r = 0,827 and r = 0,829, respectively).

The effect of rating scales on user ratings, on the contrary, is often reported in work in the domain of survey design.

In [7], the author produces some evidence that the presence or absence of a neutral point on a scale produces some distortion in the results. In particular, they found that

some respondents may choose the mid-point in order to provide a less negative answer, because of a social desirability bias. On the other hand, rating scales with no mid-point force the real indifferent to make a choice, causing a distortion the polarity of which is content-specific.

Various factors which can cause a rating scale to be biased are examined in [6], including: 1) category labels (either words or numbers); 2) effects of response alternatives on question interpretation; 3) forced choices (e.g., no neutral point is available); 4) imbalance in the number of positive and negative responses; 5) order of responses (there is evidence of a bias towards the left side of a scale) and 6) granularity.

The possible effects of numeric category labels are also investigated in [2]. In particular, the authors show that the negative-evaluation side of a scale is perceived as more negative when it is labeled with negative rather than positive numbers (e.g., -4 rather than 1), and this causes more positive evaluations and higher average ratings when scales with negative numerical labels are used.

3 A General Approach for Defining Rating Scales

In this section, we first define the three grounding concepts for our approach: rating scales, rating scale personality and user rating. Then, we describe how we deal with rating scale personality.

We define **rating scales** as complex widgets which are characterized by the following features: i) granularity, ii) numbering, iii) visual metaphor, iv) presence of a neutral position. For "granularity", we mean the number of positions of the scale: this can be coarse (e.g., a 3-point scale where only negative, neutral/intermediate and positive ratings are possible) or fine (e.g., a 10-points scale). For "numbering", we mean the numbers, if any, which can be associated to each position in a rating scale (e.g., three different 3-point rating scales might be numbered 0,1,2; 1,2,3; or -1,0,+1). For "visual metaphor", we mean the visualization form which is used to suggest the behaviour of rating, and which influences the emotional connotation of each scale: for example, a thumb rating scale shows a metaphor related to human behaviour; a star rating scale conveys a metaphor which relies heavily on cultural conventions (as with hotel ratings), while a slider rating scale is based on a technological metaphor which reminds, for example, of volume tuners. For "neutral position", we mean that an intermediate, neutral point, indicating that users have no definite opinion, is provided.

All these features contribute to define what we call the **personality** of rating scales, i.e., the way rating scales are perceived by users and affect their behaviour. In fact, we claim that rating scales are not neutral tools, but they exert an influence on people who are using them to express their preferences. Rating scales personality causes a certain rating scale to have a specific influence on user ratings, e.g., it stimulates users to express tendentially higher/lower ratings than other scales. Therefore, mappings based on mathematical proportion alone do not allow to capture the actual meaning of user ratings. We assume that rating scale personality may be defined at two levels. First, at an **aggregate** level, it is determined according to the behaviour of all users of a recommender system, and it reflects general tendencies in the use and perception of rating scales (e.g., according to [2], scales with negative numerical labels cause users to give

higher ratings on average). Second, at an **individual** level, it is determined according to the behaviour of each specific user, and it reflects personal idiosyncrasies in the use and perception of rating scales (e.g., a certain user might consistently give higher ratings when using a specific rating scale, but this behaviour might not be generalize to the whole user community). In this paper, we focused on the aggregate level.

According to our approach, **user ratings** are therefore determined by at least three elements:

- the item which is being rated;
- the personality of the user who is rating;
- the personality of the rating scale in use.

The first point is straightforward: the influence of the items being rated on user ratings is meant to represent real user preferences for such items.

By user personality we mean the fact that users may tend to use rating scales differently, for example, optimistic users may tend to assign very positive ratings for the most part. User personality has been dealt with extensively in literature (see Section 2 for references on classical approaches which adopt *average adjusted* ratings for use in collaborative filtering systems) and we do not treat it further in this paper. On the contrary, the novel aspect we focus on here is rating scale personality.

Rating scale personality should be taken into account in various scenarios. For example, in content-based and collaborative filtering recommender systems, when users are expected to change the rating scales they use over time, or to assign specific ratings, given with different rating scales, to different aspects of items (e.g., quality of food and atmosphere for a restaurant), and such specific ratings are to be somehow aggregated in a general item rating. In content-based recommender systems, considering scale personality is useful when users are expected to use different rating scales for different types of items which map to common domain categories. For example, restaurants and recipes might be mapped to a common taxonomy based on their cuisine, as for restaurants, and on their nationality or primary ingredient, as for recipes. Thus a recommender system might be able to infer the level of user interest on French (or vegetarian) restaurants based on user ratings of French (or vegetarian) recipes. Finally, in collaborative filtering systems, rating scale personality should be taken into account if different users are expected to use different rating scales from one another (even if they may not change the scale they use over time, their ratings have to be compared in order to generate recommendations).

In this paper, we investigate the impact of *aggregate* rating scale personality in two users studies, which will be presented in the following section.

4 The experiments

In this section, we present two user studies we performed:

1. A first preliminary experiment was carried out in order to: a) validate our assumption that rating scales have different personalities, i.e., they exert an influence on user ratings, and b) define numeric coefficients which formally describe rating scale personality at an aggregate level (aggregate personality coefficients) .

2. A second controlled experiment was carried out in order to further assess our approach, focusing on the scenario of a content-based recommender system where users are expected to change the rating scales they use over time.

4.1 The first experiment

The goal of our experiment was to investigate the issue of normalizing ratings given on heterogeneous rating scales. Our starting idea is that mathematical normalization is not enough for mapping user ratings expressed with different rating scales. In a previous experiment [3] we observed that ratings expressed on different rating scales depart considerably from mathematical proportion, and so that rating scales actually have an influence on user ratings. It is worth noting that in that experiment 40% of ratings departed considerably from mathematical proportion, showing that mathematical proportion is not enough to make a mapping which is able to capture the actual meaning of user ratings. We believe that each rating scale has a specific personality that may influence the rating (even if this is in contrast with other works which found different results, as described in Section 2).

In order to confirm our past results, we have designed an experiment where users have been asked to repeatedly assess a set of N recipes presented in a cuisine web site, using N different rating scales. We then confronted user ratings on each rating scale. We chose the gastronomy domain presenting common recipes as a use case since is quite likely that user has already had experience with the recipe (because she has already eaten or cooked it) and if she does not she may obtain a good idea of the recipe just reading its description (ingredients, preparation, etc.).

For this experiment, we have considered seven rating scales (see Figure 1): thumb-up/thumb-down, thumb-up/thumb-down/thumb-medium, 3-points stars, 5-points stars, 10-points stars, 3-points slider, and 10-points slider. These rating scales are different for i) the granularity they provide in selecting values: they range from a minimum of two position to a maximum of 10 position; ii) the numbering; iii) the visual metaphor (thumb, star, slider); iv) the presence of a neutral position (thumb-up/thumb-down/thumb-medium, 3-points slider). Notice that the experiment was counterbalanced in order to avoid order effects. See later for details.

Hypothesis. We have hypothesized that user ratings may vary depending on the rating scale in use and thus ratings may depart from mathematical proportion. We have also hypothesized that this deviation could be ascribed to what we defined as the "personality" of each rating scale.

Experimental Design. Single factor within subjects design. The independent variable was the rating scale manipulated according to four levels: visual metaphor, granularity, numbering, and presence of neutral position. In the first treatment condition we manipulated both visual metaphor and granularity of the rating scale asking users to perform a rating task using three rating scales differing for visual metaphor and for granularity: thumb-up/thumb-down, 10-points slider, and 5-points stars. In the second treatment condition we only manipulated the granularity using the same visual metaphor (stars). In this second condition we presented to the users the following three rating scale: 3-points stars, 5-points stars, 10-points stars. In the third treatment condition we manipulated the visual metaphor (thumb, stars, slider) leaving the same granularity (3-point

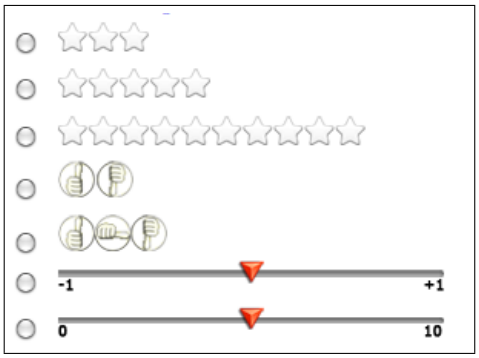

Fig. 1. The rating scales used in the two experiments.

scale), and adding the presence of a neutral position (in thumbs and slider). One rating scale (the slider) has a negative value. In this last condition we proposed to the users the following rating scales: thumb-up/thumb-down/thumb-medium, 3-points stars, 3-points slider. Subjects were randomly assigned to all the three treatment conditions.

Subjects. 21 subjects, 22-26 years old, 11 male and 10 female, students at the School of Multimedia, Arts, and Humanities, University of Turin, recruited according to an availability sampling strategy.

Measures and Material. A series of nine web pages was prepared, grouped according to the three different treatment conditions. Each page presented a set of eight recipes to be rated with a single rating scale per page, and randomly presented to each subject. We randomly varied the order of the pages in each condition, and the order of the condition served from each user. We recorded users' performance with a screen capture video, and we registered user's ratings. Users received the instruction for experimental tasks directly on the web page presenting the experiment.

Experimental task. Users were asked to read the recipes (belonging to a imaginary friend's blog) and then rate them taking into account the description, and if they would cook/eat or not the recipe. Since their friend would like to know which rating scale to use in her blog, they were asked to rate the same recipe several times with different rating scales. At the end of the test every recipe had been rated with all the eight rating scales (3-point stars were used twice).

Results. We have collected in total 1512 ratings. All scales were normalized to a zero-to-one range. We computed mean ratings (see Figure 2, first row), and we correlated original values by means of Pearson correlation in order to compare the rating behavior of the users on different scales. We found the following significant correlations (all significant at the 0.01 level). The reader should notice that we only consider correlations equal or beyond a given threshold (0.5):

– in the 3rd treatment, **thumb-up/thumb-down/thumb-medium** and **3-points slider**: r=0.861. Single values showed that users gave often the same values when using these scales. When values do not co-vary, we observed the medium value of the thumb frequently corresponds to the maximum value in 3-points slider. Even if the

granularity is the same, the negative numbering influences the rating, pushing slider values up, as sustained in [2];
- **thumb-up/thumb-down/thumb-medium** (3rd treatment) and **thumb-up/thumb-down** (1st treatment): r=0.666. Confirming general expectations, these two rating scales tend to vary together. When values do not co-vary the, 3-points scale shows mean values lower than the 2-points scale due to the presence of neutral position;
- **thumb-up/thumb-down** (1st treatment) and **3-points slider** (3rd treatment): r= 0.658. More than the previous correlations, these two scales co-vary also for higher values. The maximum value of the slider frequently corresponds to maximum value of the thumb;
- in the first treatment, **5-points stars** and **10-points slider**: r= 0.631. This correlation is lower than one could expect, and looking at single values we have noticed that, when not correlated, stars promote higher values than sliders;
- **5-points stars** (1st treatment) and **5-points stars** (2nd treatment): r= 0.616. This correlation is lower than expected: users rated the same items with the same scale but gave different ratings depending on the treatment condition. In the second treatment 5-point stars showed values slightly higher, and this trend is also confirmed by the next correlation;
- in the second treatment, **5-points stars** and **10-points stars**: r= 0.575. As in the case of stars/sliders, the correlation is lower than one could expected. Looking at single values we notice that, when not correlated, 5-points stars promote higher values.

In order to have a measure of the impact of the rating scale on the way the user rates, we have calculated a *coefficient* for each rating scales. This numeric coefficient is calculated as a ratio between the average ratings of each scale and the average ratings of 10-points stars rating scale. This scale was chosen for the recognized acceptance of the star metaphor and because scales with a fine granularity are considered more reliable [4], provided that users can handle such granularity (which certainly holds true for 10-points rating scales, see for example [5]). The coefficients we found are summarized in Table 4.1. We believe that these rating scales coefficients could represent the role of rating scales in the users ratings.

	2-p. thumb	**10-p. slider**	**5-p. stars**	**3-p. stars**	**10-p. stars**	**3-p. thumb**	**3-p. slider**
1^{st}*experiment*	1.12	0.99	1.02	1.04	1	0.84	1.17
2^{nd}*experiment*	1.05	0.92	0,98	1.08	1	1.08	0.77

Table 1. Coefficients for rating scales in the two experiments.

However, most of users ratings do not correlate, and when they do they do not correlate very well. Thus, we can affirm that these correlations do not reflect a mathematical proportion. Visual metaphor, granularity, numbering, and presence of neutral position seem to have an influence on the way the users rate. However, i) rating the same item would bias the way the user mapped her ratings on each scale, and ii) as shown in previous experiments (see [5] and [10]), the correlation between re-ratings ranges between

0.8 and 0.7. These reasons could have affected final results. Thus we have designed a second experiment wherein user can perform less constrained tasks in a more realistic setting.

4.2 The second experiment

Hypothesis. We have hypothesized that mathematical normalization fails in less biased conditions, and we wanted to validate the rating scales coefficients we calculated in our previous experiment.
Experimental Design. Single factor within subjects design. The independent variable was the rating scale manipulated according to 4 levels: visual metaphor, granularity, numbering, and presence of a neutral position. The rating scales presented to the subjects were the seven scales of the previous experiment (see Figure 1). Users were asked to choose a rating scale, then to rate five courses (appetizer, first dish, second dish, side dish, desserts) they like using that scale. After that, they had to choose another scale, and do the same tasks. Scales were presented to the user in a random order, as well as courses.
Subjects. 32 subjects, 20-69 years old, 15 male and 17 female, skilled Internet users, recruited according to an availability sampling strategy.
Measures and Material. Users were given written instructions, then they were asked to connect to I-Cook recommender system[1] and perform the experimental tasks. Users' performance was recorded with a screen capture video, they were observed in real time by the experimenter, and their ratings were registered on a database.
Experimental task. Users were asked to connect to I-Cook, then to register on the web site. After that they were asked to choose a rating scale, and to rate an appetizer, a first dish, a second dish, a side dish, a dessert they like using that scale. After that, users were asked to select another rating scale and perform the same task. User were asked to use 7 rating scales (see Figure: thumb-up/thumb-down, thumb-up/thumb-down/thumb-medium, 3-points stars, 5-points stars, 10-points stars, 03-points slider, 10-points slider) presented in a random order. At the end they had to fill in a questionnaire, and to answer a set of questions asked by the experimenter.
Results. We have collected in total 1120 ratings. We calculated mean values for every user/rating scale, then we correlated values using Pearson correlation (See Figure 2, second row). The reader should notice that in the previous experiment all values were comparable, since all the users rated the same item with the same rating scale. In this experiment mean values refer to the average values obtained from all the recipes rated by users with the same rating scales. So mean values represent the general trend of ratings obtained by using the same rating scale. The effect of the item to be rated on the

[1] I-Cook is a recommender system which suggests recipes according to user preferences inferred by the user's rating behavior. Recipes are described by several features, relating to their course (appetizer, first dish, second dish, side dish, desserts), nationality (Italian, French, Chinese, Japanese, Spanish), main ingredient (meat, fish, etc), difficulty (easy, medium, hard) and preparation time (short, medium, long). Moreover, recipes can be characterized as vegetarian and gluten free. Differently from existing systems, I-Cook allows users to use the rating scale they prefer (http://brogiroberto.altervista.org/.).

total number of ratings has been neutralized by the high number of well-known recipes to be rated. We found the following significant correlations, all significant at the 0.01 level:

- **10-points slider** and **5-points stars**: r=0.579. This correlation was already present in the previous experiment, but with an higher value. The values expressed with these two rating scales correlate quite strongly. However, when not correlated, slider values are lower then star values, as in shown the previous experiment;
- **10-points stars** and **5-points stars**: r=0.486. This correlation was already present in the previous experiment, but with an higher value. 10-points stars slightly promote higher values;
- **10-points stars** and **3-points stars**: r=0.472. This correlation was not present in the previous experiment. 3-points stars mean values tend do be higher than 10-points ones;
- **thumb-up/thumb-down/thumb-medium** and **3-points stars**: r=0.458. This correlation was not present in the previous experiment. Thumbs mean values tend do be close to the ones in 3-point stars especially when users rate medium/higher values;
- **5-points stars** and **3-points stars**: r=0.472. This correlation was not present in the previous experiment. However, when not correlated, 3-points stars values tend to be higher than 5-points stars;

	2p.thumb	10-p.slider	5-p.stars	3-p. stars	10-p stars	3-p.thumb	3-p.slider
Mean 1st exp.	0.67	0.59	0.60	0.63	0.60	0.50	0.70
Mean 2nd exp.	0.68	0.60	0.64	0.70	0.65	0.70	0.50

Fig. 2. Mean ratings in the two experiments.

More than in the previous experiment we can affirm that these correlations do not reflect mathematical proportion. To investigate the impact of single scales on ratings, we have calculated coefficient also for these rating scales. The coefficients we found are summarized in Table 4.1 (2^{nd} experiemnt row). Comparing these new values with the older ones we can make the following considerations. Some trend in the coefficients is confirmed, but with some are slight different: i) thumb-up/thumb-down promote high ratings, and the new coefficient is lower than the old one; ii) 10-points slider promotes low ratings, and the new value is lower than the old one; iii) 3-points star promotes higher ratings, and the new value is higher than the old one; iv) 5-points star values are quite close to the ones expressed with 10-points star values, but in this experiments they tend to be lower. Some other value shows an opposite trend: i) thumb-up/thumb-down/thumb-medium new values seem to promote higher ratings, while in the old experiment they promoted low ratings. As noticed above, in this experiment users exploit this scale as the 3-stars one, considering its medium value close to the one expressed

by the one of 2-points star; ii) 3-points slider promotes low ratings, while in the previous experiment promoted higher ones, as also sustained in [2]. This different trend could partly be explained by experiment design. We have noticed that users, knowing that they could change the rating scale, prefer giving negative values using the 3-points slider. Some user thinking-aloud said "I do not like this recipes, I will rate it after with the slider".

Regarding the preferences for the rating scales, the most favourite is the 5-point stars (with 16 preferences), followed by 10-points stars (with 9 preferences). All the other scales had few preferences (all 2). The worst is thumb-up/thumb-down with only one preference. This confirms the results of [13]. Most of the users (25) claimed that they appreciate the possibility to choose the rating scale. Two users did not have an opinion about this, and 5 did not like this opportunity.

We can conclude this analysis with some general insight. In general, we believe that the coefficients for capturing the actual personality of rating scales should be learnt by users' behaviour with a specific system, and cannot be calculated at priori. However, concerning the design of recommender interfaces, we notice that, in general, stars promote high ratings, especially 3-points stars, wherein 2 stars are largely used for items the user likes. Sliders promote low ratings - we can hypothesize that its design constraints encourage the criticism - and with negative labels are preferred for expressing negative ratings. Thumbs promotes high ratings, especially when used in thumb-up/thumb-down version.

5 Conclusions

In this paper we investigated the problem of how a recommender system can properly deal with values coming from heterogeneous rating scales. We present two experiments that confirm the idea that a normalization process for mapping preferences expressed with different rating scales onto a unique system representation should consider the personality of the rating scale. The main contributions of the paper are the following: i) we experimentally confirmed the idea that scales have their own "personality" and mathematical normalization is not enough, ii) we discovered that the coefficients for capturing the actual personality of rating scales should be learnt by users' behavior with a the specific system, and cannot be calculated at priori.

The benefit is that designers of recommender systems now can be aware of these issues, and should take them into account in the creation of novel enhanced recommenders.

We presented our results in a context of content-based recommender systems. However, our solution could be applied as well to collaborative filtering systems in order to compare the rating on an item given by two users using different rating scales. This could be useful to compute similarity among users, which takes into account the ratings given by the users on the same items. The coefficients we have proposed could be used to compensate the variations caused by the use of different rating scales by adjusting users' ratings.

Another aspect we should take into account in our future work is the fact that individual ratings in some case can simply depend on the evaluated item (i.e., the rating on an item

is low not for the ratings scale personality or for the user personality, but because the user does not really like the item itself). Thus, it becomes necessary to consider this aspect, in order not to confuse the effect of the rating scale with the effect of the evaluated item. For example, if the user uses an optimist scale for voting items she does not like, her ratings will be higher than using other more pessimistic scales, but the average could be low, as if the scale were pessimistic. To avoid this, some kind of semantic description of item is useful, in order to be able to compare the items and see if the users rate similar objects in a similar way.

Finally, we are planning to experiment the case of rating scale personality at the *individual* level, i.e. consider the specific rating behaviour of the individual user. Thus we will investigate to use machine learning techniques.

References

1. G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17:734–749, June 2005.
2. T. Amoo and H. H. Friedman. Do numeric values influence subjects' responses to rating scales? *Journal of International Marketing and Marketing Research*, 26:41–46, 2001.
3. F. Cena, F. Vernero, and C. Gena. Towards a customization of rating scales in adaptive systems. In P. D. Bra, A. Kobsa, and D. N. Chin, editors, *UMAP*, volume 6075 of *Lecture Notes in Computer Science*, pages 369–374. Springer, 2010.
4. J. Churchill, Gilbert A. and J. P. Peter. Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, 21(4):pp. 360–375, 1984.
5. D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '03, pages 585–592, New York, NY, USA, 2003. ACM.
6. H. H. Friedman and T. Amoo. Rating the rating scales. *Journal of Marketing Management*, 9(3):114–123, 1999.
7. R. Garland. The Mid-Point on a Rating Scale: Is it Desirable. *Marketing Bulletin*, 2:66–70, 1991.
8. K. Y. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retr.*, 4(2):133–151, 2001.
9. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.
10. W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '95, pages 194–201, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.
11. J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. The adaptive web. chapter Collaborative filtering recommender systems, pages 291–324. Springer-Verlag, Berlin, Heidelberg, 2007.
12. K. Swearingen and R. Sinha. Interaction design for recommender systems. In *Proceedings of Designing Interactive Systems 2002. ACM*. Press, 2002.
13. J. van Barneveld and M. van Setten. *Designing Usable Interfaces for TV Recommender Systems. In: Personalized Digital Television. Targeting programs to individual users*. L. Ardissono, A. Kobsa and M. Maybury editors, Kluwer Academic Publishers, 2004.