Evaluating rating scales personality

Tsvi Kuflik[1], Alan J. Wecker[1], Federica Cena[2], and Cristina Gena[2]

[1] Information Systems Department, The University of Haifa, Israel
[2] Department of Computer Science, University of Turin, Italy

Abstract. User ratings are a valuable source of information for recommender systems: often, personalized suggestions are generated by predicting the user's preference for an item, based on ratings users explicitly provided for other items. In past experiments that were carried out by us in the gastronomy domain, results showed that rating scales have their own "personality" exerting an influence on user ratings. In this paper, we aim at deepening our knowledge of the effect of rating scale personality on user ratings by taking into account new empirical settings and a different domain (a museum), and partially different rating scales. We compare the results of these new experiments with our previous ones. Our aim is to further validate in a different application context, and domain, and with different rating scales, the fact that rating scales have their own personality which affects users' rating behavior.

Keywords: rating scales, user study, recommender systems

1 Introduction

User ratings are valuable pieces of information for recommender systems: often, personalized suggestions are generated by predicting the user's preference for an item, based on ratings the users explicitly provided for other items [9]. Nowadays, with the advent of Web 2.0, which has turned users into content producers, almost all the social applications give users the opportunity to rate content (for social or for personalization purposes). Users provides their votes by means of "rating scales", i.e. graphical widgets that are characterized by specific features (e.g. granularity, numbering, presence of a neutral position, etc.). Much work in the field of survey design refers to the possible effects of rating scales on user ratings [7, 1, 8]. Reviews of psychological literature indicates that expanding the number of choice does not systematically increase scale sensitivity[3, 6]. This confirms that how people respond to different rating scales is a primarilyan issue of psychology rather than a mathematical question [5]. In the field of recommender systems there have been few works addressing the problems of how to properly translate ratings given by means of different rating scales. Cosley et al. [4] show that ratings given on different scales correlate well, and thus their approach for ratings translation from a rating scale to another is simply based on mathematical proportion. They implicitly assume that rating scales are neutral tools which do not have any influence on the user ratings themselves. Conversely, in a similar experimental task, where some users were asked to rate the same object on different rating scales [2], Cena et al. observed that 40% of the ratings departed considerably from mathematical proportion. They confirmed this insight in two subsequent experiments [9]. The results have

led the authors to define the concept "personality" of the rating scales, which exerts an influence on user ratings. Moreover, for both experiments the authors derived a series of coefficients describing the relationship between the average ratings on each rating scale and the average rating on a reference scale.

In this paper, we aim at deepening our knowledge of the effect of rating scale personality on user ratings by taking into account new empirical settings, a different domain, and partially different rating scales with respect to the experiments described in ([2] and [9]). Instead of having a small number of users rating the same item repeatedly or rating different sets of items with different rating scales, in a controlled lab experiment, we now turned to a realistic setting. We consider the case of museum guides, where real visitors rated multimedia presentations using different rating scales. Aiming at validating our previous experiment in a real setting; we compare the results of this new experiment with the previous ones in order to further validate in a different application context,and domain, and with different rating scales, the fact that rating scales have their own personality which effects users' rating behavior. The paper is organized as follows: Section 2 describes the rating scales features, while Section 3 describes the novel evaluations we carried out, presents the results, and discusses them, comparing them with previous studies. Section 4 concludes the paper with some final remarks and hints for possible future work.

2 The rating scales: an analysis

In [9] Gena et al. defined **rating scales** as complex widgets characterized by: i) *granularity*, i.e. the number of positions on the scale: coarse (e.g., a 3-points scale) or fine (e.g., a 10-points scale); ii) *numbering*, i.e. the numbers, if any, which can be associated with each position (e.g., 3-points rating scales might be numbered 0,1,2; 1,2,3; or -1,0,+1); iii) *visual metaphor*, i.e. the visualization form which influences the emotional connotation of each scale: e.g., a smiley face rating scale is a metaphor related to human emotions; a star rating scale is a metaphor which relies heavily on ranking and scoring conventions (e.g., hotel ratings); both can also convey cultural connotations; iv) *neutral position*, i.e. the presence of an intermediate, neutral point.
According to the authors, all these features contribute to define the **personality** of rating scales, i.e., the way rating scales are perceived by users and affect their behavior. Rating scales personality may cause a certain rating scale to have a specific influence on user ratings, e.g., it stimulates users to express higher/lower ratings than other scales. Rating scale personality may be also measured at two levels [9]. First, at an **aggregate** level, where it is determined according to the behavior of all users of a system, and reflects general tendencies in the use and perception of rating scales. Second, at an **individual** level, where it is determined according to the behavior of a specific user, and it reflects personal idiosyncrasies (e.g., a user might consistently give higher ratings when using a specific rating scale, but her behavior can not be generalized to the whole community). However not only the personality of the rating scale may determine final users ratings. There are at least other two elements influencing the rating, *the item which is being rated* and *the personality of the user who is rating*, e.g., optimistic users may tend to assign positive ratings.

3 The experiments

In early 2011 a museum visitors guide system was introduced to the Hecht museum[3], a small archeological museum located at the University of Haifa, Israel. The system was an advanced version of the system described in [11]. It is a web-based system that allows users to freely walk around in the museum, wearing a small proximity sensor and carrying an iPod touch. When they are detected at the vicinity of a point of interest, they are offered a selection of multimedia presentations about objects of interest. Once they selected a presentation and viewed it, they are required to provide feedback about their satisfaction from the presentation before continuing the visit (i.e. providing feedback is mandatory before the user continues to use the system).
As part of the design of the user interface 5 different feedback mechanisms designs were implemented and integrated into the system (presented in Fig.1) in order to explore whether the interface design of the rating scales have an impact on the ratings, as suggested by [9].

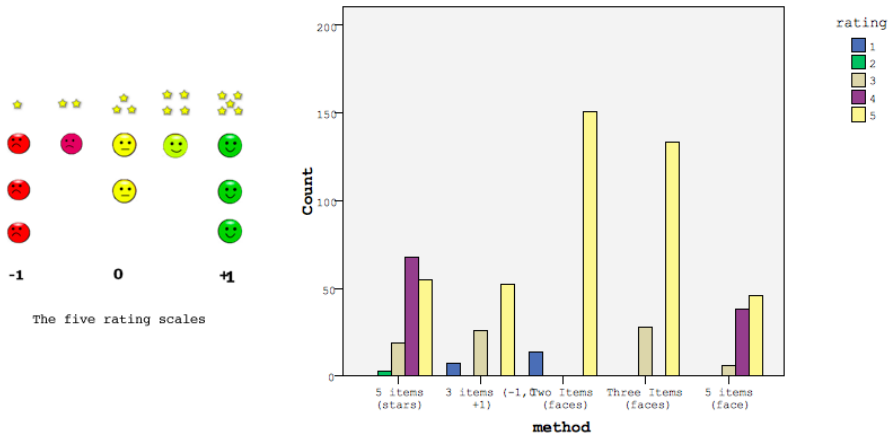

Fig. 1. The used rating scale (left side) and ratings distribution (right side).

According to the features seen in the previous section, the rating scales in this experiment differ in granularity (from 2 to 5 points), in metaphors (human emotions: the smiley faces; school marks/degrees: the numerical scale; scoring/ranking: the stars), in the presence of neutral position (present in stars, in 3-points faces, in the numerical scale) and in numbering (there a scale consisting of -1,0,1).

Experimental settings. For experimentation purposes, whenever a visitor logged in and started using the system, a randomly selected rating method was activated for her and used throughout the whole visit. The interactions of the visitors with the system

[3] http://mushecht.haifa.ac.il/Default_eng.aspx

were logged. The experimentation stated in October 2011 and by January 2012 we had 72 logs of visitors that provided more than 3 ratings[4]. These logs included two groups of visitors: regular visitors (34) and students (38) that participated in a study about indoor navigation and also used the guide to view and rate various presentations during their visit. For every rating scale we got between 14 to 16 user logs that we used for our analysis.

Experimental results. In general, all visitors scored the presentations high. Table 1 presents the average ratings of the 5 different methods used in the experiment. In order to be able to compare the various scales, all were converted to 1-5 scale in the following way: when there were 2 values, then 1 and 5 were used, when there were 3 values, then 1,3 and 5 were used (for example: -1 $\rightarrow$ 1, 0 $\rightarrow$ 3, +1 $\rightarrow$ 5). Looking at the table we can see that the average of the stars (1-5) and numerical values (-1,0,+1) are closer than all the faces, whose ratings are higher. Moreover, while there is little difference between 2 and 3 faces, the average score is a bit lower for 5 faces. This difference is statistically validated by Chi-Square=92.44, df=4, $p<0.01$ and Levene Test F=15.40, df1=4, df2=641, $p<0.01$. Using Duncan Homogenous Post Hoc testing we see that the face form a subgroup (that their harmonic means don't significantly differ) with confidence value of at least 90% and the stars and numbers form a subgroup with 78% confidence level.

Considering standard deviation, it seems that numerical values, followed by 2 smiley faces, produce somehow more noisy data (higher STD) than the other rating scales.

It is interesting to note that when considering the use of neutral point, e.g. the 3 smiley faces, users preferred the neutral face instead of using the lower value and they used the neutral value more, as compared to the 2 faces scale (see Fig. 2).

Rating Scale	Gran.	Numb.	Metaphor	Neutral pos.	Average	STD	Freq.	Coeff.
-1 0 +1	3	x	Marks/Grades	x	4.06	1.294	85	0.921
5 stars	5		Cultural conventions	x	4.21	0.744	145	0.955
2 smiley faces	2		Human emotions		4.66	1.118	165	1.058
3 smiley faces	3		Human emotions	x	4.65	0.760	161	1.055
5 smiley faces	5		Human emotions	x	4.44	0.659	90	1.008

Table 1. Rating scales described according to their personality and experimental results.

It should be noted that we have considered the average values of the rating scales at the aggregate level, namely the values reflect the behavior of all users of the system, and thus they reflect general tendencies in the use and in the perception of rating scales. However, since ratings may be also influenced by the evaluated item and by the evaluating user, we also have taken into account the following aspects:

In our experiment, visitors watched and rated 43 different presentations and provided 646 ratings for them, giving an average of 15 ratings for presentations. We calculated the average standard deviation, 0.47 (in a 1 to 5 range) with 14% of values higher than 1. Thus, presentations seem to have received quite homogeneous rates, closer to the higher values. The 5 score received 64.6% of rates, which could mean, that more often than not, users probably picked and therefore rated presentations they liked.

[4] Visitors that viewed and responded to less than 3 presentations, did not use the guide in practice, so they were discarded from the analysis, there were only a few such cases

In order to have a measure of the individual user rating trend, we concentrated on the medium standard deviation, 0.44, with 9% of values higher than 1. Thus users tend to rate presentations they like in a consistent way, namely using the same value while using the same rating scale.
We classified in Table 1, the rating scales, according to their features, their average value, and their coefficient. The *coefficient* has been calculated in order to have a measure of the impact of the rating scale on the way the user rates. It is computed as a ratio between the average ratings of each scale and the average of all the ratings.
As already noted, faces-based rating scales tend to push up the ratings. In particular 2-points and 3-points faces rating scales show a similar medium score. As seen in Fig. 1, the presence of the neutral position (in 3-points smiley face) produce some little distortion in the distribution results, as mentioned in [8]). The neutral face is used more compared to the lower face value in 2 faces scale, while the smiling face is used less compared to the 2 faces scale, thus balancing at the end the final score. In the experiment 2-points and 3-points faces rating scales also seem to correlate with the quantity of ratings, since users using these scales rated more items than other scales (11,8 and 12,6 rates per user vs. 9 rates per users of numbers and 5 faces), except for users using the stars rating scales (13,1 rates per user).
Another interesting findings is related to the low values obtained by the "-1 0 +1" rating scale. [1] found that scales with negative numbers have higher ratings since the negative number is perceived as more negative, so users tend to avoid it. In fact, the negative score has been used 2,6% of times, while the neutral point ("0") has been used 38.5% of times, probably causing the score to be so low.
Discussion. We conclude by providing some insights regarding the features of rating scales as described in Sec. 2: granularity, numbering, metaphor, and neutral position.
Regarding *granularity*, the experiment confirms the result reported in [9] that showed that rating scale characterized by a coarse granularity promoted rates higher than the average. Regarding *numbering*, the experiment shows that the explicit presence of number, even if with negative values, promotes lower ratings, similar to the sliders in the experiment described in [9](-1,0,+1 and 0-10). Regarding *neutral position*, we have observed that particularly 2-point and 3-point faces rating scales show similar score, demonstrating that the presence of the neutral position (in 3-point face) produce some little difference in results. We have noticed that, at least, that the neutral position is probably preferred to its corresponding lower values, probably due to social desirability bias, see [8].
As far as the *metaphor*, looking at the results we could conclude that rating scales sharing the human metaphor seem to correspond with higher results then other scales. Also in [9] the rating scales conveying a metaphor related to human behavior - the thumb - corresponded to rates higher than the average values.
In particular, 2-point and 3-point faces, which are very popular on the Web and on social applications, show similar trend and corresponded with the user rating more items. Another rating scale that seems to correspond with higher results is the 5-stars rating scale, which is one of the most used scale in rating-based systems. Thus we may hypothesize that the **popularity** of a rating scale is a another feature that needs to be taken into

consideration, and contributes to define the rating personality. This hypothesis needs to be verified in future experiments.

4 Conclusion

In this paper the collected data confirm that that rating scales have a "personality" which exerts an effect on user ratings. While visitors in general favored the presentations they viewed, the average ratings differed (in some cases the differences were statistically significant) between the different rating scales. The implication of our findings is that it is necessary to consider the rating scales personality when translating from one scale to another, since pure mathematical solutions are fundamentally untrustworthy [3]. In fact given the different distributions no linear transformation can exist [10]. This translation can be useful, for example, when users of system can choose the rating scales to vote on items, and thus the system must transpose ratings in a unique scale [2]. A transposition is necessary when systems exchange the user's ratings in a user model interoperability scenario. Finally, sometimes researchers need to compare scores derived from different rating scales [3]. As part of a future work, we are planning to test our ideas with more users. We are also working on finding an approach for translating user ratings that can be used by different applications.

References

1. Amoo, T., Friedman, H.H.: Do Numeric Values Influence Subjects Responses to Rating Scales? Journal of International Marketing and Marketing Research 26, 41–46 (2001)
2. Cena, F., Vernero, F., Gena, C.: Towards a customization of rating scales in adaptive systems. In: UMAP. pp. 369–374 (2010)
3. Colman, A., Norris, C., Preston, C.: Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. Psychological Reports 80, 355–362 (1997)
4. Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., Riedl, J.: Is seeing believing?: how recommender system interfaces affect users' opinions. In: SIGCHI 2003. pp. 585–592. ACM
5. Cummins, R., Gullone, E.: Why we should not use 5-point likert scales: The case for subjective quality of life measurement. In: Second International Conference on Quality of Life in Cities. pp. 74–93 (2000)
6. Dawes, J.: Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales. International Journal of Market Research 50, 61–77 (2008)
7. Friedman, H.H., Amoo, T.: Rating the rating scales. Journal of Marketing Management 9(3), 114–123 (1999)
8. Garland, R.: The Mid-Point on a Rating Scale: Is it Desirable. Marketing Bulletin 2, 66–70 (1991)
9. Gena, C., Brogi, R., Cena, F., Vernero, F.: The impact of rating scales on user's rating behavior. In: UMAP. pp. 123–134 (2011)
10. Kaptein, M.C., Nass, C., Markopoulos, P.: Powerful and consistent analysis of likert-type rating scales. In: CHI'10. pp. 2391–2394 (2010)
11. Kuflik, T., Stock, O., Zancanaro, M., Gorfinkel, A., Jbara, S., Kats, S., Sheidin, J., Kashtan, N.: A visitor's guide in an active museum: Presentations, communications, and reflection. J. Comput. Cult. Herit. 3, 11:1–11:25 (2011)