On the Use of Bipolar Scales in Preference–Based Recommender Systems

Miguel-Ángel Sicilia and Elena García

Computer Science Department. Polytechnic School.
University of Alcalá. Ctra. Barcelona km. 33.6
28871 – Alcalá de Henares, Madrid (Spain)
{msicilia, elena.garciab}@uah.es

Abstract. Recommendations in e–commerce collaborative filtering are based on predicting the preference of a user for a given item according to historical records of other user's preferences. This entails that the interpretation of user ratings are embodied in the prediction of preferences, so that such interpretation should be carefully studied. In this paper, the use of bipolar scales and aggregation procedures are experimentally compared to their unipolar counterparts, evaluating the adequacy of both techniques with regards to the human interpretation of rating scales. Results point out that bipolarity is closer to the human interpretation of opinions, which impacts the selection of recommended items.

Keywords. Collaborative filtering, e–commerce recommendations, bipolar aggregation.

1 Introduction

Recommender systems in e–commerce are aimed at helping customers by suggesting them products that could be of their interest, according to some algorithm the operates on navigation or purchase history or any other kind of data regarding products and customers. More specifically, *collaborative filtering* (CF) techniques [8, 10, 5] analyze preference data for the purpose of producing useful recommendations to customers. CF systems proceed by first matching the target user against the user database to discover *neighbors* — i.e. users that have historically had similar preferences —, and then recommending products that neighbors like, since it is assumed that the target user will "probably" also like them [9]. Other recommendation approaches are content–based, i.e. they use some kind of semantic representation of the product descriptions and use them as a source of similarities for the task of selecting recommendations. Content–based and preference–based techniques are complementary, as demonstrated in existing recommender systems, e.g. [7].

The rationale behind collaborative filtering algorithms has been said to be the automation of the process of "word-of-mouth", by which people recommend products or services to others with similar taste [10], so that preferences (either explicitly or implicitly collected) are the main source for recommendations. But in most current e–commerce systems, customers are not informed about the identity of their *neighbors*, so that "reputation" in trusting recommendations is not exploited, and in fact, it would be almost impossible to use in practice, due to the large population of users and the generalized unwillingness to reveal oneself's identity. In consequence, the mathematical models used to predict user preferences only deal with past recorded preferences, which are in most cases expressed in numerical scales, e.g. $\{1, 5\}$ or $[1, 5]$. If we look at the problem from the perspective of modeling human trust processes, it can be hypothesized that the *interpretation* of such scales and the volume of neighbors that are taken into account for each given recommendation — among other aspects — influence the trust of customers with regards to the "quality" of the recommendation. The latter aspect has been somewhat addressed in the diverse techniques designed to overcome the so–called *latency problem* — i.e. the problem of how CF system should behave when they have low volumes of historical data — , but the former one remains largely neglected.

In this paper, the polarity in the *interpretation* of numerical preference scales is studied from the perspective of its influence in the degree of trustworthiness of preference predictions, provided that explanation details for them (like those described in [3]) are showed to customers. The main objective of such inquiry is to come up with some evidence to devise CF algorithms that behave more closely to humans in the process of inferring preferences from the judgements of anonymous peers, eventually resulting in more "commonsensical" approaches to generate and explain recommendations. At the best of our knowledge, this is the first study regarding polarity in e–commerce ratings used for recommendation. The consideration of polarity in ratings may eventually result in more "conservative" recommendations, that tend to penalize the recommendation of an item that has received ratings in the "negative" part of the scale, thus avoiding compensation. This points out to the necessity of combining bipolar interpretations with a notion of 'democracy' as the one used in `RACOFI` [1].

The rest of this paper is structured as follows. Section 2 details the overall motivation for the present research, and Section 3 describes a concrete experimental study that provides evidence of the influence of polarity in the human judgement of preference predictions. Finally, conclusions and future research directions are sketched in Section 4.

2 Bipolar Aggregation versus Unipolar Preference Predictors

The goal of a CF algorithm is that of predicting the degree of preference of a given item or product for an specific user based on the user's previous likings and the opinion of other like–minded users. A typical CF setting consists on a set of users $\mathcal{U} = \{u_1, \dots u_m\}$, a collection of items $\mathcal{I} = \{i_1, \dots i_n\}$, and a collection of ratings that can be modeled as a relation $\mathcal{R}$ as defined in expression (1), where $\mathcal{S}$ denotes the rating scale used.

$$\mathcal{R} : \mathcal{U} \times \mathcal{I} \rightarrow \mathcal{S} \quad (1)$$

Typical scales are integer or real intervals, so that they can be normalized to other intervals without loosing information (this is used in this paper only for notational convenience). The relation $\mathcal{R}$ is usually incomplete, so that many ratings for pairs $(user, item)$ are actually missing (simply because users normally rate explicitly only a small portion of the item database). This leads to implementing prediction of ratings, in which $\mathcal{R}$ is considered to produce as output the explicit rating of a user (if available) **or** an estimation (prediction) based in the collection of explicit ratings, so that the relation can be defined in terms of a rating matrix $\mathcal{M}$ storing the explicit ratings, and resorting to a prediction algorithm for missing values — see expression (2).

$$\mathcal{R}(u, i) = \left\{ \begin{array}{c} \mathcal{M}[u, i] \;\; if \mathcal{M}[u, i] \neq null \\ pred(\mathcal{M}), \;\; otherwise \end{array} \right\} \quad (2)$$

Score prediction processes in collaborative filtering (i.e. *pred* functions) —like the classical Pearson correlation–based one described in [8]— can be considered as complex aggregation processes that take as input the history of ratings and produce the "expected" rating for a concrete item of a specific user, which serves as a basis for recommendation decisions. In fact, lightweight approaches like the one described in [1] are also based on historical records.

Bipolar aggregation operators [6] act on the interval [-1,1] instead of the unipolar unit interval, dealing with positive, supporting information as well as negative, excluding one. This difference may influence significantly rating predictions due to the consideration of negative ratings as inhibitors of preference matching, in what can be considered as *conservative* strategies to prediction. This has lead us to study the influence of bipolarity in CF settings. More concretely, the first two research questions addressed are described in what follows.

Hypothesis 1 *Users of e–commerce sites interpret rating scales as bipolar ones, with negatives acting as inhibitors of recommendation.*

Hypothesis 2 *The use of negative weights according to bipolar scales is interpreted by users as more adequate than unipolar interpretations.*

The first hypothesis is directly connected with the human interpretation of rating scales in e–commerce, and the second one complements it by suggesting that the bipolar interpretation positively influences prediction "appropriateness" as seen by users.

Previous work have raised research questions about the provision of explanations for CF recommendations [3] showing users the rationale for the prediction process, but always operating on a unipolar interpretation. Our current focus introduces a new variation in existing models that could affect the whole prediction process.

3 Experimental Study

In this section, the results of an experimental study aimed at gathering evidence about hypotheses 1 and 2 are described. The study used the large *MovieLens*[1] rating database that contains more than a million ratings from approximately 3.900 movies made by 6.040 users.

3.1 Experimental Design

Hypothesis 1 states that the interpretation of the rating scales is bipolar (at least in a significant proportion). This is to say that the scale is interpreted as having a "neutral" element that distinguishes between two opposite notions, as in "good/bad", rather than being interpreted in unipolar, comparative terms as in "more satisfactory than". Bipolarity has been studied in attitude measurement, and the the mid–point on bipolar scales is considered to "represent the neutral point in attitude"[11], so that we will follow a similar initial assumption for ratings.

The experimental design for this first question was based on asking participants to assess rating exemplars. Concretely, there were obtained five significant ratings for each participant, extracted from the *MovieLens* database (using the prediction procedure described in [5]), and distributed over the rating interval to prevent biases related to the distribution of predictions.

Users were asked to assess two related aspects about each of the examplars:

- To classify them as "good" or "'bad" films according to their (aggregated) rating, allowing for any arbitrary linguistic hedge to be added to the rating.
- To answer wether a "negative" (i.e. lower than the midpoint) rating should influence negatively her decision to recommend the item to other users, below the averaging of the ratings.

These questions provided a measure of the bipolar interpretation of the ratings, along with its intensity of influence in consuming decisions. To avoid biases, it was required that the users had neither watched the movies nor have heard previous comments about them. An example set of ratings for question 1 could be (1.05, 1.99, 3.07, 4.0, 4.98), which are (approximately) distributed over the rating interval.

The second question was investigated through a comparison between two lists of ratings that yield the same average rating, but with one of them having greater variance (due to some negative ratings, compensated with positive ones).

Hypothesis 2 is aimed at measuring the comparative "rationality" of predictions for two standard *pred* functions that differ in the consideration of bipolarity. In this case, experimental design requires the presentation of concrete prediction cases to users, describing the history of ratings for each concrete situation, and asking them for which prediction is seen as more acceptable. In addition,

[1] Available at http://www.cs.umn.edu/Research/GroupLens/

the number of ratings used for each prediction is fixed to a specific constant, to isolate the study from the influence of the size of the ratings database. In order to make the procedure feasible and non–biased, the details of the computation procedure are not disclosed to participants. The mathematical models used for the comparison are based in the classical *GroupLens* heuristic described in the seminal paper [5]. Expression (3) shows the model for predicting the rating to item l by user u, where correlation coefficients (between each pair of users a and b) are in the form described in (4), being $v_{x,y}$ the explicit rating element $\mathcal{M}[x,y]$ and $\overline{v_x}$ is the average rating of user x.

$$p_{u,l} = \overline{v_u} + \frac{\sum_{i\in\mathcal{U}}(v_{i,l} - \overline{v_i})w(u,i)}{\sum_{i\in\mathcal{U}}|w(u,i)|} \tag{3}$$

$$w(a,b) = \frac{\sum_{j\in\mathcal{I}}(v_{a,j} - \overline{v}_a)(v_{b,j} - \overline{v}_b)}{\sqrt{\sum_{j\in\mathcal{I}}(v_{a,j} - \overline{v}_a)^2 \sum_{j\in\mathcal{I}}(v_{b,j} - \overline{v}_b)^2}} \tag{4}$$

Bipolarity in expressions (3) and (4) can be introduced by changing the one to five scale to [-1,1] by the simple transformation $y = \frac{x}{2} - 1.5$, but this by itself do not change the interpretation of ratings under zero as negative. An additional transformation is required to differentiate the influence of negative ratings in the overall prediction. We have chosen not to change the correlation coefficient in (4) to avoid changing its robust interpretation of matching profiles, so that it is expression (3) which becomes modified. Expression (5) shows the simple change introduced, i.e.

$$p'_{u,l} = \overline{v_u} + \frac{\sum_{i\in\mathcal{U}}(\Phi(v_{i,l}) - \overline{v_i})w(u,i)}{\sum_{i\in\mathcal{U}}|w(u,i)|} \qquad \Phi(v_{i,l}) = \left\{ \begin{array}{l} v_{i,l}, , \quad v_{i,l} \geq 0 \\ \frac{v_{i,l}}{k}, , \quad v_{i,l} < 0 \end{array} \right\} \tag{5}$$

The k value in (5) acts as a parameter of the influence of negative ratings in the overall prediction[2]. Values in the [1,5] interval can be used to produce reasonable conservative variants of different intensity in large rating databases like *MovieLens*. For example, the predicted ratings for two specific user and item pairs are provided in Table 1.

user, item	k=1 (unipolar)	k=1.25	k=1.5	k=1.75	k=2	k=5
4, 1	3.28	3.18	3.08	2.98	2.88	2.48
7, 13	2.6	2.4	2.16	1.99	1.7	0.98

Table 1. Example bipolar predictions for concrete user and item pairs with diverse values of the k parameter.

[2] k could be also used to decrease the rating proportionally to its negative intensity, but we will not deal with this here.

As illustrated in Table 1, the bipolar correction provides a slight modification for cases with a low proportion of negative ratings (row one), so that a significant amount of negative ratings is required to make a difference with the standard approach(row two). In consequence, the problem of finding an “ideal value” for k is dependant on the profile of negative ratings in the database, and on the perception of users about the effect bipolarity should have in the final ratings. The second part of this study is intended to gather some initial evidence about this issue.

3.2 Results and Discussion

The profile of the users that participated in the study was that of students of Computer Science aged 20–35, and considered regular e–commerce buyers with around six to twenty purchases per year through the Web. Most often consumed product were books, music, video–games and movies. The experiment took place at one of the University laboratories. In what follows, the results and main findings are briefly described.

Hypothesis 1 Considering that a majority of e–commerce users can be properly represented by a 80% (this decision may seem controversial, but clearly represents a concept of ‘majority’ for the purposes of this study), the null hypothesis can be formulated in terms of the proportion of individuals that provided a (consistent) bipolar interpretation to samples.

Thus, we have $H_0^1 : \mathit{bipolar} \geq 0.8$ and $H_1^1 : \mathit{bipolar} < 0.8$. With a significance level $\alpha = 0.05$ and using a z–test we have that $z = \frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.77-0.8}{\sqrt{0.8 \cdot 0.2/123}} = -0.766$ which does not entails the rejection of H_0^1. In addition, if we count as successes users that do not adhere to bipolarity having the value 3 as midpoint, but at a higher or lower value, the proportion of bipolar interpretations grows to 0.91 which is consistent with even stronger null hypotheses.

Several linguistic hedges were used by more than one user. Table 2 details the frequencies of the most employed ones (translated from the original Spanish expressions). The significant but not completely consistent frequency (i.e. the use of the same height for different numerical ratings depending on the user) of use of a number of linguistic hedges suggests that the intensity of positive and negative polarities have not clear boundaries.

The second question was used to study the relation between bipolar interpretations and recommendation decisions. A simple χ^2 test between two variables, called $X = \mathit{bipolar}$ and $Y = \mathit{negative}$ respectively — with Y being the users that think that below–midpoint values should influence (negatively) the final recommendation of the item — can be used to assess such relationship, with $H_0^{1'} : \chi^2 = 0$ and $H_1^{1'} : \chi^2 > 0$. Given $\alpha = 0.05$, $\chi^2 = 10.258$, which has a significance level below 0.005 for $\mathit{df} = 1$, so that we can consider to have some degree of interaction between the criteria.

hedge	**frequency**
very	62
rather	53
extremely (translated from the slang “super”)	32
not very	12
spanish superlative	9

Table 2. Most frequently used hedges

Results for Hypothesis 1 point out that a significant proportion of users do interpret common one–to–five rating scales (like the one used in *amazon*[3]) as bipolar. Results for question 2 of hypothesis one points that bipolarity is interpreted as “negative” information influencing recommendations below average–based compensation. Both results can be considered as evidence in favor of devising bipolar approaches to recommendations, as the straightforward one studied in Hypothesis 2.

Hypothesis 2 A sample of 25 predictions from *MovieLens* were used for this part of the study, reasonably covering the domain of resulting ratings. Then, the results of the original unipolar prediction algorithm (in which $k = 1$) were put together with te results of a number of parameterized bipolar versions with $k \in \{1.25, 1.5, 1.75, 2, 5\}$. The examples were presented to the users, providing the frequencies of ratings for each value in the one to five point scale that were used to compute the predictions.

Table 3 provides the results of the study in terms of frequencies of first and second–option selection of each prediction version.

	k=1 (unipolar)	k=1.25	k=1.5	k=1.75	k=2	k=5
Preferred option frequencies (o_1)	16	27	39	26	15	0
Second–best option frequencies (o_2)	19	23	28	29	22	2

Table 3. Frequencies of preference for each of the prediction versions

Table 3 can be interpreted as a tentative degree of acceptability of bipolar interpretation intensity. Such acceptability is showed in Figure 1, in which the results from both a least–square (LS) and a fuzzy regression method [4] are depicted. The LS regression obtained the expression $a = -164.2 + 282.3 \cdot k - 93.71 \cdot k^2$, where a represents the values of $o_1 \cdot \frac{o_2}{2}$, which here is intended to represent “strength of preference” for a value of k. A fuzzy variant has been used just to try with an alternate method in which an explicit modeling of input imprecision can be used, but no significant divergences have been found.

[3] http://www.amazon.com

Such degree can be considered as an elicited parameter from users of the rating database, but further testing is required to assess its generality.

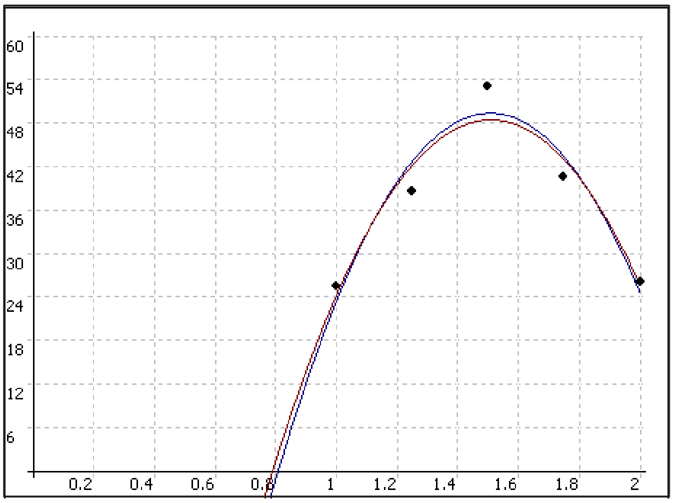

Fig. 1. Regression curves for approximate preferences on the degree of bipolarity k, where $k = 1$ represents unipolarity.

In any case, results evidence that negative interpretations are often considered as more appropriate than unipolar ones to a large extent, although the intensity of such interpretation is still subject to empirical adjustment.

4 Conclusions and Future Work

A bipolar interpretation of rating scales in the context of e–commerce entails slightly modified collaborative–filtering recommendation algorithms that are more conservative in the presence of negative ratings. A concrete study has been described in order to explore this issue and gather some initial evidence regarding the bipolar interpretation of rating scales and their perceived influence in final recommendations. Results point out that bipolarity may lead to recommendation strategies that are more consistent with the human interpretation of other's ratings. It is commonly acknowledged that the most important errors to avoid in e–commerce recommendations are *false positives* — as pointed out in [9]—, since they may lead to "angry customers". In consequence, bipolar approaches may eventually be more appropriate to reduce false positives, due to its consideration of negative ratings as inhibitors of the recommendation process.

Further studies are required to obtain a more general insight on bipolar–rating recommendations, extending both the user population and experimental setting and also covering other, more recent collaborative recommendation procedures [9]. In addition, future work should address the measure of accuracy called "*AllBut1* Mean Average Error" [1] for different values of k in existing rating databases.

Future work should also study the effect of considering bipolarity in the resulting amount of false positives generated by the recommender system, and in the concrete form of bipolar aggregation that best captures the human interpretation of positive and negative item assessment in concrete item categories and rating contexts. In the case of content–based approaches, bipolar decision operators [2] can be used to model complex situations.

References

1. Anderson, M., Ball, M., Boley, H., Greene, S., Howse, N., Lemire, D., McGrath, S. RACOFI: A Rule-Applying Collaborative Filtering System, In *Proc. IEEE/WIC COLA'03*, Halifax, Canada, October (2003).
2. Grabisch, M. and Lebreuche, C. Bi–capacities for decision making on bipolar scales. In: Proceedings of EUROFUSE Workshop on Information Systems, 185–190 (2002)
3. Herlocker, J., Konstan, J., and Riedl, J.: Explaining Collaborative Filtering Recommendations. In: Proceedings of ACM 2000 Conference on Computer Supported Cooperative Work 241–250 (2000)
4. Izyumov, B., Kalinina, E., Wagenknecht, M., Software tools for regression analysis of fuzzy data. Proceedings of 9th Zittau Fuzzy Colloquium, Zittau, Germany (2001), pp. 221-229.
5. Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J.: GroupLens: Applying Collaborative Filtering to Usenet News. Communications of the ACM 40(3): 77–87 (1997)
6. Mesiarová, A., Lázaro, J. and Calvo, T.: Bipolar Aggregation Operators. In: Proceedings of the International Summer School on Aggregation Operators and their Applications, 119–122 (2003)
7. Paulson, P. and Tzanavari, A.: Combining Collaborative and Content–Based Filtering Using Conceptual Graphs. In: J.Lawry, J.G.Shanahan and A.Ralescu (eds.): Modeling with Words: Learning, Fusion, and Reasoning within a Formal Linguistic Representation Framework, LNAI 2873, Springer–Verlag Berlin Heidelberg 168–185 (2003)
8. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl J.: GroupLens: An open architecture for collaborative filtering of netnews. In: Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, NC: ACM, 175–186 (1994)
9. Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J.: Analysis of Recommender Algorithms for E-Commerce. In: Proceedings of the ACM e–Commerce 2000 Conference. 158–167 (2000)
10. Shardanand U. and Maes, P.: Social information filtering: Algorithms for automating "word of mouth". In: Proceedings of CHI'95 – Human Factors in Computing Systems, 210–217 (1995)
11. Mehling, R. A Simple Test for Measuring Intensity of Attitudes. Public Opinion Quarterly, 23, 576–578(1959)